Toward a Route Detection Method base on Detail Call Records

Documento de Discusión
CIUP

DD1619

2016

Miguel Núñez del Prado
Profesor e investigador del CIUP
m.nunezdelpradoc@up.edu.pe

Hadrien Hendrikx
École polytechnique

UNIVERSIDAD DEL PACÍFICO
CENTRO DE INVESTIGACIÓN

Toward a Route Detection Method base on Detail Call Records

Hadrien Hendrikx
École polytechnique
Palaiseau-France
Email:hadrien.hendrikx@polytechnique.edu

Miguel Núñez-del-Prado-Cortez
Universidad del Pacífico
Avenida Salaverry 2020, Lima-Perú
Email: m.nunezdelpradoc@up.edu.pe

***Abstract*—In the last years, smartphones have become the major device for communication enabling Telco operators to capture subscribers' whereabouts. This location information allows computing geostatistics to study transportation systems, traffic jams, origin-destination matrix, etc. The first task to accomplish the aforementioned objectives is to detect routes that people use to go from A to B. Thus, in the present effort, we propose a method to extract automatically routes from CDR data relying on clustering and community detection algorithms.**

I. Introduction

Ubiquitous systems such as smartphones allow Telco operators to sense subscribers location opening new opportunities to analyze urban phenomena. This location information allows computing geostatistics to study transportation systems, traffic jams, origin-destination matrix, etc. The first task to accomplish the aforementioned objectives is to detect routes. Previous works in the literature do not have the objective to discover routes. They aim to infer transportation mode, to model movement behavior or to predict future locations. Thus, route detection is an intermediary step, which is the cornerstone for other tasks. Consequently, in the present effort, we propose a methodology to detect routes automatically. We rely on the *Vector Field K-Means* clustering and the *Infomap* community detection algorithms. Our approach receives as input CDR data to extract the set of antennas describing or covering a route of interest. Finally, the present work is organized as follows: Section II detail the related works in the state of the art, while Section III introduces some basic concepts. Then, sections IV and V describe the datasets used in this work and the methodology to extract routes, respectively. Finally, Section VI presents the results of our experiments and Section VII concludes the work and comments future research direction.

978-1-5090-1633-4/16/$31.00

II. Related Works

In the current section, we present some previous works on route detection using Global Position System (GPS) and Global System for Mobile Communications (GSM) data.

The former kind of data is a more reliable source, since it is both spatially accurate (fine-grained) and frequently collected , which allows a better resolution and precision for different inferences, such as route detection, home/work location, whereabouts prediction, *etc*. The drawback of this source is the scalability. Thus, massive collection is difficult due to GPS is battery consuming and users do not always turn it on. Nonetheless, there are some works like the study of Zheng *et al.* [2], where the authors use different models to infer transportation mode and thus routes. To perform this task, authors pre-process data by detecting walking and non-waking segments in trajectories. Then, they extract features such as speed, traveled distance, comparison to urban network topography and speed acceleration from the non-walking segments of different trajectories. Once features are extracted, these are the input of three different algorithms including Decision Tree, Bayesian Net, Support Vector Machine (SVM), which classify the different segments into bus, car and bike classes. Since segments are labeled, a Conditional Random Field (CRF) model is used to model and to predict the change of transportation mode from bus to bike, for instance. For the experiments,

authors gathered GPS (timestamp, latitude and longitude) data from 45 users covering 15 different cities in order to use 70% for training and 30% for test. The results, in term of accuracy, precision and recall show that Decision Tree (0.721, 0.867 and 0.197) outperform over Bayesian Net (0.574, 0.867 and 0.206), SVM (0.517, 0.578 and 0.095) and CRF (0.422, 0.115, 0.072).

