ARTICLE IN PRESS

Travel Behaviour and Society xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Travel Behaviour and Society

journal homepage: www.elsevier.com/locate/tbs

GIS-based episode reconstruction toolkit (GERT): A transferable, modular, and scalable framework for automated extraction of activity episodes from GPS data

Ron Dalumpines, Darren M. Scott *

TransLAB (Transportation Research Lab), School of Geography and Earth Sciences, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada

ARTICLE INFO

Article history:
Received 13 June 2016
Received in revised form 1 March 2017
Accepted 3 April 2017
Available online xxxx

Keywords:
Episode reconstruction
GIS
GPS
Map-matching
Route choice
Travel behavior

ABSTRACT

The widespread use of global positioning systems (GPS) has prompted transportation researchers to develop tools and methods to extract information from person-based GPS data. However, most of these procedures suffer from specific data requirements and complexity that limit their transferability. Further, they have a limited set of modules to extract all necessary information (e.g., route attributes), and were not specifically designed to handle huge GPS datasets. To deal effectively with these problems, this paper presents a framework based on three design principles (transferability, modularity, and scalability), along with the geographic information system (GIS)-based episode reconstruction toolkit (GERT) based on this framework, for automated extraction of activity episodes from GPS data.

About 26,000 episodes were automatically reconstructed using GERT from 47.3 million GPS points. A comparison of the episode and duration distributions reveal similar patterns between time-use diary and GPS episodes, a similarity that confirms that GERT's modules work properly in reconstructing episodes from GPS data. GERT's overall performance suggests potential because of its scalability – GERT can scale up to large GPS data; modularity – GERT has a complete set of tools to support activity analyses and route choice model estimations; and transferability – GERT's reliance on generic variables (latitude, longitude, time) makes it applicable to other places. Overall, GERT's modules provide transportation researchers with rich datasets (*stop* and travel episodes, activity locations, travel segments, route choice sets, route attributes) for improving the understanding of activity/travel patterns in general and route choice decisions in particular.

1. Introduction

Person-based global positioning system (GPS) devices capture the start and end times of activity episodes, their duration, and travel routes in greater spatial and temporal resolution than traditional recall-based surveys. Because not all episode information can be directly captured by these devices, GPS is increasingly used to supplement traditional survey methods primarily to increase the accuracy of data collection and reduce burden among respondents (Wolf, 2000; Bricka, 2008; Millward and Spinney, 2011). Several procedures have been developed to extract information from person-based GPS data in order to supplement data from recall-based surveys (e.g., Stopher et al., 2005; Chung and Shalaby, 2005; Tsui and Shalaby, 2006; Zheng et al., 2008; Schuessler and Axhausen, 2009; Bohte and Maat, 2009). However, most of these procedures suffer from specific data requirements and complexity that limit their transferability to other application environments. Further, they have a limited set of modules to extract all necessary information such as route attributes, and were not specifically designed to handle huge GPS datasets (Schuessler and Axhausen, 2009; Biljecki, 2010; Lawson et al., 2010). To deal effectively with these problems, this paper presents a framework based on three design principles (transferability, modularity, and scalability), and a GIS-based toolkit (based on this framework) for automated extraction of activity episodes from GPS data. Without an effective framework and a toolkit to implement this framework, we cannot take full advantage of GPS data for the following reasons: (1) difficult to adopt tools developed by other researchers for lack of transferability, (2) limited ability to derive more information from GPS data for lack of an integrated set of modules, and (3) high computational costs and lack of automatic procedures in dealing with huge datasets. Before elaborating on these issues (lack of transfer-

* Corresponding author.
E-mail addresses: dalumprf@mcmaster.ca (R. Dalumpines), scottdm@mcmaster.ca (D.M. Scott).

http://dx.doi.org/10.1016/j.tbs.2017.04.001
2214-367X/

ability, huge GPS data, incomplete set of tools), it is best to clarify some key terms used in this paper.

In the context of activity analysis, a person's 24-h (daily) activities can be subdivided into episodes, which are differentiated based on location (inside a building or travel mode); hence an activity episode can be a stationary episode (*stop* episode) or a travel episode (e.g., *car* episode). Travel episodes are synonymous to *trips*. In this paper, a segment refers to a sequence of GPS points similarly classified as *stop* or *trip* points (Fig. 1), while an episode refers to the diary-equivalent classification of a segment based on location (e.g., *stop* episode, *walk* episode, *car* episode, and so on) with attributes such as start time, end time, duration, and distance. In different contexts, some researchers used the term "stages" (Schuessler and Axhausen, 2009), "objects" (Dodge et al., 2009), and "segments" (Zheng et al., 2008; Gong et al., 2012) to refer to episodes.

