Comparing predefined and learned trajectory partitioning with applications to pedestrian route prediction

Mark Dimond[1], Gavin Smith 2[2], James Goulding[2],
Mike Jackson[1], Xiaolin Meng[1]

[1]Nottingham Geospatial Institute,
University of Nottingham, Triumph Road, Nottingham NG7 2TU
Tel. (0115) 823 2316
psxmd@nottingham.ac.uk

[2]Horizon Digital Economy Research Institute,
University of Nottingham, Triumph Road, Nottingham NG7 2TU

Summary: Route and destination prediction of mobile device users has become increasingly feasible in recent years due to improvements in positioning technology. In general route prediction requires identification of discrete trajectories from unprocessed histories of user positions, automation of which could improve performance and reduce data requirements for prediction. This paper investigates the assumption that user time spent at a position can be used to identify trajectory partitioning locations, providing a comparison between an automated method and a known ground truth. In addition the impact on trajectory prediction is considered.

KEYWORDS: movement prediction, spatial data mining, geographic representation

1. Introduction

Increased ownership of mobile devices with more precise positioning sensors allows new possibilities in the development of location-based services. One such possibility is the prediction and analysis of user movement through statistical analysis of movement history. Services such as targeted advertising (Krumm, 2010), communications infrastructure provisioning (Yavas *et al*, 2005), and monitoring of users with sensory impairments (Patterson *et al*, 2004) could all be improved through implementation of reliable movement prediction. While early work in the field focused on *destination* prediction (see Karbassi and Barth, 2003; Ashbrook and Starner, 2003) – attempting to identify the likely ultimate goal of the user from a history of positions – this can be extended to *route* prediction (Smith and Goulding, 2011), in order to provide more detailed predictions and facilitate the use of intermediate position estimates where destination prediction might fail.

However, raw positioning data streams that serve as the basis for predictive models are unlikely to be immediately separable to discrete trajectories. Instead it is necessary to identify geographic *locations* where streams can be split, identifying individual journey routes or *trajectories*. In many cases it will be possible to identify appropriate splitting locations from prior knowledge of the target domain, such as a feature gazetteer or a town plan. Nevertheless, there are situations where external knowledge is incomplete or unavailable due to time, cost or logistics. In such situations it is desirable to computationally identify these locations via analysis of user movement histories, to replace or augment the traditional formalisation of known locations.

To this end, this paper will use data from the DSCENT location-based game (Sandham *et al*, 2011) to compare trajectories generated computationally against those delimited via prior destination knowledge, examining the prediction performance achieved. Within the game, players are given specific roles with objectives to complete by travelling between set geographic locations.

2. Problem

As discussed, user position histories often do not include identification of different trajectories taken by a user. It is therefore necessary to perform some analysis to approximate them, using some property of the position data as an indicator of trajectory split. For this work we will investigate the use of user time spent at each position as an indicator of a trajectory splitting location, a property we refer to as *dwell time*. Use of this metric relies upon the assumption that users spend more time at places that represent the starts or ends of their trajectories: there may be instances in which this would be inappropriate, such as slow-moving traffic queues, but for simple location-based game data this is a reasonable approximation.

The aims of this work are twofold:-
- to qualitatively identify whether locations detected using the dwell-time property were topographically similar to the predefined locations used for the DSCENT data, and
- to assess the differences in prediction error between trajectories identified from predefined locations and from those mined using different enumerations of dwell-time.

If it can be shown that dwell-time is useful in approximating trajectory partitioning locations for route prediction, this will provide the basis for further investigation of route prediction without the need for prior knowledge of journey locations.

3. Related Work

While seeming intuitive, little prior work has examined the assumption that dwell time is an indicator of trajectory start and end points. For instance Ashbrook and Starner (2003) only consider one approach to automatically determining a minimum dwell, and resort to a value that seems *reasonable*. However, they do not examine the core assumption that trajectories partitioned by dwell time actually correspond to trajectory partitioning as perceived by those making journeys. Similarly Krumm and Horwitz (2006) and Froelich and Krumm (2008) use dwell-time to partition trajectories based on values chosen empirically. Once again the authors utilize dwell time without further investigation.