Another work using GPS data was done by Liao *et al.* [10]. In this research authors rely on a 3 level hierarchical Markov chain to model movement behavior. Where the highest level indicates either new places or anomalous behavior. The second level models the trip segments of trajectories (motivated by a goal) between point of interests (POI) and the lowest level estimates user whereabouts to depict routes. More precisely, authors use the Rao Blackwellized particle filters algorithm to estimate people location as well as routines in the lowest level. In the second layer, the set of locations (a segment of trajectory) is enriched with the transportation mode and using the Expectation Maximization (EM) algorithm, authors compute the transition probabilities between different transportation modes. Finally, based on those transitions they are able to monitor anomalous behavior or people visiting new places. To validate the model, authors collected 60 days (the first 30 days for learning and the second 30 days for testing) of GPS data from one person. The aforementioned model obtained an accuracy score between 0.66 and 0.98 depending on the amount of the observations over time.

The latter kind of data (GSM) is more sparse but with a constant frequency update rate. Thus, Laasonen *et al.* [7] build clusters containing a sequence of cell Ids to model physical routes. In detail, authors take a sequence (p) of antennas Ids relaying two POIs *A* and *B*. This sequence of antennas are considered as strings. Then, the clustering algorithm takes p and a Jaccard based measure as similarity function. The algorithm compares p to all known trajectories (*i.e.*, groups). If the similarity measure is small then the clustering algorithm merges p to complete the know trajectory relaying *A* and *B*; otherwise a new group is created. This process is part of a system for predicting future whereabouts. Consequently, authors did not evaluate the algorithm for detecting routes. There is also the work of Eagle *et al* [9] where the authors extract frequent trajectory patterns from GSM events.The authors consider trajectories as a set of antennas Ids. It is possible to see the set of frequent trajectories as representation of routes. However, the works on the state of the art do not evaluate directly the route extraction. From the state of the art, it is possible to observe the lack of methods applied to GSM generated data due to the spatio-temporal sparseness of the generated data. Therefore, in the next sections we explain our method to resolve this problem.

III. Basic Concepts

In the present section, we will introduce two key concepts for our methodology: clustering and community detection. The former algorithm groups trajectories (*c.f.,* Subsection III-A). While the latter discovers frequent antennas Ids, representing routes, in the clustered trajectories (*c.f.,* Subsection III-B).

A. *Vector-Field K-Means (*VFKM*)*

The clustering algorithm works in two phases [1]. The first phase computes the characteristic vector fields of a set of trajectories like the computation of the centroid in classic k-means (*c.f.,* algorithm in Figure 1).

$$E''(\alpha_i, X_j) = \int_{t_0}^{t_1} ||X_j(\alpha_i) - \alpha_i')||^2 dt \qquad (1)$$

Equation 1 describes the similarity function used in the second phase to measure the minimal distance between a trajectory α_i and the characteristic vector fields x_j. Then trajectory is added to the closest group X_j as shown in algorithm in Figure 1. The algorithm repeats these phases until no trajectory could be assigned to a different group *i.e.,* convergence.

B. *Community detection*

A community is a group of vertices in a graph, which are densely linked one to each other and sparsely connected to other communities. We will use community detection algorithm to outline some connected sets of cells representing routes. In the current effort we use the *Infomap algorithm* [5]. The idea behind this algorithm is to reduce the Map equation [6] by computing the fraction

Require: k: number of clusters, $\alpha = \{\alpha_1, \ldots, \alpha_n\}$: set of trajectories
Ensure: $V = \{X_1, \ldots, X_k\}$: Group of trajectories, $CVF = \{cvf_1, \ldots cvf_k\}$: Characteristic vector fields
$CVF \leftarrow Initialize(T, k)$
repeat
//Choose k trajectories randomly
CVF $\leftarrow$ Initialize(V,k)
//Fit characteristic into the vector field
for i=1 to k **do**
$X_i \leftarrow fitVectorField(cvf_1)$
end for
//Processes trajectories
for i=1 to n **do**
$v \leftarrow argmin_{j \in 1,\ldots,k}\ E'(\alpha_i, X_j)$
$X_j.add(v)$
end for
until convergence
return V

Fig. 1. Vector-Field K-Means algorithm.

