Travel Behaviour and Society 13 (2018) 105–117

Contents lists available at ScienceDirect

Travel Behaviour and Society

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

Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam

Danique Ton[a,*], Dorine Duives[a], Oded Cats[a], Serge Hoogendoorn[a]

[a] Department of Transport & Planning, Delft University of Technology, Stevinweg 1, PO Box 5048, 2600 GA Delft, The Netherlands

ARTICLE INFO

Keywords:
Data-driven choice set generation
Cyclists' route choice
Travel behaviour analysis comparison
Breadth-first search on link elimination
Labelling

ABSTRACT

Specifying the choice set for travel behaviour analysis is a non-trivial task. Its size and composition are known to influence the results of model estimation and prediction. Most studies specify the choice set using choice set generation algorithms. These methods can introduce two types of errors to the specified choice set: false negative (not generating observed routes) and false positive (including irrelevant routes). Due to increased availability of revealed preference data, like GPS, it is now possible to identify the choice set using a data-driven approach. The data-driven path identification approach (DDPI) combines all unique routes that are observed for one origin-destination pair into a choice set. This paper evaluates this DDPI approach by comparing it to two commonly used choice set generation methods (breadth-first search on link elimination and labelling). The evaluation considers the three main purposes of choice sets: analysis of alternatives in the choice set, model estimation and prediction. The conclusion is that the DDPI approach is a useful addition to the current choice set identification methods. The findings indicate that in analysing alternatives in the choice set, the DDPI approach is most suitable, as it reflects the observed behaviour. For model estimation the DDPI approach provides a useful addition to the current choice set generation methods, as it provides insights into the preferences of individuals without requiring network-data for additional information or generating routes. In terms of prediction, the DDPI approach is not suitable, as it is not able to perform well with out-of-sample data.

1. Introduction

In the context of travel behaviour, many choices must be made by an individual before a trip is made, e.g. destination, mode and route choice. These choices are all discrete in nature, meaning that only one option can be chosen at a time. The choice set from which an individual chooses one, forms an important aspect in the analysis of travel behaviour. Three different purposes of choice sets can be identified. First, it is essential in analysing different travel options in the network (e.g. number of alternatives, characteristics or composition of the alternatives), second it is used for demand model estimation (estimating behavioural parameters), and third it is instrumental in predicting choice probabilities and thereof flow distribution over alternatives/the network (Bovy, 2009). The size and composition of the choice set influence the results of the model estimation and prediction, and consequently the interpretation of the estimated behavioural parameters (Bovy, 2009). This issue is for example relevant in route choice analysis, as many possible alternatives can be identified by the researcher, but only few will be known to the individual, leading to possible mismatches in the choice set identification.

Route choice sets are often specified using *choice set generation algorithms* (e.g. k-shortest paths or labelling), which compute a set of routes based on characteristics of the network(-links) (e.g. distance or travel time). The use of these algorithms can introduce two types of errors in the choice set: false negative and false positive errors. False negative errors arise when the algorithm is not able to reproduce the chosen alternatives. The generated alternatives might not match the behaviour and preferences of the individual, and as a result the chosen route is not reproduced. The impact of this error decreases when the ability of the choice set generation algorithm to capture the individuals' behaviour and preferences increases. False positive errors occur when a choice set generation algorithm also generates routes that are not considered by the individual, resulting in a too large choice set. In conclusion, the use of choice set generation algorithms potentially comes with several flaws.

In recent years, large improvements have been made in revealed preference data collection methods. New data sources, such as GPS data that contain detailed spatial and temporal information on the movement pattern of individuals, help creating insights into the individuals' choice behaviour. By combining the GPS records belonging to one

* Corresponding author.
E-mail addresses: d.ton@tudelft.nl (D. Ton), d.c.duives@tudelft.nl (D. Duives), o.cats@tudelft.nl (O. Cats), s.p.hoogendoorn@tudelft.nl (S. Hoogendoorn).

https://doi.org/10.1016/j.tbs.2018.07.001
Received 28 February 2018; Received in revised form 13 June 2018; Accepted 10 July 2018
Available online 29 July 2018

individual into separate trips, the observed trips can be used for route choice research (e.g. Menghini et al., 2010; Hood et al., 2011). Next to generating the choice set based on a set of assumptions on network properties, it is then also possible to use the observed trips from GPS data to identify the choice set directly. Every trip between an origin and destination follows a certain route, the unique routes that are observed can then be combined into one choice set. Consequently, the potential false negative error associated with choice set generation algorithms cannot occur and the potential false positive error is negligible because all the routes included in the choice set have been chosen by the individual.

Governments worldwide have shown increasing interest in promoting and understanding cycling usage, due to the potential health, congestion and emissions benefits. Consequently, goals have been set to increase the cycling modal share (Pan-European Programme, 2014). Several studies investigated bicycle route choice using GPS data, primarily in areas where cycling is relatively scarce, with the goal of identifying determinants that influence route choice, so that substantiated infrastructure investments can be made (Menghini et al., 2010; Hood et al., 2011; Broach et al., 2012; Casello and Usyukov, 2014; Montini et al., 2017; Zimmerman et al., 2017; Chen et al., 2017; Li et al., 2017; Ghanayim and Bekhor, 2018). Other studies have taken in place in urban environments with a larger share of cyclists, like Copenhagen (Halldórsdóttir et al., 2014; Prato et al., 2018; Skov-Petersen et al., 2018). These studies have applied different types of choice set generation algorithms, such as labelling, stochastic methods, link elimination, and link penalty. However, none of the studies has applied a data-driven method for choice-set identification as proposed and examined in this study. This approach is applied to a bicycle route choice study for the city of Amsterdam, the Netherlands (Ton et al., 2017). Amsterdam is known for its well developed bicycle infrastructure and high share of bicycling trips (37%) (OViN, 2011). To evaluate the potential of this data-driven method for choice set identification, we compare the method using the dataset from the city of Amsterdam, to other choice set generation algorithms previously applied in the cycling route choice literature.

