Reconstructing Uncertain Pedestrian Trajectories From Low-Sampling-Rate Observations

Ricardo Miguel Puma-Alvarez and Alneu de Andrade Lopes
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - Campus de São Carlos
13560-970 São Carlos, SP, Brazil
rpuma@usp.br, alneu@icmc.usp.br

Abstract

The ever-greater number of technologies providing location-based services has given rise to a deluge of trajectory data. However, most of these trajectories are low-sampling-rate and, consequently, many movement details are lost. Due to that, trajectory reconstruction techniques have been created to infer the missing movement details and reduce uncertainty. Nevertheless, most effort has been put into reconstructing vehicle trajectories. Therefore, we study the reconstruction of pedestrian trajectories by using road network information. We compare a simple technique that only uses road network information with a more complex technique that, besides the road network, uses historical trajectory data. Additionally, we use three different trajectory segmentation settings to analyze their influence over reconstruction. Our experiment results show that, with the limited pedestrian trajectory data available, a simple technique that does not use historical data performs considerably better than a more complex technique that does use it. Furthermore, our results also show that trajectories segmented in such a way as to allow a greater distance and time span between consecutive points obtain better reconstruction results in the majority of the cases, regardless of the technique used.

1 Introduction

Currently, there are many technologies providing location-based services. Some of them are the GPS (Global Position System), RFID (Radio Frequency Identification), smartphone sensors, ultrasonic and infrared systems, etc (Feng and Zhu, 2016). All these technologies allow a large-scale generation of trajectory data of moving objects, which can be used to perform several data mining tasks. Thus, this scenario paved the way for the rise of the trajectory data mining field. There are several trajectory data mining applications such as path discovery, location prediction, behavior analysis, urban services improvement, etc. However, there are still some important challenges to be addressed regarding storage, computation and trajectory data mining (Baraniuk, 2011). Due to storage and transmission issues, these trajectories are generally collected at low sampling rates, consequently, they have long time intervals between location updates. In general, these trajectories provide a very limited representation of the real paths. This type of trajectories are called uncertain trajectories (Zheng and Zhou, 2011).

Most often, trajectories can be tracked very accurately with GPS-embedded devices like smartphones or automotive navigation systems. Nevertheless, a recent study has demonstrated that, aiming at reducing energy consumption, the majority of taxis of big cities use sampling intervals of two minutes (Wei et al., 2012).

The high energy consumption of GPS impairs its use in smartphones for long periods of time. Furthermore, most social networks provide check-in services, which allow user location sharing. Thus, it is possible to create trajectories by sorting these check-ins chronologically. In a similar way, trajectories can be generated from geotagged photos in photo sharing sites like Flickr [1]. In spite of that, the location updates generated through these sites are low-sampling-rate.

Addressing this issue is very important for several different trajectory data mining applications.

[1] https://www.flickr.com/

For instance, trajectories generated from geo-tagged photos could be reconstructed and used in itinerary recommendations applications. Additionally, other tasks like indexing and querying processing efficiency can be affected (Zheng et al., 2012).

Motivated by this problem, many works on trajectory reconstruction have been published. Most of them use road network information through a graph whose nodes represent intersections and terminal points, and the edges depict road segments. On the other hand, there are also some works that do not take into account this kind of information (Wei et al., 2012). These works aim to reconstruct trajectories in rural areas where there is no road network, and, also, trajectories of animals and certain natural phenomena like hurricanes. However, here we are focused on pedestrian trajectory reconstruction in urban areas. An example of a method of reconstruction that uses road network information is Infertra (Banerjee et al., 2014). This technique, instead of predicting the most likely route, returns an edge-weighted graph that summarizes all probable routes. The trajectory reconstruction process employs Gibss sampling by learning a Network Mobility Model (NMM) from a database of historical trajectories. Other works that also use road information are Hunter (2013), Zheng (2012), Li (2015) and Chiang (2013), to cite a few.