TABLE I
DATASETS SUMMARY

Attribute	Subscribers	Events	days
Small level	17 500	1 700 000	1
Medium level	48 000	69 000 000	15

of times a random walker visits a node. Based on those visits, a Greedy search algorithm is used to find a partition [16]. Once the partition is found, the results is refined using a simulated Annealing approach [15].

IV. DATASET DESCRIPTION

The present section introduces the two datasets we use for our experiments. The Call Detail Record (*CDR*) was provided by a Telco operator in 2014. All events were gathered within in Paris.

a) Small dataset. This dataset contains the country level CDR of mobile phones users in of a weekday day, which represents about 1.7 millions events of 17 500 subscribers.

b) Medium dataset. The second dataset has around 69 millions of events of 48 000 subscribers that registered at least one in Paris. This data set was collected for 15 days.

TABLE II
DATASETS SUMMARY

timestamp	msisdn	imsi	mcc	mnc
13701045	915463	208103	1326403	210

lac	ci	latitude	longitude
11	48000	48.8534100000	2.3488000000

In both cases events are recorded each time a user cross a set of cells called *IRIS*. The datasets are summarized in Table I. They have events containing a *timestamp, msisdn* and *imsi* corresponding to subscriber identity and antenna identifiers (*mcc,mnc,lac,ci*) in addition of antennas location (*i.e., latitude and longitude*) as illustrated in Table II. It is worth noting that datasets are anonymized for preserving subscriber privacy.

V. METHODOLOGY

The dataset described in Section IV contains a huge amount of low quality events due to the sparseness and incompleteness of the *CDR*. Indeed, there is no a constant *sampling rate* for collecting events in the Telco network and there are spatio-temporal gaps in the data. These factors makes complicate to extract reliable and relevant trajectory. Thus, our methodology is composed of three steps: pre-process, trajectory patterns extraction and routes detection.

a) Pre-process. The first step of our methodology is to delimit the zone where we want to detect routes by placing a ***bounding box*** (*BB*). Then, the algorithm works only with all events within the box. Once events within the *BB* are kept and order chronologically, we have to deal with the *Ping-Pong* effect. This phenomenon occurs when a subscriber is between two or more antennas and his cell phone connect to them in short times simulating movements from one antenna to another randomly. To deal with this problem, we rely on the filter proposed by Lovan *et al.* [12]. Thus, the ***Ping-Pong filter*** discards all events with a speed higher than V_{max} or all events describing and angle between $180 \pm \theta_{max}$ We found empirically a suitable threshold of $45Km/h$ and $20°$ for V_{max} and θ_{max}, respectively.

Since events are less noisy after the pre-process phase, we apply an algorithm to **extract trajectories**. The idea behind this algorithm is that stops

mark off subscribers' trajectories. Consequently, the algorithm takes as input a *time threshold* Δt to verify if a subscriber has left a given antenna after Δt to end a trajectory and to begin a new one (*i.e.,* $t_{out} - t_{in} < \Delta t$). To refine trajectories for the *VFKM* clustering algorithm, we discard all trajectories containing few sequence of antennas (*i.e.,* small trajectories). Finally, we compute the average speed of the remaining trajectories.

b) Trajectory patterns extraction. We use *VFKM* to find global trajectory patterns directions such as north to south. The *VFKM* algorithm takes as input: the number of k clusters, the weight of the eigenvectors, λ_L and the resolution *R* of the algorithm. Then, we perform two consecutive clustering. The first one is executed over all the trajectories within the bounding box. The second one is carried out over the trajectories belonging to the first group of the previous clustering phase. Consequently, at the end of the second *VFKM* clustering algorithm, we have a set of vectors representing movements within the bounding box.

c) Routes detection. After the clustering phase, trajectories representing routes are still coarse. In order to obtain a fine representation of the routes, we build a transition matrix T, using the obtained vectors from the present phase, where $T_{i,j}$ is equal to the number of trajectories that have a segment going from antenna i to antenna j. Once T is built, we can weight the values in T by the number of trajectories, the number of users or the logarithm of the number of unique users passing from i to j. Then, to reduce the size of the matrix and computation time, we filter all antennas having a transition value less than an activity threshold $minActivity$. Finally, we use T to build a graph representing flows from one antenna to another for applying the *Infomap* community detection algorithm. As a result of this process, we are able to extract the set of trajectories describing different routes.