Alternative, but related, work is that which seeks to identify *significant locations* (Marmasse and Schmandt 2000; Liao *et al*, 2007). This work is relevant since determining significant locations is the first step in performing trajectory partitioning. In Marmasse and Schmandt (2000), the loss of GPS signal multiple times within a fixed radius is used to identify significant locations. The assumption here is that the signal is lost when entering places such as buildings. While interesting, much like the assumption of dwell time in the aforementioned work, this assumption is simply utilized and not investigated further.

In summary, while dwell time has seen significant use in trajectory partitioning as a pre-processing step to movement prediction, limited work exists investigating its validity and the specific way in which it is used. It is this aspect we seek to address. The DSCENT data presents an unusual opportunity in this respect: though it took place on a relatively small scale (400x200 metres), the locations which represent users' actual goals are available for comparison.

4. Implementation

Data used for this work originated from the DSCENT location-based games (Sandham, 2011), within which a player's position is recorded in the format *(lat, long, time)*, in addition to some further information such as game and player number and an estimate of the GPS accuracy. An observation of visiting a game location is determined using a shapefile of polygons representing the locations. If the coordinates are not contained by any location polygon, the observation is assigned a location ID of 0. This information can later be used with a simple database query to enumerate the individual observations (a list of cell occupancies) representing travel between locations.

In order to transform observations in continuous space to contiguous regions of similar dwell time, a workflow was developed to discretise observations. Below we outline the steps taken in this process, including the assumptions made and parameters used for different stages of the location region modelling. Extended implementation details are omitted for brevity.

1. Player dwell times are allocated to cells in a 5 metre grid, forming a *heatmap* of occupancies. These heatmaps can be built per-player, per-game or across all games. In this work we use data from all games. The use of a grid addresses the problem of GPS accuracy to some extent, since nearby observations are collated.
2. Cells above a certain percentile of dwell time are clustered using the DBSCAN density-based clusterer (Ester and Kriegel, 1996), with a density parameter (ε-distance) of 10 metres. The number of cells to be retained can be adjusted – see Figure 2.
3. Detected clusters are enclosed using a convex hull operation, removing empty areas in clusters. The resulting polygons are the trajectory partitioning locations. Figure 3 shows the output of this stage.
4. Player trajectory histories are built using location polygons and the spatial containment test described above. A *trajectory* is defined as a sequence of points in which a user starts within a location, leaves that location, and subsequently enters any location (including the start location).


(a)   (b)
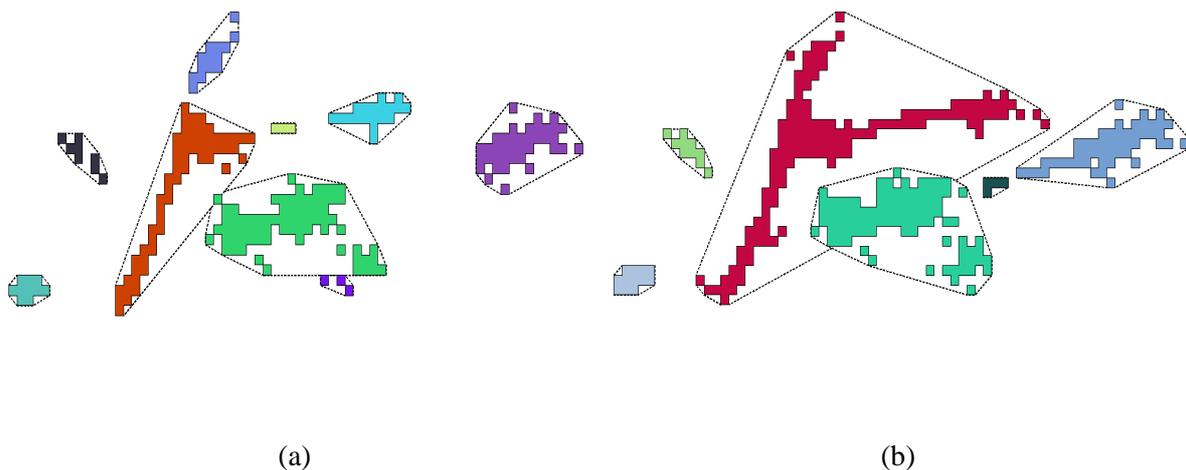
Figure 2. Clusters from the a) 85th and b) 95th percentile of dwell times, without the convex hull applied. Convex hull output is marked with a dotted line. Output from other levels is omitted: for example the 95th percentile produced only three small clusters.

