Computers, Environment and Urban Systems 36 (2012) 140–153

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

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

Exploratory calibration of a spatial interaction model using taxi GPS trajectories

Yang Yue [a,b,c,*], Han-dong Wang [a,b,c], Bo Hu [a,b,c], Qing-quan Li [a,b,c], Yu-guang Li [a,b,c], Anthony G.O. Yeh [d]

[a] Transportation Research Center, Wuhan University, Wuhan, Hubei 430079, China
[b] Engineering Research Center for Smart Acquisition and Applications of Spatiotemporal Data, Ministry of Education, Wuhan, Hubei 430079, China
[c] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei 430079, China
[d] Centre of Urban Studies and Urban Planning, The University of Hong Kong, Hong Kong SAR, China

ARTICLE INFO

Article history:
Available online 3 November 2011

Keywords:
Model calibration
Spatial interaction model
Trading area analysis
GPS data
Trajectory analysis

ABSTRACT

Model calibration is the cornerstone of spatial interaction models in many geographic, transportation and marketing analysis. Conventional questionnaire approaches that collect data for model calibration are both labor-intensive and time-consuming, and generally show a poor response rate. This study takes advantage of increasingly available vehicle GPS trajectory data to conduct spatial interaction model calibration. A Huff model for retail trading area analysis was used as an example. Model calibration and parameter validation were conducted based on over 63,000 taxi GPS trajectories for seven major shopping centers in Wuhan, a large city in China. The results were positive and in general showed satisfactory descriptive and predictive capability. This study demonstrated the feasibility of using the emerging technology to calibrate spatial interaction models (and also showed the potential for use in other related studies). The main advantage of using these new data sources is that they allow efficient use of increasingly available positioning data, which is easier to collect than conventional customer surveys, and usually with larger data sizes. It also allows inferences to be made about distance-decay rates based on accurate computation of travel time and distance. This could save both time and expense in many related areas of research, while achieving high quality model calibration results.

1. Introduction

The calibration of spatial interaction models is a significant challenge. Without accurate calibration, a model is almost useless for both description and prediction (Rodrigue, Comtois, & Slack, 2009). Model calibration usually requires a large amount of high quality spatial interaction data to ensure that the values of estimated parameters can "best fit" the observed data. Conventionally, data is collected using surveys at a chosen destination, such as a store, or in questionnaires administered to people who have chosen a particular service. For example, data for transportation analysis is often obtained through interviews at stations, parking lots, or during on-board transit (Thill & Horowitz, 1991). The data collection for model calibration, therefore, is both labor-intensive and time-consuming, and generally shows a poor response rate (O'kelly, 1999). Moreover, most empirical analyses use household-level scanner panel data, such as location, income, and the number of family members; thus data availability is often a problem. Similar problems also exist in other data-intensive studies.

With the development of sensor technology and location-based services (LBS), there is increasing availability of positioning data, such as vehicle GPS trajectory data, which can assist research in many fields. In recent years, taxis in many large cities in China have been equipped with GPS and data transmission devices that can be used to obtain taxi trajectories. Such trajectory data are usually used for taxi management, allowing for real-time traffic information as well. This is a rich informative data source, and could be very valuable for many applications, such as urban, business and human behavior analysis. However, at present, most studies that have used this type of data are in transportation fields (Liu, Andris, & Ratti, 2010; Yue, Zou, & Li, 2009b).

The purpose of this study is to investigate the feasibility of using the taxi GPS trajectories to calibrate a spatial interaction model, specifically retail trading area delimitation. Trading area analysis is used to help to increase understanding of market areas around stores or office locations, discover where customers are coming from, and, further, predict trading areas for new locations (Huff, 1963, 1964). To our knowledge, there is only one previous study addressing a similar question using a similar trajectory dataset. Kawasaki and Axhausen (2009) utilized individual GPS traces to identify shopping trips and shopping destination choices, and showed that the generated data was comparable with travel behavior survey data. The contribution of our study is twofold.

* Corresponding author at: Transportation Research Center, Wuhan University, Wuhan, Hubei 430079, China. Tel.: +86 27 68778035 x8107; fax: +86 27 68778043.
E-mail address: yueyang@whu.edu.cn (Y. Yue).

0198-9715/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2011.09.002

First, it shows an alternative for spatial interaction model calibration. Positioning data is increasingly available and has the potential to improve calibration efficiency and accuracy. Meanwhile, we also list the advantages and disadvantages that may be important for related studies in the future. Second, we show how the GPS trajectory data can allow inferences to be made about distance-decay rates based on accurate computation of travel time and distance. This paper does not discuss calibration methodology. Related topics can be found at Batty and Mackie (1972), Hall (1975), Wu and Sitter (2001), and Roy and Thill (2004).