VI. Results

In the present section, we apply the methodology introduced in Section V to the datasets in Section IV. More precisely, we fine tune the parameters for the *VFKM* algorithm and test the proposed methodology.

In order to find the most suitable parameters, $k, R,$ and λ_L for the *VFKM* algorithm. We use the *small dataset* introduced in Section IV. First, we fix the λ_L to 0.05 as indicated in [1]. Next, we test different configurations varying k from 2 to 7, where k is the number of routes we are interest in and R from 4 to 10. Thus, using visualization of the trajectories over a cartography, we observed that the best parameters for our datasets are $k = 4$ and $R = 6$. Consequently, we use this configuration for the rest of the experiments.

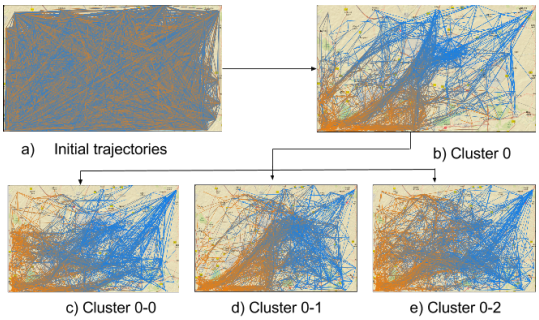

Fig. 2. *VFKM* iterative clustering process

Regarding the methodology, Figure 2a shows the raw trajectories after the pre-processing, which is the input of the *VFKM* algorithm. Each trajectory is represented by a line, and the color fades from blue at the beginning to orange at the end. This allows observing whether trajectories follow global trends. We note the initial dataset containing trajectories without a common patterns. They are among both the main routes as well as the residential areas near to the airport. In order to extract patterns, we apply *VFMK* clustering algorithm, with the parameters $k = 4$, $R = 6$ and $\lambda_L = 0.05$, obtaining four cluster $k = \{c_0, c_1, c_2, c_3\}$ with 6 914, 63 509, 9 567 and 352 243 trajectories, respectively. Thus, when examining clusters in detail, we discover that c_0, the smallest cluster represents smooth and homogeneous vector fields from the north-east to the south-east trend within the bounding box (*c.f.,* Figure 2b). On the contrary, the other groups c_1, c_2 and c_3 do not present clear patterns. We can also observe that c_3 contains 80% of the dataset.

Once the first clustering is done, we take the group with the smallest number of trajectories C_0 to apply the *VFMK* clustering algo-

rithm. Accordingly, we get four clusters $k = \{c_{0-0}, c_{0-1}, c_{0-2}, c_{0-3}\}$ and we show three in figures 2c, 2d and 2e, respectively. Cluster c_{0-3} represent the group of noisy trajectories. In the cluster c_{0-0}, trajectories follow some routes, such as the train line *RER D*, the *A1* highway and the train line *RER B*. Cluster c_{0-2} is quite similar to cluster c_{0-0}. However, c_{0-2} is noisier and it contains the commuting patterns between the west zone and the airport. Further, sub-clustering has been done in order to identify more precisely different routes. Nevertheless, further clustering does not provide significant information. Cluster c_{0-1} depicts the *RN104* route, *A1* highway and the train line *RER B* as well as commuting patterns between the east zone and the airport.