Nevertheless, most works are focused on reconstructing vehicle trajectories. This is mainly due to the fact that some pedestrian routes comprise small alleys and trails that are so narrowed to be traversed by other transportation mode different from walk. Despite of that, free collaborative maps like OpenStreetMap[2] allow the addition of these type of routes exclusively traversed by pedestrians to the road network. This way, it would be possible to reconstruct pedestrian trajectories using road network information. Considering that, we aim to study the reconstruction of pedestrian trajectories using road network information. Consequently, we depict a framework to reconstruct pedestrian trajectories composed by three phases. Firstly, we segment trajectories by using three different settings in order to study their influence over the quality of the reconstruction. Secondly, we perform a map matching task on these segmented trajectories using a free tool, thereby generating a set of network-constrained trajectories. Thirdly, we apply two different trajectory reconstruction techniques on this new trajectory set. We compare these two techniques, one of them a simple technique that only takes into account the road network information, and a more complex one that besides the road network structure uses historical trajectory data. We show that, under limited data conditions, the simpler technique greatly outperforms the more complex technique in pedestrian trajectory reconstruction. Furthermore, our results also demonstrate that trajectories segmented in such a way as to allow a greater distance and time span between consecutive points obtain better reconstruction results in the majority of the cases, regardless of the technique used.

[2] https://www.openstreetmap.org

2 Reconstructing uncertain pedestrian trajectories

In this section, we describe the framework used to reconstruct pedestrians trajectories.

2.1 Trajectory Segmentation

GPS logs generally record people's movement for long periods, in which the person could make multiple trips. When the person stops for a relatively long time, this could indicate the end of a trajectory and the start of the next one. Therefore, in order to reflect the real pedestrian intention as well as possible, we segment the GPS logs into effective trajectories, which have specific source and destination stay points.

In the Table 1, we observe three different segmentation settings based on the degree of tolerance, which is based on two criteria, the time between one GPS point to the next one and the distance between them. A GPS point is the representation of a location update in terms of space and time by means of geographic coordinates (latitude and longitude) and a timestamp.

	Medium	**High**	**Low**
Distance (m)	300	600	150
Time (mins)	10	20	5

Table 1: Trajectory segmentation settings

The rationale behind this arrangement of values lies mainly in the medium tolerance setting. This setting states that if the distance between a location update and the next one is greater than the combined lengths of three average city blocks (100m (HARRIS et al., 2008; Yeang et al., 2000; of the German Aerospace Center et al., 2012)), the previous location update is considered as the end of a trajectory and the last one as the start of the next trajectory. Likewise, if the time between two location updates is more than 10 minutes, the first location update and its successor are considered as the end and the start, respectively, of two different and successive trajectories. The idea behind the period of 10 minutes is to assume that a pedestrian can make some small stops due to external factors such as a quick conversation with some unexpected acquaintance on the way or waiting for the traffic light to change to cross a street, which far from meaning a source or destination, are just trip interruptions. Thus, finally, the half and the double of the values of these time and distance thresholds are allocated to the low and high tolerance settings respectively.

2.2 Map matching

The second phase of this framework is the map matching process, which aims to transform our set of GPS trajectories into network-constrained trajectories by matching each GPS point to an edge of the road network of a certain city. As already mentioned, with the use of free collaborative maps, now, these edges can also represent small alleys and trails walked exclusively by pedestrians. Map matching is an important research topic and there are many works focused on it (Lou et al., 2009; Greenfeld, 2002; Yuan et al., 2010). Additionally, there are free tools available that perform map matching tasks as Graphhopper [3]. Using a correct map matching method to align GPS points onto the road segments is relevant because the GPS points do not reflect their true position due to the GPS measurement error. Finally, this way, we used the Graphhopper tool to create a set of network-constrained trajectories that are used by the reconstructing methods in the next phase.

2.3 Reconstruction

One of the best techniques of trajectory reconstruction using road network information is InferTra (Banerjee et al., 2014). This method outperforms other state-of-the-art techniques by a large margin. Infertra is composed by two phases. Firstly, this technique uses the historical network-constrained trajectories and a road network to create a generative model called Network Mobility Model (NMM), which is a weighted directed graph whose edge weights denote the probability of the corresponding road segment being traversed. Hence, NMM learns the mobility patterns in a road network from a database of historical trajectories. Secondly, given an uncertain trajectory (a trajectory with low-sample-rate location updates), NMM is used to generate a weighted subgraph that depicts the probabilities associated to each possible trajectory arising from the uncertain trajectory location updates. On the other hand, we also used the Shortest Path technique to establish contrast with InferTra. The Shortest Path is a much simpler technique compared to InferTra, so this comparison can reveal whether a simple or more complex approach performs better when it comes to reconstructing pedestrian trajectories using road network information.