(a)   (b)
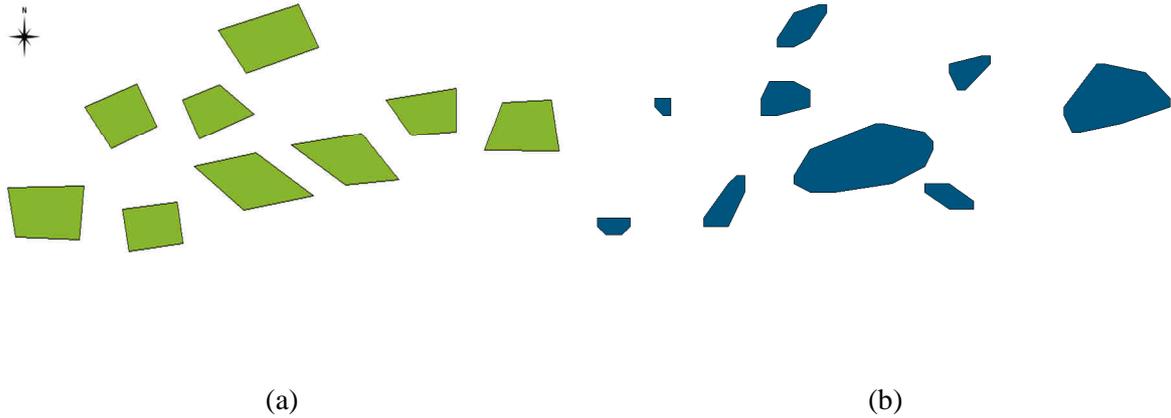
Figure 3. a) Predefined regions and b) mined journey regions from the 90th percentile of dwell cells, with the convex hull applied.

5. Evaluation

The trajectory partitioning locations used for the initial DSCENT games and those mined from the 90th percentile of dwell-time grid cells are shown in Figure 3. This level was selected empirically as being particularly similar to predefined locations. Qualitative inspection of these shows that detected locations:

- are mostly similar in position to the predefined locations
- are generally smaller in area

Table 1 gives some statistics of the different regions' geographies. Detected locations are overlapped by predefined locations for approximately 60% of their area on average. However the average area of predefined locations is over double that of detected locations. This is possibly because users were unlikely to visit the whole area of a location with equal probability; our detection method can only find cells that were commonly visited.

	Predefined	Detected
Mean area	1 547 m^2	700 m^2
Overlap vs detected	27%	-
Overlap vs predefined	-	58%

Table 1. Area statistics for the trajectory partitioning regions

The similar positioning of the automatically detected locations compared to manually identified locations provides evidence that dwell time can provide an approximation of trajectory partitioning locations. The approximation is not perfect, however, with the reduced size of the locations leading to trajectories of longer length than those using predefined locations (see Figure 5). This occurs as the smaller locations mean that users sometime are not marked as entering a location - in this case one long trajectory is created. This is in contrast to when the large, manually identified, locations are used

and two shorter trajectories are created. For the use of the trajectories in the target application of route prediction this has the potential to make route prediction more difficult.

Additionally there is an instance where the detection process identifies a single location where two predefined locations exist. Once again considering the target application of route prediction, this means trajectories between these locations are not created. In this case the ability to predict between these locations is lost. However, this could be a problem of sensitivity of the method used to enumerate dwell time, investigation of which is interesting future work.

