Towards a recommender system for bush taxis

Sébastien Gambs
Université de Rennes 1 - INRIA / IRISA
Campus Universitaire de Beaulieu
35042 Rennes, France
Email: sgambs@irisa.fr

Marc-Olivier Killijian,
Miguel Núñez del Prado Cortez
and Moussa Traoré
CNRS ; LAAS ,
7 avenue du Colonel Roche
F-31077 Toulouse, France
Email: {killijian, mnunezde, mtraore}@laas.fr

***Abstract*—To improve the transport efficiency and to reduce the traveller stress, we introduce a recommender system for bush taxis in Ivory Coast whose main objective is to propose to pedestrians potential means of transportation in their neighborhood whose destination match their own destination. The prediction of the next location relies on a mobility model called Mobility Markov Chain. One of the strength of the proposed recommender system is that it is fully automatic as a user does not need to explicitly express his next destination but rather the system tries to infer it based on his past mobility behavior. Moreover, the recommendation algorithm is biased towards suggesting the means of transportation that are the cheapest (if one is available). The preliminary evaluation of the recommender system conducted on one of the D4D dataset shows that approximately 99% of the time in less than 30 minutes, the system is able to suggest a mean of transportation that is at most 1 kilometer away from the current position of the user for an accuracy of the prediction of the next location that is between 30% and 50% depending on the complexity of the mobility behavior of the user considered.**

***Keywords*-Location-based service, Recommender system, Next place prediction, Mobility Markov chain.**

I. Introduction

Transport is a vital economic component for developing countries. Indeed, when a performant transport infrastructure exists, people are able to take advantage of the wide variety of business opportunities, thus increasing their income level and improving their standard of living. However, in many countries, the access to transport remains an unsolved challenge, in particular in fast growing urban areas. Therefore, governments have incentive to develop innovative transportation services that are able to cope with the wide variety of the mobility needs of the citizens of these countries.

In this paper, we tackle this issue by introducing a recommender system for bush taxis whose main objective is to propose to pedestrians potential means of transportation in their neighborhood whose destination match their own destination. More precisely for each user, we first build a mobility model summarizing their behavior called a Mobility Markov Chain (MMC) [4]. The MMC is learnt from the mobility traces of the user, which in the context of the D4D challenge correspond to calls made from cell towers. Afterwards, once a MMC has been learnt, it can be used to predict the next location visited by a user based on the knowledge of his current position. In order to suggest a potential mean of transportation, the recommendation system first starts to guess the next location of the user and then scan the neighborhood for potential vehicles (*e.g.*, bush taxis, colored woro-woros, gbakas and traditional taxis) whose destination matches the next location of the user. The destination of a mean of transportation is also predicted by learning a MMC representing his mobility. One of the main advantage of our recommender system is that it is fully automatic, in the sense that a user does not need to explicitly express his next destination but rather the system tries to infer it based on his past mobility behavior. Moreover, the recommendation algorithm is biased towards suggesting the means of transportation that are the cheapest (if one is available). The preliminary evaluation of the recommender system shows that approximately 99% of the time in less than 30 minutes, the system is able to suggest a mean of transportation that is at most 1 kilometer away from the current position of the user.

The outline of the paper is the following. First, in Section II we describe the related work on recommender systems for taxis. Then, respectively in Sections III and IV, we review briefly the different transportation modes available in Ivory Coast as well as the D4D datasets. Afterwards, we introduce in Section V the Mobility Markov Chain (MMC), which is the mobility model that we used to predict the next location visited by a user before conducting a mobility analysis of the D4D dataset in Section VI. The recommender system for bush taxis is presented in Section VII and then we conduct a preliminary evaluation of its performance in Section VIII, before concluding in Section IX.

II. Related work

In recent years, mobility traces of GPS-enabled vehicle, including taxicabs, have been collected on a massive scale [7]. This collection of mobility traces of taxicabs has fostered the research in urban vehicle transportation, in particular in taxi recommender systems.

For instance, Lee, Shin and Park [8] have proposed a recommender system for taxis. These researchers have analyzed

80,000 traces stored in the taxi telematics system of Jeju island (Republic of Korea). Basically, their positions as well as their speed are reported by the taxis to a central server along with their current status (*i.e.*, free or occupied). From these mobility traces, the researchers first extracted potential pick-up points for taxis by observing the changes of status of taxis from free to occupied. Afterwards, the classical k-means algorithm [9] was used to cluster together pick-up points that are close to each other and the center of each cluster represents a potential recommendation for a pick-up point (in practice the value of k was set to 100 in their experiments). The analysis was also performed by taking into account the time dimension in order to discover time-dependent pick-up patterns. The resulting recommender system is able to propose the nearest beneficial pick-up points to vacant taxis.

Ge and co-authors [6], [5] have developed a recommender system for taxi drivers suggesting rides (*e.g.*, sequence of pick-up points or parking places) so as to maximize the probability of picking-up passengers. The pick-up points are learned from the trail of mobility traces of taxi drivers that are the most successful. Then, the pick-up points are clustered using the k-means algorithm (here also the chosen value of k was 100 for the different experiments). A probability is also associated to each generated centroid measuring the frequency of pick-up events when taxi cabs pass across the corresponding cluster. To avoid the overload of the road due to taxicabs in the same area following the same recommendation, a load balancing approach is applied to distribute empty taxis through multiple paths. Moreover, to reduce the risk of fraud from taxi drivers (*e.g.*, greedy taxi drivers who overcharge passengers by taking unnecessary detours), the trajectory of each taxicab is analyzed in order to identify the ones that are unusually long, combining evidences of frauds through the Dempster-Shafer theory.