This paper evaluates the use of a data-driven approach for choice set identification in travel behaviour analysis. The goal is to investigate whether a data-driven approach can be a valuable addition to the current choice set identification methods. Bicycle GPS data from Amsterdam, the Netherlands, is used to identify the choice set and this choice set is used in the estimation and validation of a route choice model. The evaluation of the data-driven approach is done by means of a comparison study, where it is compared to two commonly used choice set generation methods, to assess and compare their performance and results. Based on computation time, sensitivity to false negative errors and, number of applications, two approaches have been selected: the breadth-first search on link elimination (BFS-LE) introduced by Rieser-Schussler et al. (2013) and the labelling approach introduced by Ben-Akiva et al. (1984). The evaluation is performed on the three abovementioned purposes of choice sets; 1) analysing the composition of the choice set, 2) understanding behaviour (model estimation) and 3) application of the model on out-of-sample data (model validation).

The rest of the paper is outlined as follows. Section 2 reviews contemporary choice set generation procedures. In Section 3, the data-driven approach is elaborated upon in terms of requirements of data, opportunities, limitations of the method, and sensitivity with respect to data collection duration. Section 4 describes the methodology for evaluating the specified choice sets as well as the route choice model estimation and validation. Section 5 provides background on the data that was collected and prepared for this study. Section 6, then details the evaluation of the generated choice sets in comparison to the observed routes and Section 7 covers the evaluation of the route choice model estimation and validation. Finally, Section 8 concludes the paper and provides directions for future research.

2. Choice set generation methods

This section discusses different choice set generation methods that have been proposed in the past and selects two methods as reference for the evaluation of the data-driven approach.

Many different methods have been proposed for identifying route choice sets (for detailed reviews see Fiorenzo-Catalano (2007) and Ramming (2002)). Bovy (2009) and Prato (2009) identify four categories of choice set generation methods: deterministic methods, stochastic methods, probabilistic methods and constrained enumeration methods. Most choice set generation methods belong to the *deterministic category* and consist of repeated shortest path searches in the network. These shortest path methods have different input variables such as search criteria, route constraints and link impedance (Prato, 2009). They are computationally attractive due to the efficiency of shortest path algorithms. *Stochastic methods* are also based on repeated shortest path searches, but additionally the computation of optimal paths is randomised based on link impedances or individual preferences drawn from probability distributions, mostly done using simulation. These methods have been applied in the bicycle route choice context by Hood et al. (2011), Halldórsdóttir et al. (2014), Ghanayim and Bekhor (2018), and Prato et al. (2018). *Constrained enumeration methods* are not only based on shortest routes, but also make additional behavioural assumptions (Prato, 2009). These assumptions reflect different behavioural thresholds that can be specified, e.g. excluding loops and only including links that bring the individual closer to the destination. These methods have been applied in the bicycle route choice context by Halldórsdóttir et al. (2014), but did not prove to outperform the deterministic or stochastic methods. *Probabilistic methods* assign a probability for each alternative to be included in the choice set. A fully probabilistic approach, as proposed by Mansky (1977), which includes the choice set generation and selection in the utility function, is often deemed infeasible due to its computational complexity. As a consequence, these methods have not yet been applied in the bicycle route choice context.

Recently, two alternative approaches have been proposed that address the choice set identification implicitly (i.e. no need for explicit enumeration of alternatives). The first is the *sampling approach* (Frejinger et al., 2009; Flötteröd and Bierlaire, 2013), that assumes a universal choice set and by means of importance sampling selects a subset of these routes. The second approach is the *link-based approach* (Fosgerau et al., 2013), which assumes that individuals make successive choices at each node. The link-based approach was applied in the bicycle route choice context by Zimmerman et al. (2017).

Due to its prevalence in the general and bicycle route choice literature, computational efficiency and deterministic nature (which relates more to the cognitive aspects of the decision-maker rather than conceived a computational instrument), *deterministic methods* are selected as reference methods for comparison in this study. Four categories of deterministic methods are identified: shortest paths, link elimination, labelling and link penalty. Previous findings suggest that the *shortest path methods* have the lowest performance in terms of reproducing the observed routes (Bovy, 2009). Furthermore, the *link penalty methods* are known for their large computation times (Bekhor et al. 2006). Therefore, the focus lies with the link elimination and labelling methods.

The *link elimination method* iteratively removes links that are on the shortest path and finds new shortest paths (Bellman and Kalaba, 1968). Prato and Bekhor (2007), Bekhor et al. (2006), and Ghanayim and Bekhor (2018) evaluated this approach and found that in about 40% of the cases false negatives are produced. Azevedo et al. (1993) proposed an alternative approach, where the entire shortest path is eliminated, after which a new shortest path is calculated. This approach is more drastic, as it eliminates overlap but can result in an unrealistic choice set (e.g. large detours). Rieser-Schussler et al. (2013) adapted the link elimination method by applying a breadth-first search technique on link