Transportation Research Part C 17 (2009) 381–393

Contents lists available at ScienceDirect

Transportation Research Part C

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

Characterising and predicting car ownership using rough sets

Stephen D. Clark *

Institute for Transport Studies, University of Leeds, LEEDS LS2 9JT, United Kingdom

ARTICLE INFO

Article history:
Received 3 March 2008
Received in revised form 15 January 2009
Accepted 16 January 2009

Keywords:
Car ownership
National Travel Survey
Knowledge discovery
Income

ABSTRACT

This paper applies the relatively new knowledge discovery technique of Rough set analysis to identify the factors that influence the level of car ownership in a household. The study uses the detailed Great Britain National Travel Survey data set which contains information on both household and individual travel behaviour. The knowledge extraction is done using the theory of Rough sets and is presented in the form of easily understood if-then statements or rules which reveal how each attribute influences car ownership behaviour. These rules can then be used to predict household car ownership from information held about previously unseen households and the classification performance of the rules can be assessed. The performance of this classification task is shown to be on a par with other reported studies in this area.

1. Introduction

Within the transportation field there exists many informative and detailed data sets that reveal a great deal about the travel and transport behaviour of households and individuals. In many cases, however, it is this sheer volume and potential complexity of data that has meant that these data sets have not been given the scrutiny they deserve.

The primary presentational medium for these data sets are as single or cross tabulations that attempt to show how an attribute of interest varies according to the classification of another attribute, e.g. how the level of car ownership varies according to the household income or tenure. The structure of these tabulations, both in the attribute being quantified (e.g. car ownership) and the classifications (e.g. income), are in place because of some perception of the revealing nature of this structure. In practice however these tabulations are often provided for historic continuity purposes, with only the occasional innovation of a new tabulation appearing or an old one being made redundant.

A contrasting approach to this rather rigorous structure for presenting information is to allow the data to determine or suggest what significant relationships are contained within its values. This can be done in a number of ways. One common approach is to hypothesis a relationship between an attribute of interest (the dependent attribute) and a number of explanatory (independent attributes) and use standard statistical techniques to estimate the strength and legitimacy of this relationship. This parametric approach relies on certain distributional and modelling assumptions regarding the nature of the data, which can be tested for, but are commonly supposed. Another approach, and the approach adopted in this paper, is to embark on an exercise termed data mining or knowledge discovery, which makes few or no assumptions about the statistical nature of the data.

Knowledge discovery is not a single technique – the phrase acts as an umbrella term for many techniques and approaches. What they all have in common is an ability for the data to suggest the relationships it contains using as few assumptions on the nature of these relationships as possible. This information is usually contained in a series of rules that when they are

* Tel.: +44 (0) 113 343 5333; fax: +44 (0) 113 343 5334.
E-mail address: s.d.clark@its.leeds.ac.uk

0968-090X/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.trc.2009.01.006

evaluated to be true suggest a definite outcome or outcomes. These rules can be expressed in the form of if-then statements or in a tree like structure. In this tree structure the internal nodes are decision tests; branches are paths from these decisions and terminal (or leaf) nodes are the outcome (Witten and Frank, 2005). Other representations of the relationship between attributes in the data are also possible, e.g. neural networks or Bayesian networks. For this study, the knowledge is contained in the form of if-then clauses. The technique for arriving at these rules comes from the area of fuzzy set theory (Zadeh, 1965) and in particular the Rough set application of this theory (Pawlak, 1982).The characteristic of interest selected for the application of this theory is the number of motorised vehicles (cars and vans, but covered by just the term "car" here) owned by a household.

In this paper, the following section provides a short review concerning the current understanding of household car ownership behaviour and in particular the evidence for other characteristics that help to determine the level of car ownership by a household. The third section follows on by describing the Great Britain (England; Scotland and Wales) National Travel Survey data set used in this study. Section four introduces, with illustrative examples, the technique of Rough set analysis and how it can be used to generate classification rules. In the following section the two strands of data and technique are brought together to provide the insight into understanding car ownership by the application of Rough set theory. The penultimate section compares the performance of the Rough set classification technique with other classification and prediction techniques and the final section provides some concluding remarks.

2. Determinants of car ownership

Probably the two most commonly analysed areas of travel behaviour are travel mode choice and car ownership. To a large extent the two behaviours are linked in that, for example, an individual is more likely to use a car for their travel if they, or their household, own a car.

The most consistently quoted determinant of the level of car ownership is household income (Dargay, 2001 and Whelan, 2007). This could be for two reasons. The first reason is the practicality of requiring a household to have sufficient income to purchase, maintain and run a car. The second is less tangible and is linked to the acquisition of a car as a status symbol to indicate to others that you or your household has achieved sufficient wealth to own a car.

Another common determinant is the size and structure of the household (Barker and Connolly, 2005). A household with a large number of adults is not only wealthier (see the above comment on the effect of household income) but also requires more trips to be undertaken. Also in a household with children, the inconveniences and expense of using alternative modes of travel makes the ownership of a car a more attractive proposition.

The availability of alternative modes of travel may also have an impact on car ownership decisions (Dargay, 2002). In densely populated urban areas there will usually be alternative motorised modes of travel available (buses or trains) or services such as shops or schools are located close enough to permit walking and cycling as viable alternatives to car travel. In more rural areas these alternative modes may not be available and the services people wish to use may be located some distance away from the household.

3. National Travel Survey

The National Travel Survey (Department for Transport, 2007) is a continuous household travel diary survey commissioned by the United Kingdom Department for Transport. Initially the survey was commissioned on an adhoc basis, but since 1988 it has been a continuous survey and, in 2001 the set sample size of the survey was tripled from around 5800 households per year to 15,000. The survey is conducted using a clustered, stratified approach (by using region, car ownership and population density information) to select the localities. Within each locality, households are selected at random to participate in a seven day long travel diary survey. The survey is able to provide reliable estimates of travel behaviour and other transport characteristics at the national; regional and, in certain instances, sub-regional geographies.

3.1. Structure of the survey

The data is hierarchical in nature and Fig. 1 illustrates this structure. The top level is the primary sampling unit (PSU) which is the sampling cluster and is based on the UK postcode geography. The PSUs to sample within are selected to ensure that the survey returns are representative (e.g. London addresses are oversampled because the eventual response rates from such addresses is poorer than elsewhere) but there is also a quota effect where, each year half the PSU's are retained for sampling next year. No PSU can, however, be carried forward for more than one year. The information about the PSU is supplied by the survey organisation and provides generic information about the nature of the locality, e.g. population density and what type of concessionary fares are available.

Within each PSU a number of households are selected for inclusion in the survey and a letter is sent to the household to inform them that an interviewer will be visiting them. The selected households are then visited by an interviewer who briefs the members of the household on the purpose of the survey and its requirements. During this initial visit by the interviewer information is collected on the household; the individuals in the household and the characteristics of cars owned or used by the household. These interviews should ideally be face-to-face but information supplied by an adult on another member of