Zheng, Liand and Xu [17] have designed an application guiding users to locations in which they can wait for a vacant taxi with the main objective of reducing their waiting time. The mobility behaviors of vacant taxis in the streets of Beijing are modeled as non-homogeneous Poisson processes. The application predicts to users the nearest road segment in which they will find a vacant taxi as well as estimate of the average waiting time (30% of simulation error).

Jing and co-authors [15], [16] have introduced a recommender system for both taxi drivers and passengers. Their model describe the probability for taxi drivers to pick-up passengers while going to adjacent parking places (*i.e.*, places in which drivers wait for passengers), the average waiting time depending on the time of the day as well as the average distance for the next trip of the driver. The passenger recommendation algorithm provides two services. The first service returns a list of nearby parking places with the average corresponding waiting time while the second service outputs a nearby road segment located at walking distance for the current position of the passenger as well as the average associated waiting time before finding an empty taxi.

While most of related works have used very precise mobility traces collected by GPS devices installed directly on taxis, our approach relies on coarser mobility traces from phones providing granularity at the level of GSM towers, which is a more challenging task. However, contrary to previous works, we did not limit ourselves to recommend only standard taxicabs but we also push suggestions for other transportation modes in Ivory Coast. These other transportation modes are detailed in the next section.

III. Modes of transportation in Ivory Coast

Due to the fast growth of the population in recent years in Ivory Coast the population of urban cites have also experienced an important increase, mainly due to the galloping urbanization and rural exodus. These fast-growing cities face enormous challenges in terms of the development of their infrastructure as well as the need to cope with the increasing demand for transport. For instance, Ivory Coast has witnessed the rising of informal modes of transport complementary to the traditional buses and taxi services. Thereafter, we briefly review these existing informal modes of transportation.

1) A *gbaka* is a minibus of 14 to 22 seats covering relatively long commuting distance between the outskirts of a city and its urban center. The gbaka has a predetermined trajectory but can deviate from it due to the occurrence of a traffic jam and some suggestion done by the apprentice-gbaka (*i.e.*, the assistant of the driver of the gbaka). Indeed the apprentice is responsible for spotting potential passengers on the way and makes the driver stop if they find one. Gbakas represent around 27% of the public transport offer in Abidjan [10].
2) The (colored) *woro-woro* is an informal taxi riding only within the limits of a particular city whose color represents the municipality to which it belongs. A colored woro-woro has a maximum capacity of 5 seating places and is continuously patrolling for clients unless he is already full. White woro-woro also exist that are not directly attached to a particular city but rather act as a shuttle between neighboring cities and have predefined and well-known parking places. However, as we believe that their mobility behavior is quite similar to the one of gbakas, for the rest of the paper we will make no distinction between gbaka and white woro-woro. This transportation mean counts for 32% of the public transport offer in Abidjan [10].
3) *Bush taxi* is the cheapest mode of transportation for long distance as well as the most common mean of transportation for inter-urban travel. A bush taxi usually pick-ups passengers at a fixed station but without

	Intra-communal	Inter-communal				Inter-rural	
Trajectory	Woro-woro (colored)	Bus	Taxi	Gbaka	Woro-Woro (white)	Bush-taxi (Badjan)	Bush-taxi (504)
Fixed trajectory	✗	✓	✗	✓	✓	✓	✓
Number of seats	5	32	5	12-32	5-8	12-32	8-10
Vehicle type	Sedan	Bus	Sedan	Bus	Bus	Bus	Van

Figure 1. Summary of the transportation modes available in Ivory Coast.

following any predefined schedule. More precisely, a bush taxi generally start its journey when all its seats are filled. Nevertheless, a bush taxi may stop anywhere on the way to pick up or drop off passengers if required. Two types of bush taxis exist: the Badjan, which correspond to Toyota vans, can accommodate up to 32 passengers while the other type of bush taxis drive Peugeot 504, thus having a more limited capacity of 10 passengers at the maximum.

Figure 2. Toyota bush taxi.

4) Traditional (*i.e.*, metered) *taxi* (similar to European taxis) is a transportation mean that is widely available as this taxi is always patrolling. These last years, the business of traditional taxis has suffered from the development of informal modes of transport such as gbakas, woro-woros and bush taxis. While traditional taxis are the most comfortable mean of transportation, they are also the less preferred one due to their expensive price. A traditional taxi can contain up to 5 passengers at a time and is not restricted to operate within the limits of a particular city. However, these taxis are notorious for overcharging clients. This category counts for 17% of the public transport supply in Abidjan [11].
5) Finally, *buses* is the only organized and formal public transportation system serving the city of Abidjan in Ivory Coast. In particular with the exception of Abidjan, other cities do not have any public transportation system and therefore have to rely on the other informal modes of transport detailed before. Buses are operated by SOTRA (*Société des TRansports Abidjanais*). The existing bus lines cover all the districts of Abidjan. However, buses do not have fixed schedules, are quite disorganized and are poorly maintained. Until the beginning of 2010, buses had the monopole on public transportation in Abidjan, but due to the emergence of informal modes of transport, today SOTRA does not meet more than 12% of the public transport demand [10].

