

## Calibrating Route Choice Sets for an Urban Public Transport Network using Smart Card Data

Shelat, Sanmay; Cats, Oded; Oort, Niels Van; Van Lint, Hans

**DOI**

[10.1109/MTITS.2019.8883366](https://doi.org/10.1109/MTITS.2019.8883366)

**Publication date**

2019

**Document Version**

Accepted author manuscript

**Published in**

MT-ITS 2019 - 6th International Conference on Models and Technologies for Intelligent Transportation Systems

**Citation (APA)**

Shelat, S., Cats, O., Oort, N. V., & Van Lint, H. (2019). Calibrating Route Choice Sets for an Urban Public Transport Network using Smart Card Data. In *MT-ITS 2019 - 6th International Conference on Models and Technologies for Intelligent Transportation Systems* Article 8883366 IEEE.  
<https://doi.org/10.1109/MTITS.2019.8883366>

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

# Calibrating Route Choice Sets for an Urban Public Transport Network using Smart Card Data

Sanmay Shelat, Oded Cats, Niels van Oort, Hans van Lint  
*Dept. of Transport and Planning*  
*Delft University of Technology*  
Delft, The Netherlands

S.Shelat@tudelft.nl, O.Cats@tudelft.nl, N.vanOort@tudelft.nl, J.W.C.vanLint@tudelft.nl

**Abstract**—Identifying the set of alternatives from which travellers choose their routes is a crucial step in estimation and application of route choice models. These models are necessary for the prediction of network flows that are vital for the planning of public transport networks. However, choice set identification is typically difficult because while selected routes are observed, those considered are not. Approaches proposed in literature are not completely satisfactory, either lacking transferability across networks (observation-driven methods) or requiring strong assumptions regarding traveller behaviour (uncalibrated choice set generation methodologies (CSGM)). Therefore, this study proposes a constrained enumeration CSGM that applies the non-compensatory decision model, elimination-by-aspects, for choice set formation. Subjective assumptions of traveller preferences are avoided by calibrating the decision model using observed route choice behaviour from smart card data, which is becoming increasingly available in public transport systems around the world. The calibration procedure also returns two key insights regarding choice set formation behaviour: (i) the ranking of different attributes by their importance, and (ii) the acceptable detours for each attribute. To demonstrate the methodology and investigate choice set formation behaviour, the tram and bus networks of The Hague, Netherlands are used as a case study.

**Keywords**—route choice, choice set generation, smart card, public transportation, The Netherlands

## I. INTRODUCTION

There is widespread agreement in the marketing field that consumption choices occur in a two-stage process whereby consumers first form a consideration choice set and then make the final choice from this set [1]. Choice set composition and size can affect the ultimate decision in a number of ways [2], with the most obvious being that the exclusion of an alternative from the choice set means that it cannot be selected. Identification of choice sets may be straightforward when the number of alternatives are limited but it becomes more difficult as this number increases. In these cases, correctly identifying the choice set is important not only for real-world application of estimated choice models but also for the estimation of choice models from revealed preferences where choices are observed but choice sets are not.

Public transportation provides vital, sustainable transportation in many regions, making their planning, maintenance and operation a priority for authorities. In order to provide an appropriate level of service, understanding traveller behaviour to correctly model network flows has become increasingly important. Amongst other traveller decisions, route choices have a significant impact on network flows. Therefore, for both, estimation and application of route choice models, identification of route choice sets is a

crucial step [2].

However, identifying route choice sets for origin-destination (OD) pairs in a network is a non-trivial task for several reasons. First, due to the combinatorial nature of the problem, the number of available and attractive routes is usually large. Second, public transport characteristics, such as fixed routes, schedules, and headways, which are usually time-dependent, add to the complexity of the task. Finally, the existence of different forms of travel costs, for instance, transferring or in-vehicle time, mean that traveller preferences have to be taken into account when identifying route choice sets.

Given the importance and complexity of route choice set identification, several studies in transportation literature have either entirely focussed on or have employed some form of route choice set identification methodology. These methodologies can be broadly divided into: (i) direct identification of choice sets and (ii) choice set generation methodologies (CSGMs).

Direct identification of choice sets may be based on reporting or observations of non-selected and selected alternatives, respectively. In the former, surveyed travellers are asked to report alternatives to their chosen route that they did not select but considered. This method has the obvious advantage that researchers do not have to guess what travellers have in their mind and the consideration set is known at the individual level. However, this reporting is subject to a number of errors (e.g., forgetfulness) and, as suggested in [3], is ‘at best a subset of the true choice set’. Furthermore, such interview techniques are time consuming and difficult to implement when choice sets are required for network-wide analysis.