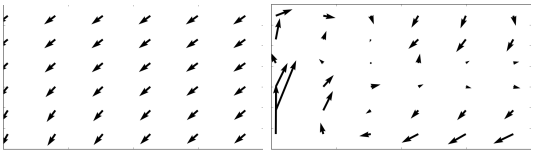

Fig. 3. Vector field corresponding to clusters c_{0-0} and c_{0-3}, respectively

Since the group of trajectories from the *VFKM* clusters are coarse, we need to refine them by applying the *Infomap* community detection algorithm. The question about the cluster to be passed as input is raised. Consequently, as shown in Figure 3, we choose the group of vector fields minimizing the entropy. In our case the most homogeneous vector field corresponds to cluster c_{0-0}. Thus, Figure 4 depict the remaining antennas to detect routes. It is worth noting that the activity of antennas (*i.e.,* the number of different cell phone attach to a given antenna) changes over time. Hence, we are able to filter low activity antennas to keep only high activity antennas, which should represent main roads. The different colors in Figure 3 represent the antenna activity, where blue, green and red correspond to low, medium and high antenna activity, respectively.

Taking into account only high activity antennas, we build a transition matrix T. This transition matrix, contains the number of vector fields from antenna i to antenna j. Relying on the T matrix, we build a graph G, where nodes are the antennas, weighted by the number of different cell phone

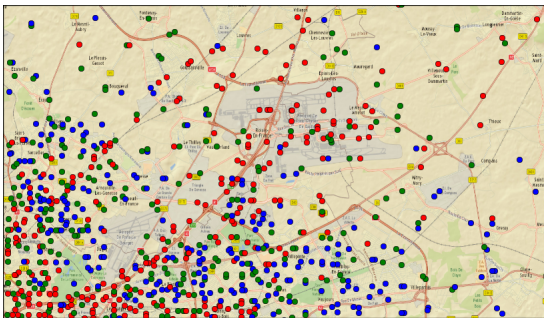

Fig. 4. Antenna activity corresponding to cluster c_{0-0}.

attach to the antenna i and the edges are the number of flows from antenna i to j.

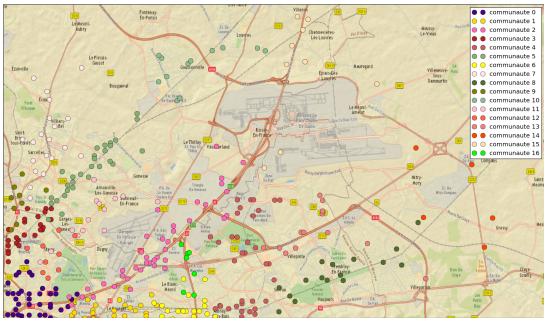

Fig. 5. Result of the *Infomap* community detection algorithm over cluster c_{0-0}.

Accordingly, the graph G is the input of the *Infomap* community detection algorithm. Figure 5 illustrates the result of the community detection algorithm to find more fine routes. The algorithm found 15 communities $C = \{c_0, \ldots, c_{14}\}$, which represent different routes.

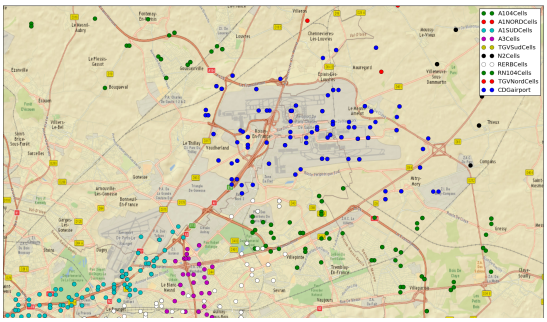

Fig. 6. Ground truth of set of antennas covering routes.

TABLE III
CORRESPONDENCE BETWEEN REAL ROUTES AND DETECT COMMUNITIES.

Routes	Communities
A104	c9 (0.16), c11 (0.05)
A1SUD	c0 (0.29), c1 (0.27)
A3	c6 (0.17), c14 (0.17)
N2	c10 (0.46)
RN104	c5 (0.17)
RER B	c2 (0.14), c3 (0.19)
Others	c4,c7,c8,c12c,13

To evaluate the performance of our approach, we rely on *Jaccard Similarity* $J = A \cap B / A \cup B$, where A and B are a set of antennas represented by either trajectories or communities. We use this metric to measure the similarity between detected communities and ground truth extract manually from the Telco antennas database (*c.cf.,* Figure 6). We obtain different values for the communities ranging from 0.05 to 0.46, as presented in Table III. We observe that computed communities represent routes. The limitation of the *Infomap* algorithm is the resolution. Since we are not able to stop the algorithm to obtain the optimal resolution vis-a-vis the routes, we need to merge complementary communities, which are partial routes, to detect a complete route. For instance, *A104* route is composed of c9 and c11. Finally, with the presented method, the intervention of an expert is reduced to merge visually communities to complete the set of antennas representing routes.