Figure 1 summarizes the characteristics of these different modes of transportation.

IV. Description of the D4D datasets

Thereafter, we briefly review the four datasets available for the Data for Development Orange challenge (D4D challenge) [12], [2].

1) *Aggregated communication between cell towers.* For each period of one hour, this dataset summarizes the number of calls as well as the total communication time that has occurred between each possible pair of cell towers. This dataset also contains the identity of the cell tower that has initiated the call. Note that the calls that have started in a particular one-hour time period are associated with this time period irrespective of the time at which they stop.
2) *Mobility traces (high-resolution dataset).* This dataset contains a random sample of 50 000 active users whose traces have been recorded over a period of 2 weeks. These mobility traces are composed of the identifiers of the cell towers from which users have made a call or send a SMS as well as the corresponding timestamp. This process is repeated for 10 two-weeks period but with other samples of 50 000 users randomly selected, which makes a total of 500 000 users overall.
3) *Mobility traces (low-resolution dataset).* This dataset is composed of the mobility traces extracted from a random sample of users (50 000) over a period of 5 months. Each trace has associated to it a timestamp and some location information corresponding to the position at which the call was made or the SMS

was sent. Contrary to the previous dataset, the only information revealed about the location is at the level of the sub-prefecture instead of the cell tower, which is a much coarser information. A sub-prefecture is an administrative unit and Ivory Coastis actually composed of 255 sub-prefectures. A table containing the location of the centers of these sub-prefectures is also given as part of this dataset.

4) *Communication sub-graphs.* This dataset is composed of the communication graphs of 5 000 users. More precisely, the communication graph of each user is generated by observing all the communications that he had with his contacts as well as the communications that have occurred between his contacts and the contacts of his contacts (which means that the graph considered up to two degrees of separation). A communication graph is obtained by aggregating all the communications that have happened over a two-week period. Thus, for the total period of 5 months, 10 different communication graphs are generated for each user. The identity of the contacts of a user have been pseudonymized by assigning a random identifier to each contact that remains unchanged over the total observation period.

While all these datasets contained valuable information, we have focus only on the dataset containing mobility traces (*i.e.* user id, timestamp, antenna id, antenna latitude and longitude) with high-resolution in the remaining of the paper. Before building on this dataset, we have first tried to analyze the mobility behaviors of the users contained within this dataset. Note that mobile phone datasets can be considered as *sporadic* location data in the sense that they expose a limited amount of data concerning the mobility of users contrary to other geolocated datasets containing the movements of individuals recorded at a high frequency (*e.g.*, every minute or every 5 minutes).

In order to differentiate between individuals and vehicles, we have segmented the users by taking into account the minimal number of different antennas they have visited as well as the minimal number of mobility traces they have in their history. As a result, Figure 3 illustrates the number of users of the dataset having respectively at least 20, 40, 100 and 500 mobility traces with 5 or 10 distinct antennas visited.

For the remaining of the paper, we have considered only users that have a "rich enough" mobility behaviors, which we define as user that have a least 40 mobility traces in their history and that have visited at least 5 different antennas. There are approximately 150 000 users corresponding to this profile in the high-resolution dataset. We choose to disregard users that do not match these characteristics, because for each user we will use the first 20 mobility traces as the *training set* from which their mobility model is learnt while the rest of the traces forms the *the testing set*, which is used

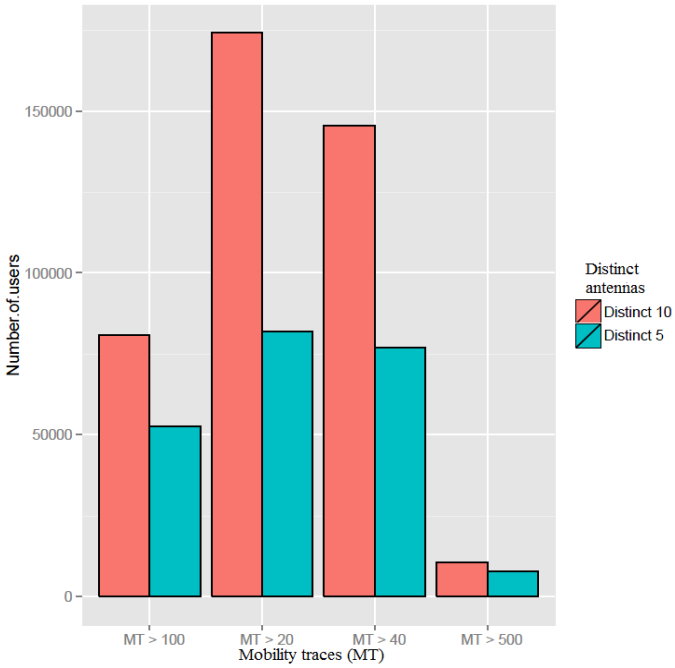

Figure 3. Distribution of the users according to the minimal number of mobility traces and the minimal number of different antennas they have visited available in their history.

for evaluating the performance of the recommender system for bush taxis. In the next section, we briefly review the mobility Markov chain [4], which is the model we used in to capture and represent the mobility of users.