1.1. Challenges in developing tools and methods for extracting episodes from GPS data

Existing procedures in extracting or reconstructing episodes from GPS data can be categorized into several modules: preprocessing (data filtering and smoothing), extraction of episodes (stages or segments), mode detection (assignment of mode to travel episode), route detection (map-matching), and purpose detection. To the authors' knowledge, original attempts to automate the extraction of episodes in the transportation literature (Chung and Shalaby, 2005; Stopher et al., 2005; Schuessler and Axhausen, 2009) suggest the lack of transferability of existing modules (e.g., non-generic variables used in preprocessing), an incomplete set of modules (e.g., no modules for purpose detection in two studies while not fully automatic for route detection), and only one study was specifically designed for large GPS data (Schuessler and Axhausen, 2009) (for related studies in other fields, see Biljecki (2010) and Bolbol et al. (2012)). Only a few studies (Schuessler and Axhausen, 2009; Bohte and Maat, 2009) specifically addressed the development of tools and methods in extracting travel episodes and trip purposes from large-scale GPS datasets.

1.1.1. Lack of transferability

Existing methods used unique inputs or variables to filter valid points for extracting and classifying episodes. Because procedures are often designed to make the most out of the data available, the decision rules vary depending on the characteristics of the input data. For example, some researchers (e.g., Wolf, 2000; Stopher et al., 2005; Chung and Shalaby, 2005) used the number of satellites, heading, and horizontal dilution of precision (HDOP) in a preprocessing module to remove outliers and invalid GPS points; in the absence of the above inputs, Schuessler and Axhausen (2009) instead used the known altitude of Switzerland to remove low quality or erroneous GPS points. Other researchers (e.g., Chung and Shalaby, 2005; Bohte and Maat, 2009; Gong et al., 2012) used proximity measures (e.g., distances to bus, subway, and railway stations) to determine probable travel modes; however, threshold distances vary significantly among studies. For rule-based procedures, there is absence of clear guidelines in finding optimum threshold values as implemented in various modules because often the cut-off values are based on subjective judgment and the quality of GPS data.

Researchers have focused more on the effectiveness of their tools and methods, and little attention had been given to the transferability of the tools and methods. Eventually each researcher has developed a unique set of tools using unique inputs for specific purposes. Transferability of the existing methods remains a challenge. This issue had been identified by Lawson et al. (2010) based on their experiment that aims to replicate currently used methods on mode detection. Their experiment provides an objective examination of currently used mode detection algorithms to date. They recognized the difficulty of directly comparing different approaches because of different data used in developing these methods, not to mention the different variables required by each approach. At this juncture, the existing methods are mostly not generic, making it difficult to apply the same methods in different environments with minimal effort.

1.1.2. Processing demands of huge GPS data

Current personal-based GPS devices are becoming widespread because of better accuracy, more portability (pocket-size), and lower costs. In recent years, we observed an increase in large-scale GPS data used for travel episode (trip) extraction: more than 20,000 km of GPS trajectories (Zheng et al., 2008); 64.5 million GPS points (Schuessler and Axhausen, 2009); 17.6 million GPS points (Bohte and Maat, 2009; Biljecki, 2010); 47.3 million GPS points (Millward and Spinney, 2011); and perhaps more. Manual procedures are no longer practical in dealing with huge GPS data that span millions of records (Schuessler and Axhausen, 2009). The availability of huge GPS data and the high potential to collect more make it necessary to come up with efficient procedures that can automatically extract information from these data.

1.1.3. Incomplete set of tools

From the perspective of activity analysis, most of the existing methods did not fully capture valuable information from GPS trajectories for these methods were focused more on mode detection, that is, the extraction of travel episodes and classifying these episodes into several types based on travel modes (Stopher et al., 2005; Schuessler and Axhausen, 2009; Gong et al., 2012). Hence no modules were specifically developed to extract information associated with activity locations (*stop* episodes), wherein more information can be extracted with the aid of additional data such as land use and potential activity locations (PAL), and information on observed routes (road attributes) connecting these locations. Although Stopher et al. (2005) and Bohte and Maat (2009) also tried to capture trip purposes (the former asked respondents to provide addresses of activities while the latter used points of inter-

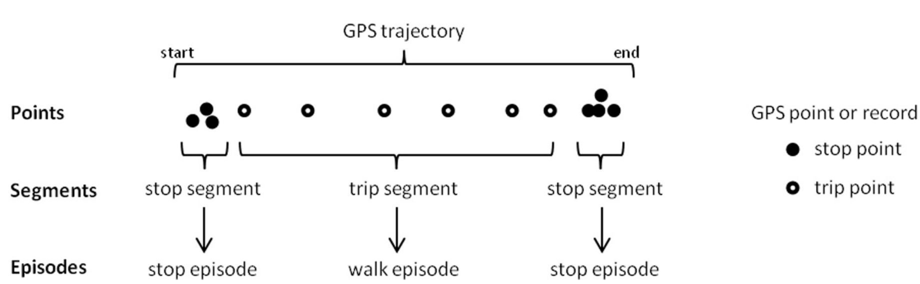

Fig. 1. GPS trajectory subdivided into points, segments, and episodes.