VII. CONCLUSION

In the present work, we describe a methodology to extract sets of antennas covering a route of interest. More precisely, our methodology detects automatically important routes for geostatistics studies like origin-destination matrix construction, urban planning, etc. In this study, we have used *CDR* data from a Telco company to extract important routes connecting *Paris* and *Charles de Gaulle* airport obtaining a Jaccard Similarity score between detected routes and ground truth ranging from 0.165 to 0.46. In the future, we will analyze how to merge complementary communities automatically and test our methodology in an urban environment.

REFERENCES

[1] N. Ferreira, J.T. Klosowski, C.E. Scheidegger, C.T. Silva, *Vector Field k-Means: Clustering Trajectories by Fitting Multiple Vector Fields*. Eurographics Conference on Visualization (EuroVis) 2013

[2] Y. Zheng, L. Liu, L. Wang, X. Wie, *Learning Transportation Mode from Raw GPS Data for Geographic applications on the Web*, Proceedings of the 17th International World Wide Web Conference (WWW 2008), pp. 247-254, Beijing, China

[3] C.Kang, S. Gao, X. Lin, Y. Xiao, Y. Yuan, Y. Liu, X. Ma, *Analyzing and Geo-visualizing Individual Human Mobility Patterns Using Mobile Call Records*

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment 2008, P10008 (2008)

[5] M. Rosvall and C. T. Bergstrom, *Maps of information flow reveal community structure in complex networks*, PNAS 105, 1118 (2008). http://dx.doi.org/10.1073/pnas.0706851105, http://arxiv.org/abs/0707.0609.

[6] M. Rosvall, D. Axelsson, and C. T. Bergstrom, *The map equation*, Eur. Phys. J. Special Topics 178, 13 (2009). http://dx.doi.org/10.1140/epjst/e2010-01179-1, http://arxiv.org/abs/0906.1405.

[7] K. Laasonen, *Clustering and prediction of mobile user routes from cellular data*, In PKDD, pages 569576, 2005.

[8] K. Laasonen, *Route prediction from cellular data*, In CAPS, pages 147158, 2005

[9] N. Eagle, M. Ali Bayir, M. Demirbas, *Discovering Spatiotemporal Mobility Profiles of Cellphone Users*, World of Wireless, Mobile and Multimedia Networks and Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a - 2009

[10] Liao L. Patterson D.J., Fox D., Kautz H., 2007 *Learning and inferring transportation routines*, Artifical Intelligence (AIJ),171(5-6):311-331, 2007

[11] Nabavi Larijani A., Olteanu-Raimund A., Perret J., Brdif M., Ziemlicki C., 2014 *Investigating the mobile phone data to estimate the origin destination flow and analysis; case study : Paris region*, 4th International Symposium of Transport Simulation-ISTS'14, 1-4 June 2014, Corsica, France.

[12] C. Iovan, A.M Olteanu-Raimond, T. Couronn, Z. Smoreda, *Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies*

[13] Bahoken, F. and Olteanu-Raimond AM., 2013, *Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths: The effect of spatiotemporal filtering on flow measurement*, In proceedings of 23rd International Cartography Conference, 2013.

[14] Frias-Martinez V., Soguero C., Frias-Martinez E., *Estimation of Urban Commuting Patterns Using Cellphone Network Data*, ACM SIGKDD Workshop on Urban Computing, Beijing, China, 2012

[15] Guimera R., Amaral L., *Functional cartography of complex metabolic networks* Nature (7028) 433:895-900, 2005.

[16] Clauset A., Newman M., Moore C., *Finding community structure in very large networks* Physical Review E 70, 066111, 2004.