V. Mobility Markov chain

A *Mobility Markov Chain* (MMC) [4] models the mobility behavior of an individual as a discrete stochastic process (more precisely as an ergodic regular Markov chain) in which the probability of moving to a state (*i.e.*, point of interests) depends only on the previously visited state and the probability distribution on the transitions between states. More precisely, a MMC is composed of the followings elements.

- A set of *states* $P = \{p_1, \dots p_n\}$, in which each state is a point of interest (POI). POIs are usually learned by running a clustering algorithm on the mobility traces of an individual, acquired for instance through a GPS-enable device or by taking the list of visited antennas identifiers obtained from his mobile phone records. These states generally have an intrinsic semantic meaning and therefore semantic labels such as "home" or "work" can potentially be inferred and associated to them.
- *Transitions*, such as $t_{i,j}$, represent the probability of moving from state p_i to state p_j. A transition from one state to itself is possible if the individual has a non-null probability from moving from one state to an occasional

location before coming back to this state. For instance, an individual can leave home to go to the pharmacy before coming back to his home. In this example, it is likely that the pharmacy will not be extracted as a POI by the clustering algorithm, unless the individual visits this place on a regular basis.

- The *stationary vector* π (also known as the steady state or the stationary distribution) is a column vector obtained by multiplying an initial column vector ((e.g.), a uniform vector) with the transition matrix repeatedly until convergence. The precise meaning of this vector depends on the type of mobility traces that have been used to build the MMC. For instance, if the POIs have been extracted by running a clustering algorithm on traces acquired through a GPS then p_i represents the percentage of the time spent by an individual in the i^{th} POI. However, in the case of D4D challenge, which consists of mobility traces obtained from call logs, p_i symbolizes rather the probability to send or receive a call or SMS while being location in the antenna corresponding to the i^{th} POI.

In a nutshell, building a MMC is a two-steps process. During the first phase, an algorithm extracts the POIs from the mobility traces. In the context of the D4D challenge extracting POIs amounts simply to read the antennas that have been visited by a particular user while for mobility traces obtained from a GPS-enabled device, this phase is more complex as it involves preprocessing the data in order to remove moving points before applying a clustering algorithm to obtained the POIs. For instance, in a previous work [4], we have used a clustering algorithm called Density-Joinable Cluster (DJ-Cluster) to discover the POIS.

During the second phase, once the POIs (*i.e.*, the states of the Markov chain) are identified, the probabilities of the transitions between states can be computed. To realize this, the trail of mobility traces is examined by chronological order and each mobility trace is tagged with a label that is either the number identifying a particular state of the MMC or the value "unknown". Finally, when all the mobility traces have been labeled, the transitions between states are counted and normalized by the total number of transitions in order to obtain the probabilities of each transition. We refer the interested reader to [4] for the details of the algorithm on how to build a MMC.

A MMC can be either represented as a transition matrix (Table I) or in the form of a directed graph (Figure 4) in which nodes correspond to states and arrows represent the transitions between along with their associated probability. When the MMC is represented as a transition matrix of size $n \times n$, the rows and columns correspond to states of the MMC while the value of each cell is the probability of the associated transition between the corresponding states.

	Home	Work	Gym	Bar	Restaurant
Home	0	1	0	0	0
Work	0,3	0	0,2	0,2	0,3
Gym	1	0	0	0	0
Bar	1	0	0	0	0
Restaurant	0.5	0	0.5	0	0

Table I
REPRESENTATION OF A MOBILITY MARKOV CHAIN AS A TRANSITION MATRIX.

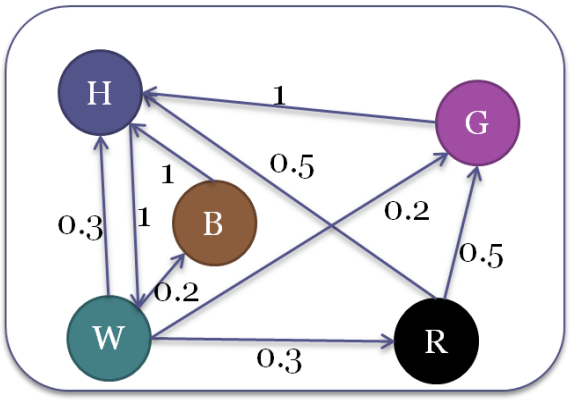

Figure 4. Representation of a Mobility Markov chain as a directed graph.

VI. MOBILITY ANALYSIS OF THE DATA

To be able to recommend different modes of transportation, we first need to understand how to categorize the mode of transport of a user. We choose to define as the possible categories for a user, the ones described previously in Section III (with the exception of buses), namely individual, gbaka, colored woro-woro, bush taxi and traditional taxi. To distinguish between these categories, we propose to rely on the three following mobility characteristics: (Shannon) entropy, predictability and average traveled distance per day. Thereafter, we briefly review the notions of entropy and predictability.