The rest of the paper describes the data that was used; explains the differences between the dataset used in this study and choice-based samples which are typically used for calibrating spatial interaction models; provides the data preprocessing method used to generate the choice-based sample; and presents a case study to illustrate the feasibility of model calibration using the GPS trajectory data. The advantages and disadvantages of the approach are covered in the discussion and conclusions.

2. Taxi GPS trajectory data

2.1. Data description

Table 1 shows the typical format of taxi trajectory data, including taxi location (*Longitude, Latitude*), speed, direction (*Angle*), passenger pick-up and drop-off information (*Status*), with associated time information. The data collection time interval is generally around 30 s–1 min. Delays or missing data may occur depending on the GPS signal, and additional records are collected when taxi load status changes.

The data used in this study was from a 24 h dataset collected on 28 December 2008, totaling over 63,000 trajectories from around 12,000 taxis. Although data from some successive weekend days would make the calibration more convincing, this was the only experimental dataset we were able to collect when conducting the analysis. However, December 28, 2008 was the last Sunday before the New Year and shopping trips were active and massive. So, we believe the data is representative in terms of customer attendance.

The dataset excludes pedestrian visitors, a limitation that means this study can only identify the vehicle-based retail trading area. However, this does not affect the generality of the study and the same approach could also work for pedestrian-based trajectory data collected by mobile phone or other LBS applications.

2.2. Gaps between GPS trajectory data and choice-based samples

Choice-based samples are normally used in calibrating spatial interaction models, which are designed on an endogenous variable, e.g., a group that has chosen to visit a certain center or destination (O'kelly, 1999). The ultimate goal of using the choice-based samples is to make inferences about the population at large. However, taxi trajectory data is random sampling per s. A number of gaps remain regarding the suitability of the trajectory data vs. choice-based samples:

Table 1
Typical taxi trajectory data format.

Field	Data type	Example	Remark
ID	Number		Taxi ID
Date	Date	2008-12-28	
Time	Time	10:17:27	
Longitude	Number	114.179301	
Latitude	Number	30.465626	
Speed	Number	13.28	Instant speed (km/h)
Angle	Number	285.39	Heading
Status	Boolean	1	0: unloaded; 1: loaded

2.2.1. Stratified (exogenous) random sampling vs. choice-based (endogenous) sampling

The taxi GPS trajectory data is exogenously stratified sampling data in terms of travel demand, or more strictly, stratified in terms of population income and shopping distance (mainly due to taxi fares), rather than choice. The data contains no explicit information regarding choice, and includes a large proportion of samples that do not participate in the activity of interest; whereas choice-based samples are endogenously stratified samples. The gap, therefore, is how to generate a choice-based sample from the exogenous sample.

2.2.2. GPS positioning error vs. actual customer patronage

The GPS positioning error in this dataset was around 10–30 m according to the data specification. However, the positioning error may have varied at different spots due to signal block or reflection of buildings and trees. So, it is usually difficult to accurately associate positioning data with a certain destination, especially in high-density areas, which was the case in the present study area. Thus, the data is more suitable for an area-based study instead of a point-based study. Choice-based samples, however, can be collected at very fine spatial granularity. In this case, the gap is in determining how to define customer patronage as accurate as possible with inherent GPS positioning error.

2.2.3. Single transportation mode vs. multimode

Another problem is that the dataset was limited to taxis, which excludes other transportation modes. This may limit the applicability of the taxi-based data in other areas with limited amount of taxi use and high taxi fares. As mentioned above, pedestrian data was not available. Therefore, the question arises of how to make best use of the single transportation mode data when multimode data is not available.

Although taxi trajectory data is stratified by travel demand, it disperses over a large geographical area. Such large-scale dispersion is useful for examining the penetration of a subject shopping center into the market from a variety of zones. However, given the inherent disadvantages listed above, the key problem therefore is how to extract customer-patronage data given the inherent disadvantages. In the next section, we demonstrate how to generate a choice-based sample from the taxi GPS trajectories for model calibration purposes.

3. Generating a choice-based sample from taxi trajectories

Since our dataset has passenger load and unload records, the problem turns to how to identify shopping trips from passenger pick-up and drop-off points. Due to GPS positioning error on one hand, and human behavior randomness on the other, it is not possible to relate every pick-up and drop-off point to a certain shopping activity in a high-density area. Therefore, it is difficult to pinpoint these drop-off points to a single store, except with the assistance of other data, such as in-door positioning data or conventional customer surveys.

In reality, anchor stores play an important role in shopping center attractiveness (Finn & Louviere, 1996). Our previous study (Yue, Zhuang, & Li, 2009a) has shown that taxi passenger pick-up and drop-off points in the study area are spatially aggregated, especially within major shopping centers. Therefore, we propose to define a buffer radius based on the anchor store (in our case, a shopping mall) to embed GPS error and human behavior randomness, as well as to reflect shopping center neighborhood impact.