Applied Soft Computing 34 (2015) 106–119

Contents lists available at ScienceDirect

Applied Soft Computing

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

The integration of classification tree and Sequence Alignment Method for exploring groups of population based on daily time use data

Natachai Wongchavalidkul, Mongkut Piantanakulchai*

School of Civil and Engineering Technology Sirindhorn International Institute of Technology Thammasat University, Thailand

ARTICLE INFO

Article history:
Received 19 September 2013
Received in revised form 3 September 2014
Accepted 15 April 2015
Available online 23 April 2015

Keywords:
Activity-travel pattern analysis
Classification tree
Sequence Alignment Method

ABSTRACT

Searching homogenous groups of individuals is one of the important steps in activity based travel demand modeling development. This study proposes an Integration of Classification tree And Sequence alignment method (ICAS) as a new classification method. The main advantage is the ability to explore all sources of lifestyle variations that have various data types including: sequential data, continuous variables, and discrete variables. These data are, for example, activity sequential patterns, socio-economic characteristics, and socio-demographic characteristics. Results from ICAS can also be used as both an activity classifier and an activity generator in an activity based travel demand modeling system. The proposed ICAS concept was evaluated with real world data, using the 2004 Bangkok time use data from Thailand's National Statistical Office (NSO).

1. Introduction

For more than a decade, the activity based modeling approach has increasingly received attention in transportation research. The activity based modeling approach extends the conventional trip based travel demand model by considering the real factors that cause people to travel, the activities that people carry out at different places and times. In this case, modeling inputs require disaggregate data of individuals rather than aggregate data of traffic analysis zones.

Therefore, the approach of searching homogenous groups of individuals who have similar activity patterns and lifestyles prior to developing the model is obviously important. This research classifies approaches in dividing homogenous groups of populations into two categories, activity schedule analysis (ASA) and Lifestyle classification (LSC). ASA is rooted from the famous Hägerstrand's time geography concept and mainly concentrates on the activity schedule data. On the other hand, LSC mainly focuses on various elements based on activity and travel behavior, which are related to fundamental human values and needs. Each of these alternatives has certain advantages and disadvantages in providing effective measures of individual activity patterns. The following sections discuss ASA and LSC in detail.

1.1. Activity schedule analysis (ASA)

ASA mainly focuses on a measurement of human activity characteristics and patterns. The earliest contribution is the development of a space-time path by Hägerstrand [1]. The space-time path is a three-dimensional orthogonal system. The system consists of a two-dimensional plane which indicates positions of an individual (spatial dimensions) and a vertical dimension axis which represents time taken by the individual at each position (a temporal dimension). Hägerstrand's concept of a space-time path, later known as time-geography, helps researchers to demonstrate human activities under constraints such as physical limitations and time [1–4]. Some transportation research adopts the time geography concept in various contexts such as improvements of the space-time path visualization [5–7] and the exploration of human activities and interactions [8,9].

Advances in information technology and the increasing number of available activity datasets push research on activity schedule analysis, such as the exploration of activity patterns from a large dataset. The objective is to answer questions such as who shares similar activity patterns. In this case, the classical time-geography concept is found to lack a quantitative method which is powerful enough to analyze a huge volume of information supplied by activity diaries [10]. For this reason, applications of quantitative methods were proposed. These applications treat an activity

* Corresponding author at: School of Civil Engineering and Technology, Sirindhorn International Institute of Technology, Thammasat University, 99, Paholyothin Rd., Klong Nueng, Klong Luang, Pathum Thani, 12121 Thailand. Tel.: +66 2 9869009x1911; fax: +66 2 9869009x1911.
E-mail address: mongkutp@gmail.com (M. Piantanakulchai).

http://dx.doi.org/10.1016/j.asoc.2015.04.035
1568-4946/

dataset as multidimensional data. Most conventional applications used in this research are the multivariate group identification methods, such as clustering or pattern recognition algorithms. Using cluster analysis, the activity patterns are represented in relatively small numbers of homogenous classes which are generalized based on the selected variables [11–15]. The clustering methods are promising and can be used to classify groups of individuals who share similar activity patterns. However, two arguments exist. Firstly, the methods cannot consider sequences of activities during the day. Secondly, they do not consider the information of activity locations, leaving off the information from the time-geography concept. Sequence Alignment Methods (SAMs) and Multiple-Sequence Alignment Methods (MSAMs) were proposed to solve these two arguments.