[3] https://www.graphhopper.com

3 Experiments

We use two data sets to perform the reconstruction of pedestrian trajectories. In each data set, the three trajectory segmentation settings previously depicted are used, low, medium and high tolerance, in order to study their influence over the performance of the reconstruction. Finally, we compare the performance of Infertra (Banerjee et al., 2014) and Shortest Path for different sampling intervals.

3.1 Data sets

The data sets considered in our experiments are (i) RadrPlus and (ii) Geolife (Zheng et al., 2009, 2008, 2010). RardPlus is a location-based social network developed in the University of São Paulo, campus of São Carlos, Brazil. This social network has the unique characteristic of being focused on communities. Therefore, RadrPlus provides functionalities not just for individual users like traditional social networks, but for groups of users as well, within a geolocated environment. The RadrPlus data set comprises trajectories of a group of 15 users in a period of 9 months. These trajectories were recorded in different parts of the city of São Carlos, but mainly around the campus of the university and its surroundings. Additionally, RadrPlus data set trajectories were labeled with two transportation modes, car and walk. The second data set is provided by the Geolife project, a location-based social network developed by Microsoft Research Asia. The Geolife data set contains trajectories of 182 users in a period of over five years. These trajectories were recorded in 30 cities of China and some cities in USA and Europe; however, most trajectories were recorded in the city of Beijing, China. In addition to that, a group of 73 users labeled their trajectories with transportation modes like walk, bike, bus, car, train, airplane and others. In spite of the variety of trajectory data of both data sets, we only select for our experiments the trajectories made by pedestrians, *i.e,* trajectories labeled with walk as transportation mode.

3.2 Experimental Results

We present and discuss the results of using the InferTra and Shortest Path techniques in the reconstruction of pedestrian trajectories. Each technique was implemented in Java. We segment trajectories from RadrPlus and Geolife data sets by using the low, medium and high tolerance settings and for each resulting trajectory set, we use the Graphhopper tool to transform these trajectories into network-constrained trajectories. Finally, we apply the aforementioned reconstruction techniques on the six different data sets generated as showed in Table 2. The main characteristics depicted in that table are the Number of Trajectories (NT), Average Length (AL), which is the average number of points per trajectory (pp), and the Average Duration (AD) of the trajectories.

	NT	**AL (pp)**	**AD (secs)**
RadrPlus (LT)	76	14.12	487.48
RadrPlus (MT)	116	17.31	604.15
RadrPlus (HT)	155	21.06	1045.61
Geolife (LT)	3895	13.06	721.74
Geolife (MT)	3601	13.71	819.57
Geolife (HT)	3134	13.79	877.2

Table 2: Main characteristics of each resulting trajectory data set after applying segmentation settings Low Tolerance (LT), Medium Tolerance (MT) and High Tolerance (HT) on RadrPlus and Geolife data sets.

We notice that, in the case of Geolife, the lower the tolerance, the greater the number of trajectories, and the shorter the trajectory length and duration. However, in the case of RadrPlus, the lower the tolerance, the lesser the number of trajectories, and the shorter the trajectory length and duration. These results are interesting due to the fact that Geolife's data was collected with a fixed, short sampling interval (Static Duty Cycle (Wu et al., 2011)) while RadrPlus' data collection process used a dynamic system that allocates short and long sampling intervals depending on the context (Dynamic Duty Cycle (Wu et al., 2011)). Therefore, we observe how the election of the data collection method affects the main characteristics of the resulting segmented trajectories.