For network-wide identification of choice sets, observations of selected alternatives offer a more suitable data source. In this method, the sets of unique routes observed are assumed to be the choice set for the respective OD pairs. The argument is that, if such data is collected over a long period of time it should include all routes considered by travellers. Practically, this is facilitated by the creation of large data sources as an increasing number of public transport services turn to automatic fare collection (AFC) technologies. As a result, several studies using smart card data employ this method for the identification of choice sets. However, this technique precludes the possibility of taking into account why some non-selected but feasible routes are never chosen [4]. Moreover, the transferability of behaviour parameters, estimated with choice sets thus obtained, is precarious because matching choice set generation methodologies are not available for other public transport networks [5].

Some drawbacks of direct identification of route choice sets can be overcome by using CSGMs. The aim with this approach is to develop a generic algorithm, that satisfies requirements associated with the purpose of the choice set [6], for identification of route alternatives. Thus, route CSGMs are suitable for network-wide application and by nature more transferable than direct identification techniques. These methods are typically classified into: (i) deterministic and stochastic shortest path, (ii) constrained enumeration, and (iii) probabilistic approaches [6, 7]. Although, a complete literature review is out of scope (see [6, 7]); below, the most important approaches are discussed and the comparison of their performance is reviewed.

Shortest path-based methodologies, which compose the largest group of models, search for optimal routes in the network and assume them to be the route choice set. Variations are based on the link impedances optimized, route constraints, and other search criteria [6, 7]. Approaches in this category that are based on either purely topological criteria or use only travel time have the drawback that the choice sets do not reflect traveller preferences. On the other hand, methods that do have some degree of behavioural sophistication, such as the link-labelling approach [8], are criticised for their dependence on analyst judgments to make assumptions regarding traveller behaviour for the definition of objective functions [6, 7, 9]. Furthermore, shortest path methods tend to produce more homogenous routes and are, therefore, typically unable to reproduce all observed routes.

Unlike the above approaches, constrained enumeration methodologies are based on rules other than minimum cost paths. Since these methods aim to generate all possible routes between OD pairs whilst being constrained by some rules, they usually perform better in terms of reproducing observed routes than the shortest path CSGMs [6]. Constraints used to reduce the number of irrelevant routes generated may be based on logic or common sense, feasibility, degree of choice set heterogeneity, or behavioural preferences [3, 10]. The disadvantages of this approach include the high computational effort required for route enumeration and the fact that, here too, the method depends on the definition of behavioural constraints which have been typically based on the expertise of analysts. Despite this drawback, in a comparison of various (uncalibrated) route CSGMs, a branch-and-bound based enumeration with threshold-based behavioural constraints performed better than other shortest path approaches on all the validation criteria considered [10, 11].

From the common disadvantages of the above approaches, it is clear that calibration of behavioural parameters is an important aspect of route CSGMs. Yet, while studies often validate their models against observed (selected) or reported (non-selected) route alternatives, calibration is rarely performed. A Scopus search<sup>1</sup> for studies that perform such calibration returned only five relevant studies, including two studies that considered public transport modes [12, 13]. The latter studies use trial-and-error methods to calibrate their models on the basis of

analyst judgments and, observed and reported route alternatives. However, a shortcoming of these studies is that sample sizes of the data used are relatively small in comparison to the networks considered (which may be at least in part due to data collection difficulties).

Given the importance of identifying route choice sets in public transport networks and the drawbacks of existing studies, we propose a methodology that adopts an intuitive and accepted behavioural model of choice set formation, and includes calibration of parameters of the same using smart card data. The proposed CSGM takes a constrained enumeration approach similar to those used (and proven to perform well) in [3, 10]. However, the methodology developed here avoids (almost completely) the need for any subjective assumptions regarding traveller preferences by delaying the application of behavioural constraints until after all logical and feasible routes have been generated. Instead of assumptions, behavioural constraints are directly obtained from AFC data, the increasing availability of which makes it possible to more easily collect network-wide route choice observations. Moreover, the constraints, which are based on a non-compensatory decision model, offer an intuitive insight into travellers' choice set formation preferences (section II). To demonstrate the methodology, it is applied to the urban public transport network of The Hague, Netherlands.

In the next section, the behavioural model used for choice set formation is discussed. Section III describes the choice set generation methodology which is applied to the urban public transport network of The Hague, Netherlands in section IV. Section V concludes the paper with a summary of the contributions, key insights from the case study, and potential paths for future research.