SAMs are used to determine the degree of difference between sequences of information. The methods have been widely applied in bioinformatics to analyze similarity in sequences of DNA, RNA, or protein [16,17]. Several studies in other fields such as social science [18,19], tourism [20], geography [21], and pattern recognition [22] also applied the methods to identify groups of behavior patterns according to their interest. In transport research, Wilson [10] firstly applied SAM to analyze travel behavior patterns. The study concluded that the method can successfully identify groups of behavior patterns in a population dataset. Following his study, many researchers applied SAM to analyze activity schedule data [21,23,24]. However, SAMs only consider the uni-dimensional data and fail to consider other dimensions of activity patterns such as activity locations, which also exist in a sequential pattern and directly relate to the activity schedule. This missing ability of SAM is solved by the application of MSAMs [25–27]. In addition, the motif search is also another approach to alleviate the limitations of SAMs [28]. On the other hand, to avoid the computational burden when applying SAMs, the integration of a cluster analysis method and a sequential analysis method was suggested [12,29].

Previous research on ASA showed several alternatives to derive homogenous groups of individuals who have similar activity patterns. After gathering the homogenous groups of the sampled population, further lifestyle variations are then considered, based on the person or household characteristics. In each homogenous group, the activity patterns are related to the individuals' or households' attributes and used as a response variable in the activity travel behavior model [5]. By considering only activity patterns prior to develop the models, there are possibilities that persons with different attitudes and preferences are categorized into the same homogenous group. For example, Fig. 1 presents the daily activity of two persons from Bangkok's time use data. They have different socio-demographic characteristics, but conducted similar activity patterns. ASA will recognize these two in the same homogenous groups. Hence, failure to integrate person or household characteristics in the classification process may lead to impractical models. For example, discrete choice models assume inputs from homogenous group of populations with similar socio-economic and household characteristics. However, it is impractical to calculate average values of socio-economic and household characteristics from a group classified by ASA and apply them directly to discrete choice models.

1.2. The Lifestyle classification (LSC)

In LSC, the activity analysis is defined as "a framework in which travels are analyzed as daily or multi-day patterns of behavior, related to and derived from differences in lifestyles and activity participation among the population". Lifestyles are also defined as the representation of individual preferences of his/her daily activity and travel decisions [30]. Kitamura [31] defined the term "lifestyles" occurring in the literature with two meanings: (a) activity and time-use patterns and (b) values and behavioral orientation. Hence, variables that are used to describe lifestyles depend on data availability and study objectives. These variables can be the activity characteristics, time use data, socio-demographic characteristics, and socio-economic characteristics. Additionally, there are three main approaches in the literature that were used to classify homogenous group of population with lifestyle variations [14]. The cluster analysis methods which were applied to these alternatives differ on choices of input variables. The first approach only concentrates on individual and household socio-economic and demographic characteristics, which is opposite to ASA, as discussed in Section 1.1. While ASA only concentrates on evaluations of activity schedule data, this alternative takes all considerations of lifestyle characteristics such as household structure, work participation, and household type. The second approach uses short-term daily activity-travel characteristics and long-term household socio-economic and demographic characteristics to classify lifestyles. Finally, the third approach mainly concentrates on the additional use of long-term information of individual activity participation. Among these approaches, the second approach is often used because the first approach fails to consider the activity schedule data while the third approach requires costly input data which are normally not available. More information of these approaches can be reviewed from Lin et al. [14] and Kitamura [31].

Further, the cluster analysis methods are good in classifying groups of sampled individuals but they cannot be used to allocate un-sampled individuals, the synthetic population, to the lifestyle clusters [14]. Presently, attention to this modeling capability has increased because of the growth in demand of the agent based travel demand model. To this problem, Support Vector Machine (SVM) was purposed to allocate those synthetic populations into the created lifestyle clusters [14]. Using results of the lifestyle clusters created from the cluster analysis methods, SVM can be used to develop a classification function that is only based on the socio-economic and socio-demographic characteristics. Additionally, Classification and Regression Tree (CART) is another alternative that can be used to classify group of sampled individuals with lifestyle variations and to allocate the un-sampled individuals to the previous classified group. CART is also used as the activity generator model in TRansportation ANalysis SIMulation System (TRANSIMS) [32,33].

From previous discussions on the past research in LSC, the cluster analysis methods are generally applied to create lifestyle clusters. Additionally, SVM and CART also have advantages over the cluster analysis methods by providing the activity classifier and the activity generator. These modeling approaches deliver both efficient classification results and a competent synthetic population classifier. However, as the modeling techniques are based on the cluster analysis or the classification methods, the models are unable to consider similarity of the activity sequential patterns. Even though this problem is already well recognized in ASA research by applying SAMs or MSAMs, the problem still exists in LSC research. LSC research requires variables such as age, household income, employment status, and gender. Since these variables are not in sequential format, they cannot be directly considered using SAMs and MSAMs.

1.3. Problems and needs

The advantages and disadvantages from both ASA and LSC research were already discussed in the previous sections. Three groups of algorithms are found in the literature including (a) cluster analysis, (b) Classification and Regression Tree, and (c) sequence/multiple sequence alignment method. The strength and weakness of each algorithm can be summarized as follows.