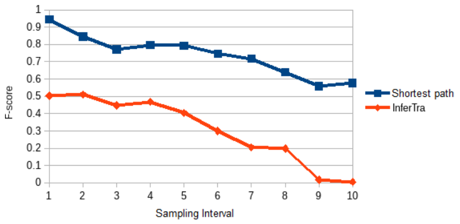
	Sampling Interval	**LT**	**MT**	**HT**
RadrPlus (SP)	1	0.944	0.930	**0.963**
	2	0.846	0.852	**0.903**
	3	0.770	0.790	**0.857**
	4	0.796	0.771	**0.800**
	5	**0.793**	0.748	0.762
	6	**0.746**	0.689	0.732
	7	**0.718**	0.680	0.717
	8	0.639	0.636	**0.700**
	9	0.558	0.567	**0.676**
	10	0.578	0.558	**0.673**
Geolife (SP)	1	0.959	0.960	**0.962**
	2	0.902	**0.909**	0.907
	3	0.862	0.868	**0.870**
	4	0.818	0.835	**0.837**
	5	0.775	0.789	**0.797**
	6	0.754	0.782	**0.783**
	7	0.746	0.766	**0.771**
	8	0.721	0.751	**0.761**
	9	0.694	0.717	**0.732**
	10	0.671	0.705	**0.727**
RadrPlus (IT)	1	0.502	0.608	**0.654**
	2	0.510	0.481	**0.593**
	3	0.446	0.430	**0.539**
	4	0.467	0.404	**0.507**
	5	0.404	0.350	**0.434**
	6	0.298	0.323	**0.411**
	7	0.205	0.303	**0.353**
	8	0.197	0.237	**0.315**
	9	0.018	0.230	**0.273**
	10	0.005	0.154	**0.231**
Geolife (IT)	1	0.558	**0.569**	0.558
	2	0.484	**0.494**	0.490
	3	0.437	**0.455**	0.450
	4	0.395	0.421	**0.424**
	5	0.363	0.382	**0.401**
	6	0.342	0.371	**0.386**
	7	0.314	0.344	**0.369**
	8	0.281	0.326	**0.346**
	9	0.264	0.300	**0.326**
	10	0.250	0.285	**0.309**

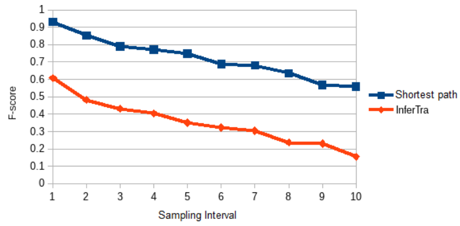
Table 3: Results of trajectory reconstruction on RadrPlus and Geolife data sets under trajectory segmentation configurations High Tolerance (HT), Medium Tolerance (MT) and Low Tolerance (LT) for different sampling intervals.

On the other hand, since Infertra reconstructs a trajectory as a weighted graph, we use the adapted F-score measure described in Banerjee (2014) to evaluate Infertra performance, whereas the standard F-score was used in the case of Shortest Path.

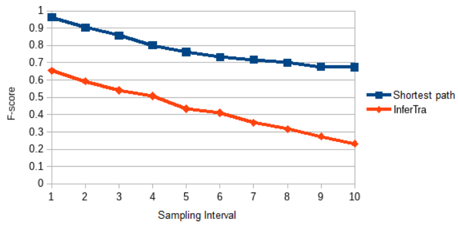
These two techniques are evaluated for different sampling rates expressed in minutes.

From Figures 1 and 2, we can easily observe



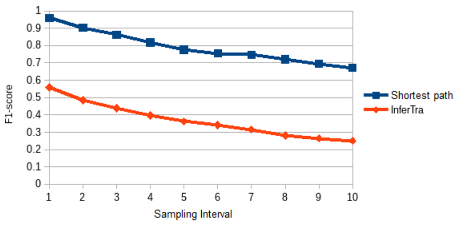


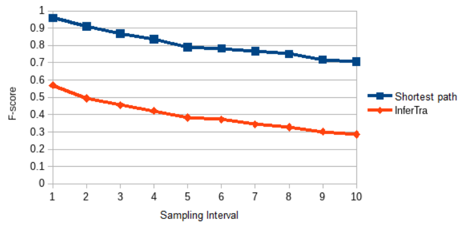
(a) (b)



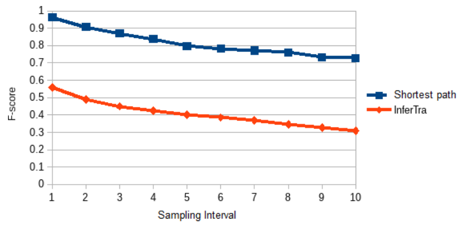
(c)

Figure 1: Results of the trajectory reconstruction on the RadrPlus data set using InferTra and Shortest Path algorithms. Each point in the figure indicates the value of the F-score for a certain sampling interval, expressed in minutes, under the segmentation settings (a) High Tolerance, (b) Medium Tolerance and (c) Low Tolerance.