To investigate the effect of the automated partitioning in the context of the target application, route prediction, the different methods of trajectory partitioning were used to build user trajectory histories and trajectory prediction performed. Trajectory prediction was based on a variable length Markov Model with an identified trajectory being split to provide the input and a ground truth to which the prediction was then compared. Comparison between ground truth and prediction was made using a variation of the Fréchet distance (Eiter and Mannila, 1994). Table 2 and Figure 4 provide some descriptive statistics and insights into the different partitioning geographies. The number of cases where prediction failed (no prediction made) is also given. The results show that, as expected, prediction performance suffers from the automated trajectory partitioning. Encouragingly, however, the amount by which prediction suffers is not so great as to discourage the pursuit of further refinements to the technique.

| Geography	Regions	Trajectories detected	Mean trajectory length (points)	Prediction tests	Predictions within 10m	Predictions not made
Predefined	9	1681	17.3	2498	16.7%	32.9%
Mined (all games, above 90th percentile dwell times)	9	1518	19.7	2263	12.8%	39.1%

Table 2. Region statistics
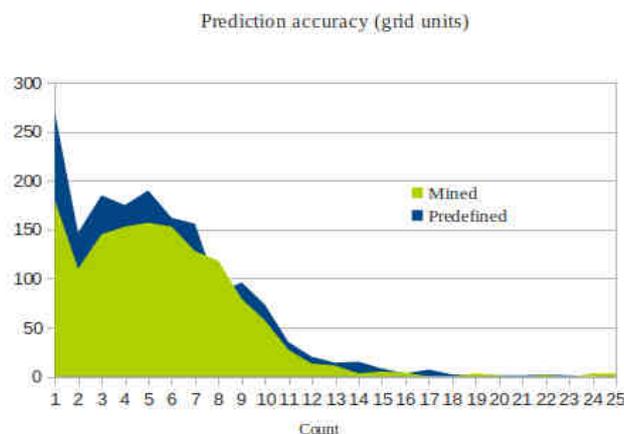
Figure 4. Prediction accuracy scores (in grid units) for each type of partitioning location
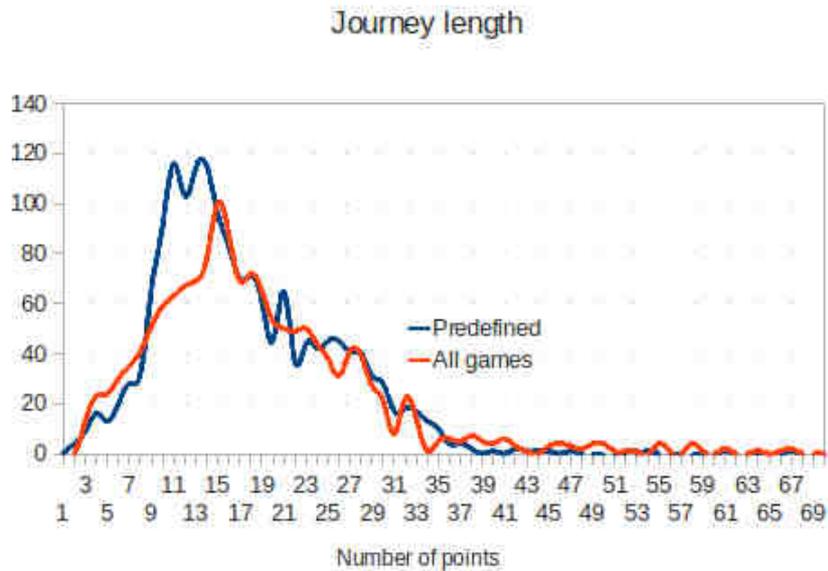
Figure 5. Trajectory lengths

6. Conclusions and Future Work

Qualitative comparison of trajectory locations mined from the top 10% of dwell times shows promising overlap with predefined partitioning locations. Differences in prediction accuracy based on trajectories created automatically similarly shows that while not perfect, the approach has potential.

As noted there are clear limitations to the method used in this case to detect trajectory splitting locations for a location-based game, with future work planned to address this. Close predefined locations are detected as a single region, meaning trajectories are not detected between them. Parameter setting is a difficulty, particularly identification of an appropriate proportion of enumerated dwell-time to use for location building. The setting of grid cell size for enumeration and the DBSCAN clustering radius has not been addressed and better understanding of these parameters would be valuable. Further work will therefore specifically look at parameterisation of the dwell score threshold, and the effects that resulting different location sizes have upon route splitting.