- In general, the *(Shannon) entropy* is a measure of uncertainty regarding the output of a random variable [13]. In the context of mobility, the entropy of a user quantifies the spatial uncertainty about the exact location of a user. For instance, it can be defined as the average number of binary questions that one needs to ask in order to predict the particular POI (*i.e.*, antenna) on which the user is currently located. Considering a particular user u, we can compute his entropy by applying the following formula

$$H(u) = -\sum_{i=1}^{n_u} p_{i,u} \log_2 p_{i,u} \quad (1)$$

in which p_i represents the probability to be located in the i^{th} POI for user u while n_u corresponds to the number of POIs visited by this user. For instance,

consider the situation in which Alice has visited four different POIs forming the following set $\{A, B, C, D\}$. For each of this POI, the number of recorded mobility traces is respectively 40, 20, 10 and 10 mobility traces. Therefore, we have $p_A = 50\%$, $p_B = 25\%$, $p_C = 12.5\%$ and $p_D = 12.5\%$. Applying Equation 1 leads to an average entropy of 1.27 bits for Alice as illustrated in Table III.

POI	Nb of mobility traces	Probability	Entropy
A	40	0.5	-0.51
B	20	0.25	-0.11
C	10	0.125	-0.24
D	10	0.125	-0.41
Total	80	1	1.27

Table II
EXAMPLE OF THE COMPUTATION OF ALICE'S ENTROPY.

- The *predictability* [3] is a theoretical measure quantifying how predictable is an individual based on his MMC model (*cf.* Section V). For instance, consider the scenario in which Bob is currently located on the "Home" POI then based on his MMC, the probability of making a successful guess of his next location is theoretically equal to the maximal outgoing probability transition leaving from this state (*i.e.*, POI), which could be for instance the transition going to the "Work" POI. More formally, the predictability $Pred$ of a particular user u is computed by using the following formula:

$$Pred(u) = \sum_{i=1}^{n_u} (\pi_{i,u} \times p_{max\_out}(i, u)), \qquad (2)$$

which corresponds to the sum of the product between each element i of the stationary vector $\pi_{i,u}$ computed from the MMC of user u, in which $\pi_{i,u}$ is the probability of being in a particular state (for n_u, the total number of states of the MMC of user u) and $p_{max\_out}(i, u)$ represents the maximum outgoing probability leaving from the i^{th} state.

Based on these three mobility characteristics (*i.e.* entropy, predictability and average traveled distance per day), we have analyzed the dataset in order to observe their spatial distributions (Figure 5). First, with respect to the distance, we can observe that most of the individuals do not travel more than 250 kilometers per day on average. Second, we looking at predictability we can see clearly that a threshold of 35% seems to divide the population into two groups. Finally, the distribution of the entropy seems to be spread rather uniformly between 0 and 6 bits.

Based on these observations as well as on the description of the different modes of transportation (Section III), we propose to partition the users contained in the dataset into the following categories whose characteristics are summarized in Table III. We recognize that this proposition of classification is rather heuristically for now and maybe

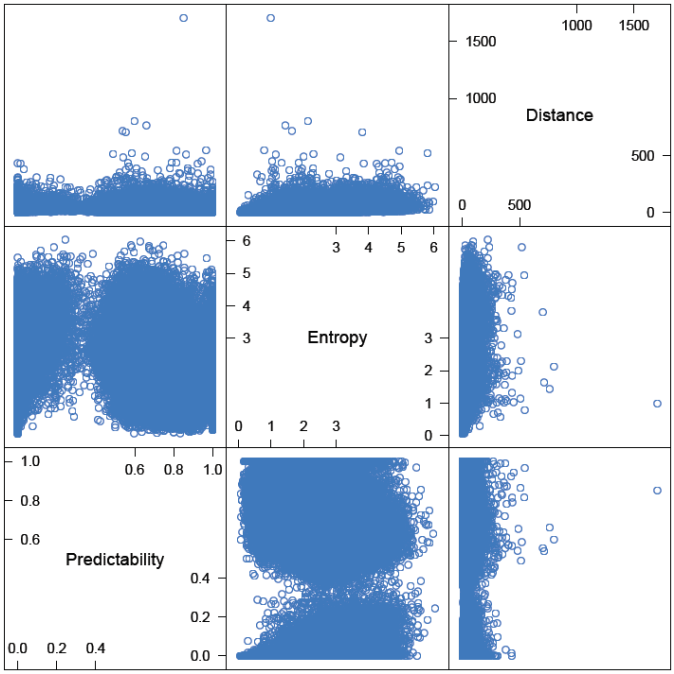

Figure 5. Comparison between the entropy, predictability and average traveled distance per day for the users of the dataset that have at least 40 mobility traces in their history and have visited at least 5 different antennas.

suggest to discussion, thus we consider it only as a first step towards discovering more sophisticated groups of users in the population sharing the same mobility patterns and characteristics. We leave the discovery and characterization of these more refined categories as future work.

