

Data Mining of Travel Surveys Using Bayesian Network Learning

Issues and Approaches

René van Hulle, Theo Arentze, Harry Timmermans

Data mining techniques are potentially useful to discover relationships in data that may be overlooked in more theory driven or parametric approaches. In this paper we consider the problem of discovering dependency relationships in travel diary data by using Bayesian Network learning algorithms. We propose and compare two strategies which consider context, situational and socio-demographic variables either simultaneously with travel-related variables or as independent variables for explaining/predicting the behavioral variables. An application of these strategies on MON data – a national travel data set from The Netherlands – illustrates the potentials and limitations of the two approaches depending on the purpose of the analysis.

1 Introduction

The dominant approach in analyzing large-scale travel surveys is to a priori decide on a model specification and then estimate its parameters. When the dependent variable refers to choice probabilities, often a discrete choice model is used [1, 2, 3]. The a priori choice of a model implies that analysts may not be fully exploring the associations that can be found in the data. Realizing that many analyses are more data-driven than one likes to believe, the question is whether more systematic exploration of travel survey data may have some potential advantages in the sense that more effort would be spent on finding associations in the data, allowing a better understanding of underlying travel behavior.

Such data exploration is the realm of data mining [4]. Although this approach has gained rapid momentum in for instance marketing, applications of data mining approaches in travel behavior research, where large activity-travel data sets are common (e.g., [5], [6]), are still scarce. For example, Keuleers *et al.* [7] used association rules to find relationships in activity diary data. Janssens *et al.* [8] applied a Bayesian belief network as an alternative to the CHAID algorithm used in Albatross [9, 10] to find decision rules underlying activity schedules. These examples indicate that existing work on data mining in travel behavior research has either searched for simple association patterns or has been motivated by an exploration of alternative representations, avoiding more systematic and full exploration of more complex causal structures in the data.

Any systematic data mining approach needs to address at least two issues: (i) what general approach should be used to guide the data mining process, and (ii) how to deal with the fact that in activity travel surveys one has information about activities, trips, tours, etc. In this paper, we will report the result of different approaches, using the recently collected MON data of the Netherlands. We describe how Bayesian Network models can be developed and used to address questions in travel-behavior research and policy making. The paper is structured as follows. First, we discuss the data and techniques involved and the different approaches in more detail. Next, we describe the results of an application highlighting how the models can be used for several analyses. Finally, we conclude the paper with discussing the major conclusions.

2 Data, techniques and approaches

2.1 MON Data

The MON data represent the latest large scale trip survey in the Netherlands. MON involves continuous data collection about mobility of the Dutch population. Questions are asked to all members of a household about personal and household characteristics and about all trips that are made during a single day. The travel survey consists of 116 variables relating to 29,221 households, 66,482 persons and 206,499 trips. The following groups of variables are included: household and personal characteristics; spatial characteristics of the household's location (e.g., province, urban density); person-day travel characteristics (e.g., number of trips conducted); trip characteristics (e.g., main transport mode) and trip-stage characteristics (e.g., travel distance, transport mode).

2.2 Bayesian networks and structure learning techniques

A Bayesian network is a network representation of the inter-relationships and conditional dependencies between a set of variables [11, 12]. Formally, it is a directed acyclic graph (a DAG) denoted by $B(G, \Theta)$, where $G = (N, A)$ is a DAG, N is a set of nodes representing a set of variables $X = \{X_1, X_2, \dots, X_n\}$, A is a set of arcs representing the directed dependency relationships between nodes, and Θ denotes a set of conditional probabilities associated with node set N and arc set A . The arcs are often interpreted in terms of cause-effect relationships (but one should realize this is still based on statistical methods) and link a variable, called child variable, with the set of its immediate predecessors, called parent variables. Conditional probabilities related to a child node specify the exact influence of the set of parent nodes and are arranged in a so-called Conditional Probability Table (CPT). The CPT of variable X_i can be written as $\theta_i = P(X_i | \Pi_i)$ where Π_i is the parent set of X_i ($\theta_i \in \Theta$ and $i = 1, 2, \dots, n$). Given the full set of conditional probabilities, standard algorithms can be used to compile the network and determine probabilities of the states of each variable representing current beliefs. Using the same algorithms, beliefs can be updated when evidence for certain nodes becomes available and is entered to the network.

The structure of the network may be based on expert knowledge. In a data mining context, however, the network is