(a) (b)



(c)

Figure 2: Results of the trajectory reconstruction on the Geolife data set using InferTra and Shortest Path algorithms. Each point in the figure indicates the value of the F-score for a certain sampling interval, expressed in minutes, under the segmentation settings (a) High Tolerance, (b) Medium Tolerance and (c) Low Tolerance.

that Shortest Path greatly outperforms InferTra, regardless of the data set, trajectory segmentation setting and sampling interval used. Additionally, as expected, we also observe that the best results correspond to the shortest sampling intervals. Nevertheless, it is not clear if there is a difference among trajectory segmentation settings. Thus, to analyze the impact of these settings over the reconstruction, we organize the data so that we can easily compare the obtained results for each setting. This way, in Table 3, we observe a clear evidence that the High Tolerance (HT) setting obtains better results for the majority of the cases. This setting only presents lesser values of F-score in the 17.5% of the cases.

4 Conclusion

We studied the reconstruction of pedestrian trajectories using road network information. Three different segmentation settings were proposed to study their influence over the reconstruction. These settings were established based on the concept of tolerance which was defined based on two criteria that comprise the distance and time span between points in a trajectory. Moreover, two state-of-the-art techniques were tested. One of these techniques uses both road network information and historical trajectories, whereas the simpler one only uses the road network structure. Empirical analysis of these two techniques on two data sets shows that the simpler technique performs better under limited data conditions than the more complex one when it comes to pedestrian trajectories. Additionally, the high tolerance segmentation setting proposed obtains better reconstruction results in a majority of the cases for both techniques.

Acknowledgments

This work was supported by FAPESP grant 2015/14228-9, CNPq grants 302645/2015-2, 162262/2014-0 and CAPES grant PROEX-9152559/M.

References

Prithu Banerjee, Sayan Ranu, and Sriram Raghavan. 2014. Inferring uncertain trajectories from partial observations. In *2014 IEEE International Conference on Data Mining*. IEEE, pages 30–39.

Richard G Baraniuk. 2011. More is less: signal processing and the data deluge. *Science* 331(6018):717–719.

Meng-Fen Chiang, Yung-Hsiang Lin, Wen-Chih Peng, and Philip S Yu. 2013. Inferring distant-time location in low-sampling-rate trajectories. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 1454–1457.

Zhenni Feng and Yanmin Zhu. 2016. A survey on trajectory data mining: Techniques and applications. *IEEE Access* 4:2056–2067.

Joshua S Greenfeld. 2002. Matching gps observations to locations on a digital map. In *Transportation Research Board 81st Annual Meeting*.

DMJM HARRIS, AECOM, and Clarke Caton Hintz. 2008. New jersey long range transportation plan 2030. technical memorandum. task 11: Local street connectivity redefined. Technical report, The New Jersey Department of Transportation, New Jersey.

Timothy Hunter, Pieter Abbeel, and Alexandre M Bayen. 2013. The path inference filter: model-based low-latency map matching of probe vehicle data. In *Algorithmic Foundations of Robotics X*, Springer, pages 591–607.

Mu Li, Amr Ahmed, and Alexander J Smola. 2015. Inferring movement trajectories from gps snippets. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, pages 325–334.

Yin Lou, Chengyang Zhang, Yu Zheng, Xing Xie, Wei Wang, and Yan Huang. 2009. Map-matching for low-sampling-rate gps trajectories. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM, pages 352–361.

Transportation Studies Group of the German Aerospace Center et al. 2012. *Urban Block Design guideline / manual to best practice - Project METRASYS*.

Ling-Yin Wei, Yu Zheng, and Wen-Chih Peng. 2012. Constructing popular routes from uncertain trajectories. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 195–203.

Chao-Lin Wu, Yu-Te Huang, Cheng-Lung Wu, Hao-hua Chu, Polly Huang, and Ling-Jyh Chen. 2011. An adaptive duty-cycle scheme for gps scheduling in mobile location sensing applications. *Proc. of PhoneSense* .

Llewelyn Davies Yeang et al. 2000. Urban design compendium. *English Partnerships/Housing Corporation, London* .

Jing Yuan, Yu Zheng, Chengyang Zhang, Xing Xie, and Guang-Zhong Sun. 2010. An interactive-voting based map matching algorithm. In *Mobile Data Management (MDM), 2010 Eleventh International Conference on*. IEEE, pages 43–52.

Kai Zheng, Yu Zheng, Xing Xie, and Xiaofang Zhou. 2012. Reducing uncertainty of low-sampling-rate trajectories. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, pages 1144–1155.

Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, pages 312–321.

Yu Zheng, Xing Xie, and Wei-Ying Ma. 2010. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* 33(2):32–39.

Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*. ACM, pages 791–800.

Yu Zheng and Xiaofang Zhou. 2011. *Computing with Spatial Trajectories*. Springer New York.