Category	Predictability	Entropy	Distance
Individual	High $(0,35 \leq p \leq 1)$	Weak $(e < 3)$	Weak $(0 \leq d < 50)$
Gbaka	High $(0,35 \leq p \leq 1)$	Weak $(e < 3)$	Average $(50 \leq d \leq 150)$
Woro-woro	Weak $(0 \leq p < 0,35)$	Weak $(e < 3)$	Average $(50 \leq d \leq 150)$
Bush taxi	High $(0,35 \leq p \leq 1)$	High $(3 < e)$	High $(150 < d)$
Traditional taxi	Weak $(0 \leq p < 0,35)$	High $(3 < e)$	Average $(50 \leq d \leq 150)$

Table III
SUMMARY OF THE CHARACTERISTICS DEFINING EACH CATEGORY OF TRANSPORTATION MODE.

When analyzing the dataset with this characterization, we obtain respectively 77 674 individuals, 1 977 gbakas, 4 861 woro-woros, 8 565 bush taxis and 2 842 traditional taxis.

VII. RECOMMENDER SYSTEM FOR BUSH TAXIS

In this section, we propose a recommender system suggesting to individuals a transportation mode among the following set $TM = \{bush\_taxi, woro\_woro, gbaka, traditional\_taxi\}$, which corresponds to a vehicle corresponding to this mode

of transportation whose position is spatiotemporally close to the one of individual considered and whose destination matches the next location of an individual. Consider for instance, the scenario in which Alice is currently at "Home" at 8AM and she wants to go to the POI corresponding to "Work". She starts the application corresponding to the recommender system on her smartphone in order to retrieve a transport passing in her neighborhood (*e.g.*, at most 1 kilometer for her current position) in the next 30 minutes and whose destination matches her own destination.

More precisely, the recommendation algorithm works in two phases. During the first phase, the recommender system takes as input the location l of the individual and looks for other users belonging to the one of the categories of the set TM that are spatiotemporally close to l. Afterwards, during the second phase, the recommender system predicts $next\_loc\_user$, the next place that will be visited by the user, as well as $next\_loc\_transport$ the next destinations of the neighboring transports in order to find a potential match (*i.e.*, $\mathsf{Dist}(next\_loc\_user, next\_loc\_transport) < \delta$, for δ a predefined distance threshold). In order to minimize the cost of the travel for the user, the system tries to find first a match from the "bush taxi" category. If at least one match is found in this category, then the recommender system returns the match from this category whose destination is the closest to the destination of the user. If no match is found, then the recommender system looks for a match in the following categories until one is found: "woro-woro", then "gbaka" and finally "traditional taxi".

We consider the three following strategies in order to compute the spatial closeness between individuals.

1) The first strategy, which is based on the *current position* of the user, searches for transports located from distance d from the current position l of the individual in the next 30 minutes.
2) The second strategy based on the *trajectory* takes as input the actual position l of the individual and computes the minimal Euclidean distance d from the current position to the path obtained by joining the point of origin and the point of destination of the transport.
3) The third strategy, which is called the *anywhere* strategy, models the situation in which a vehicle is close to the user position (*e.g.*, a taxi picks up a passenger and drives him directly to his next location).

Based on these strategies, we have designed different methods for picking up and dropping off passengers, modeling the difference between the transports that have a fixed or a flexible route. We have applied these strategies to the different transportation means existing in Ivory Coast as summarized in Table IV. For the pick up method, we rely on the current position strategy for all transportation modes. However, for the drop off method, the trajectory strategy is used for bush taxis and gbakas while the current position strategy is applied for the woro-woros and the anywhere strategy is used for traditional taxis.

Transportation	Pick up method	Drop off method
Bush taxi	Current position	Trajectory
Woro-woro	Current position	Current position
Gbaka	Current position	Trajectory
Traditional taxi	Current position	Anywhere

Table IV
PICK UP AND DROP OFF METHODS DEPENDING ON THE TRANSPORT MODE CONSIDERED.

The recommendation algorithm takes as input P, a set of pedestrians (*i.e.*, individuals) and V a set of potential vehicles belonging to the four categories of transportation mode $TM = \{bush\_taxi, woro\_woro, gbaka, traditional\_taxi\}$. For each pedestrian $p \in P$, the algorithm predicts $next\_loc\_user$, the next location visited by the user. Then, the recommendation algorithm looks for possible transports that are spatiotemporally closed to p based on the current position strategy mentioned previously starting from the "bush taxi" category. Once all the potential vehicles have been identified in this category, the next destination $next\_loc\_transport$ is predicted for each vehicle of this category that are spatiotemporally close to the user until a match is found (*i.e.*, $\mathsf{Dist}(next\_loc\_user, next\_loc\_transport) < \delta$). If a match is found, then the recommendation algorithms suggests it to the pedestrian, otherwise this process is then repeated for each transportation mode continuing with the woro-woro, gbaka and then finally the traditional taxi until a match is found. More precisely, the goal is to recommend a transportation mean to a user while at the same time minimizing the travel cost, which explained why we use this particular order. If the recommender system is not able to suggest a transport to a given pedestrian, then it is considered to have failed to provide a recommendation for this particular input.

On order to predict the next location visited by a user (pedestrian or transport), the recommendation algorithm relies on the MMC learnt from the mobility traces of this user. More precisely, the algorithm finds the state of the MMC corresponding to the actual position of the user and then predicts as the next location visited by the user the POI corresponding to maximal outgoing probability leaving from the current state (*i.e.*, ties are broken arbitrarily). In the next section, we will evaluate the practical performance of the recommendation algorithm on the D4D dataset.