Using careful examination of prediction results and through qualitative assessment of built partitioning regions, dwell time can be used as an identifier of breaks in a user's journey. If a workable automatic method could be identified for selecting optimal partitioning locations from dwell times, that did not rely upon comparison with known geographic information, this would be of significant utility to route prediction.

7. Acknowledgements

8. References

Ashbrook D and Starner T (2003) Using GPS to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput*. 7, 5 (October 2003), 275-286.

Eiter, T and Mannila, H (1994) Computing discrete Fréchet distance, *Tech. Report CD-TR 94/64*, Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria

Ester M, Kriegel HP, Sander J, and Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, volume 1996, pages 226-231. Portland: AAAI Press

Froelich, J and Krumm, J (2008) Route Prediction from Trip Observations. In *Intelligent Vehicle Initiative (IVI) Technology Controls and Navigation Systems: Special Publications of the SAE World Congress*, p53-66. SAE, Washington DC.

Karbassi, A and Barth, M (2003) Vehicle route prediction and time of arrival estimation: techniques for improved transportation system management. In *Intelligent Vehicles Symposium, 2003 Proceedings*. IEEE, pp 511-516.

Krumm, J and Horwitz, K (2006) Predestination: Inferring destinations from partial trajectories. In Paul Dourish and Adrian Friday, eds., *Ubicomp 2006: Ubiquitous computing. Lecture Notes in Computer Science*, vol 4206, 2006

Krumm, J, (2010) Ubiquitous Advertising: The Killer Application for the 21st Century. *IEEE Pervasive Computing*, PrePrints (99).

Marmasse, N and Schmandt C (2000) Location-aware information delivery with commotion. In *Proceedings of the 2nd International Symposium on Handheld and Ubiquitous Computing*, HUC '00, pp 157-171, London, UK. Springer-Verlag.

Patterson, D.J., et al., (2004). Opportunity Knocks: A System to Provide Cognitive Assistance with Transportation Services. In: N. Davies, E. Mynatt and I. Siio, eds. UbiComp 2004: Ubiquitous Computing., Vol. 3205 of Lecture Notes in Computer Science Springer, 433–450

Sandham, A et al (2011). Scent trails: countering terrorism through informed surveillance. In *HCI International*, Orlando, Florida, USA.

Smith, G, and Goulding, J, Pedestrian Route Prediction using Augmented Cover Trees, NEMO, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, September 2011.

Yavas G, Katsaros D, Ulusoy O, and Manolopoulos Y (2005). A data mining approach for location prediction in mobile environments. *Data Knowl. Eng*. 54, 2 (August 2005), 121-146.

9. Biography

Mark Dimond is a PhD student funded through the Horizon Doctoral Training Centre programme, and is based at the newly formed Nottingham Geospatial Institute. Mark's research interests include spatial data mining and machine learning for location-based applications. Previous work included database administration and IT systems support.

Gavin Smith is a Research Fellow at the Horizon Digitial Economy Research Institute at the University of Nottingham. Gavin's interests are in data mining and machine learning, with a recent focus on pedestrian route prediction.

Dr. James Goulding is a Research Fellow at the EPSRC funded Horizon Digital Economy Institute at the University of Nottingham. His background is was in Econometrics before progressing to Computer Science. He is internationally published in the fields of Data Theory, Information Retrieval, Collective Intelligence and Geospatial Science.

Prof. Mike Jackson is Professor of Geospatial Science at the University of Nottingham and Founder/Director of the Centre for Geospatial Science 2005-2011. He is Chair of the Association of Geographic Information Laboratories of Europe (AGILE) and a Director of the Open Geospatial Consortium (OGC). Previous posts include being Director, Space Division, QinetiQ and CEO, Laser-Scan.

Dr. Xiaolin Meng is Associate Professor at the Nottingham Geospatial Institute at the University of Nottingham in the UK. His research interests include Network-based GNSS, Location Based Services and Intelligent Transportation Systems and Service.