VIII. EMPIRICAL EVALUATION

In this section, we detailed the preliminary empirical evaluation that we have conducted to assess the efficiency of the recommender system for bush taxis. Note that contrary

to most of the previous works detailed in Section II, our goal is not to maximize the occupancy of taxis or to minimize the waiting time for passengers. Rather, our main objective is to provide each pedestrian with the recommendation about a transport going in the same direction as his next destination while at the same time selecting the cheapest transportation mode among the following categories: bush taxi, woro-woro, gbaka and traditional taxi. One of the main difficulty compared to previous work is that we do not have detailed GPS traces for the vehicles, instead we only have access to the sporadic exposure of their location data. In order to evaluate, the accuracy of the predictions made by MMCs for next location as well as the recommendation algorithm, we designed two metrics, the *prediction accuracy* as well as the *coverage*, that we described thereafter. These metrics have been assessed on the categories of users introduced in Section IV.

The *prediction accuracy* quantifies the capacity of a mobility model (in our case a MMC) to predict the next location visited by a user based on its current location. In our experiments, we have considered only users that have at least 40 mobility traces and that have visited at least 5 antennas. For each user, we split the trail of mobility traces into two sets of same size: the *training set*, which is used to build the MMC, and the *testing set*, which is used to evaluate its accuracy in predicting the next location visited by a user. Contrary to the predictability (*cf.* Section VI), which is a theoretical measure, the prediction accuracy of a MMC is computed on the testing set, which is completely disjoint from the training set on which the MMC is learnt.

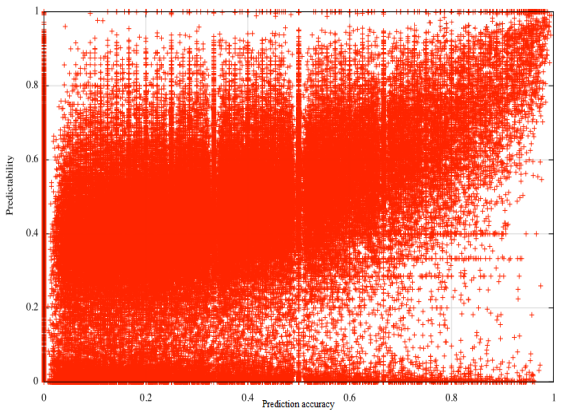

Figure 6. Comparison of the prediction accuracy versus the predictability for the MMCs.

Figure 7 depicts the correlation between the (theoretical) predictability and the (empirical) prediction accuracy. The average predictability is approximately of 50% while the average prediction accuracy is of 30% with an average absolute difference between the two of 28% and a variance of 6.5%. These results are explained by the fact that mobility GSM traces are coarser and less regular than GPS traces. Comparing to previous work done by the authors [3] the prediction accuracy is less than when a MMC is trained on richer traces (*e.g.*, mobility traces obtained from GPS-enabled devices). We observe that some users might display a poor predictability but a high prediction accuracy, we conjecture that this is due to the fact that these users have few antennas in their test set. In contrast, another group of users displays a high predictability but a poor prediction accuracy, due to the presence in the test set of antennas (*i.e.*, states) that the MMC has never seen in the training set when it was learnt.

In a nutshell, the *coverage* measures the percentage of users that have been served by the recommender system. More precisely, considering a set composed of n individuals, we consider for each user the mobility traces of their testing set and we try for each of these traces to suggest a potential transport for the user based on the recommendation algorithm described in Section VII. If a potential match is found then we consider that the system has succeeding in proposing a recommendation to the user, while otherwise we consider that it has failed. The coverage is then averaged over all the pedestrians and all the mobility traces of their test set. To evaluate the coverage, we have used 10% of the users coming from the "individual" category sampled at random (roughly 8000 users), which we believe correspond mainly to pedestrians. The spatiotemporal closeness considered for the recommendation process is respectively of 30 minutes for the time and 1 kilometer for the distance. Recall that the recommendation algorithm looks first for a match in the bush taxi category, before going through the gbaka, then the woro-woro and finally the traditional taxi categories. Overall, we observe a high coverage of 99% for the recommender system, which is decomposed respectively over 72% for bush taxis, 15% for gbakas, 1% for the woro-woros and finally 11% for the traditional taxis. Therefore, we observe that almost 3 times out of 4 the recommender system is able to provide a recommendation that corresponds to the cheapest mode of transportation. Of course, the results we have obtained are only preliminary and we plan to validate them by running the experiments on the complete set of users who have more than 40 mobility traces on their history and that have visited a least 5 different antennas.

IX. Conclusion

In this paper, we have introduced a recommender system for bush taxis in Ivory Coast whose main objective is to propose to pedestrians potential means of transportation in their neighborhood whose destination match their own destination by relying on a mobility model called Mobility Markov Chain. One of the main advantage of our recommender system is that it is fully automatic, in the sense that a user does not need to explicitly express his next destination but

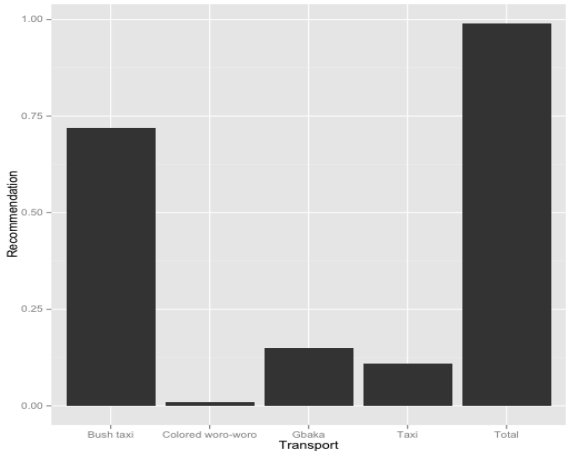

Figure 7. Distribution of the coverage over the different possible modes of transport.

rather the system tries to infer it based on his past mobility behavior. Moreover, the recommendation algorithm is biased towards suggesting the means of transportation that are the cheapest (if one is available). The preliminary evaluation of the recommender system shows that approximately 99% of the time in less than 30 minutes, the system is able to suggest a mean of transportation that is at most 1 kilometer away from the current position of the user for an accuracy of the prediction of the next location that is between 30% and 50% depending on the mobility category to which this user belongs.

As future works, we plan to validate the results of the preliminary analysis on the complete part of the D4D dataset composed of users that have at least 40 mobility traces in their history and that have visited a least 5 different antennas. We also want to evaluate the possible trade-off between the accuracy of the prediction and the coverage of the recommendation algorithm, possibly by combining them into a global metric *à la* F-measure quantifying the success of the recommender system. With respect to the mobility model considered for the prediction of the next location, we also want to apply more sophisticated variants of MMCs that incorporate the notion of time slices (*i.e.*, they can predict the next location depending on the current period of the day) or that can remember the k last visited locations by a user. Moreover, we also want to conduct an evaluation measuring the economical impact on the average cost of travel for a user of using different recommendation strategies. Finally, we acknowledge that the current idea of a recommender system for bush taxis is still very crude and we are currently working on refining it by making it more adaptive and flexible to the current context.

REFERENCES

[1] Daniel Ashbrook and Thad Starner. Learning significant locations and predicting user movement with GPS. In *Proceedings of the International Symposium on Wearable Computers*, volume 6, pages 101–108, Sardina, Italy, February 2002.

[2] Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *Computing Research Repository*, 1210(137):1–10, September 2012.

[3] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility Markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, volume 3, pages 1–6, Bern, Switzerland, April 2012.

[4] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show me how you move and I will tell you who you are. In *Transactions on Data Privacy*, volume 4, pages 103–126, August 2011.

[5] Yong Ge, Chuanren Liu, Hui Xiong, and Jian Chen. A taxi business intelligence system. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 735–738, New York, NY, USA, August 2011.

[6] Yong Ge, Hui Xiong, Alexander Tuzhilin, Keli Xiao, Marco Gruteser, and Michael Pazzani. An energy-efficient mobile recommender system. In *Proceedings of the 16th ACM international conference on Knowledge discovery and data mining*, volume 10, pages 899–908, New York, NY, USA, July 2010.

[7] Junghoon Lee, Gyung-Leen Park, Hanil Kim, Young-Kyu Yang, PanKoo Kim, and Sang-Wook Kim. A telematics service system based on the linux cluster. In *International Conference on Computational Science*, pages 660–667, Krakow, Poland, June 2007.

[8] Junghoon Lee, Inhye Shin, and Gyung-Leen Park. Analysis of the passenger pick-up pattern for taxi location recommendation. In *Proceedings of the Fourth International Conference on Networked Computing and Advanced Information Management*, pages 199–204, Washington, DC, USA, September 2008.

[9] James MacQueen et al. Some methods for classification and analysis of multivariate mbservations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, CA, USA, July 1967.

[10] UITP International Association of Public Transport. Aperçu du transport public en afrique subsaharienne. http://www.uitp.org/knowledge/pdf/transafrica_fr.pdf, 2009.

[11] UITP International Association of Public Transport. Public transport in sub-saharan africa, major trends and cadse study. http://www.uitp.org/knowledge/pdf/PTinSSAfr-Majortrendsandcasestudies.pdf, 2010.

[12] Orange. D4d challenge dataset. http://www.d4d.orange.com/learn-more.

[13] Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 23(3):379–423, October 1948.

[14] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 259–270, Uppsala, Sweden, March 2011.

[15] Jing Yuan, Yu Zheng, Liuhang Zhang, XIng Xie, and Guangzhong Sun. Where to find my next passenger. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 109–118, Beijing, China, September 2011.

[16] Nicholas Jing Yuan, Yu Zheng, Liuhang Zhang, and Xing Xie. T-finder: A recommender system for finding passengers and vacant taxis. *IEEE Transactions on Knowledge and Data Engineering*, 99:1–14, August 2012.

[17] Xudong Zheng, Xiao Liang, and Ke Xu. Where to wait for a taxi? In *Proceedings of the International Workshop on Urban Computing*, pages 149–156, New York, NY, USA, August 2012.