## II. CHOICE SET FORMATION BEHAVIOURAL MODEL

When a large number of alternatives are involved, consumers are likely to apply heuristic decision rules, rather than perform a comprehensive evaluation, when forming their considered choice set. These choice set formation heuristics are usually more reasonable because of the relatively high (cognitive and explicit) costs of complete evaluations [1]. Therefore, since the number of route alternatives available in transportation networks is typically large, travellers can be reasonably expected to use such heuristics to identify their choice sets [6, 7].

While complete evaluations are typically compensatory in nature, heuristics involve non-compensatory decision rules. Compensatory models take into account trade-offs between alternative attributes whereas non-compensatory models only apply constraints on individual attributes. A number of non-compensatory decision models have been proposed in literature, such as: (i) disjunctive, (ii) conjunctive, (ii) lexicographic, and (iv) elimination-by-aspects [14]. Some of these have been used in the route choice set generation literature.

Disjunctive and conjunctive rules both set minimum thresholds for all important attributes. The former accepts alternatives that comply with at least one requirement while the latter needs all attribute thresholds to be met. Most route CSGMs that apply detour thresholds to different attributes (e.g., [3, 10]) are applying conjunctive rules. In these studies,

<sup>1</sup>Search term: ( TITLE-ABS-KEY ( calibr\* ) AND TITLE-ABS-KEY ( ( route OR path ) AND TITLE-ABS-KEY ( ( "choice set\*" OR "consideration set\*" ) ) ); Access date: 8 February 2019

thresholds are usually set as multiplicative factors (greater than one but not necessarily integers) of the attribute value of the best performing alternative (for that particular attribute).

Under lexicographic decision-making, first, attributes are ranked by importance; then alternatives are selected on the basis of their performance of the top-ranking attribute. In case of a tie, the performance on the second-best attribute is checked, and so on. Since this method does not set thresholds, desired choice set sizes need to be defined for their formation (lexicographic and conjunctive decision rules become the same if choice set size is defined for both) [1]. The link-labelling route CSGM [8], which assumes that travellers optimize paths for different attributes, is an example of this category.

Elimination-by-aspects (EBA) combines attribute ranking and setting of thresholds. Although the original version [15] was proposed as a probabilistic rule, most applications for choice set formation use a deterministic version [1]. For the choice set formation, first the most important attribute is selected and alternatives that do not meet its threshold are eliminated. This is repeated until all attribute thresholds have been checked although in another version, elimination stops once the required choice set size has been achieved [1]. Based on the literature review conducted here, no route CSGM could be found that uses this behavioural model. A possible reason could be that in the absence of a calibration method, because this model combines ranking and setting thresholds, researchers are required to make more assumptions regarding traveller behaviour.

This study assumes deterministic EBA as the behavioural model for route choice set formation. In the version employed here, no assumptions are made regarding choice set size. Deterministic EBA implies that the choice set for an OD pair at a given time is the same for all travellers. Attribute values are obtained from the general transit feed specification (GTFS) data. Therefore, attributes included in the process are limited to those observable in this data.

The output of the methodology proposed here are route choice sets per OD pairs and time periods. Each alternative in the route choice set is defined uniquely by the sequence of alighting stations and the common lines (lines passing through the same sequence of stations) connecting the respective stations. Although common lines are thus accounted for, issues concerning partial route overlap are assumed to be handled at the next stage of choice modelling. In addition to the route choice sets, calibration of the choice set formation model returns two insights regarding traveller behaviour: (i) the importance ranking of attributes and (ii) the acceptable detour threshold for each attribute.

### III. ROUTE CHOICE SET GENERATION METHODOLOGY

To give structure to the complexity of route choice set identification, a hierarchy of route choice sets (for a given OD pair and time period) is proposed in [16] and presented from traveller and researcher perspectives in [3]. Similar to those, for the methodology presented here, the following hierarchy is used (Fig. 1, right hand side): (i) complete network containing the universal set of all possible paths from origins to destinations, (ii) logical routes per OD pair, (iii) feasible routes per OD pair for different times (OD-T), (iv) considered routes per OD-T, and finally (v) chosen routes. Here, the consideration route set is obtained from the

generated-feasible and observed route sets. The following sub-sections describe the steps in the proposed methodology (Fig. 1, centre) that take some inputs and produce the desired outputs (Fig. 1, left hand side), by progressively moving down the hierarchy.

#### A. Inputs

Two main data sources are required for the route CSGM proposed here: (i) GTFS and (ii) AFC. GTFS data contains information regarding the service layer of the network and its properties. These define public transport lines connecting different sequences of stops in the network, the in-vehicle travel time (time taken by a vehicle) between OD pairs, and the frequencies (vehicle departures per hour) of each line. Although, a frequency-based system is assumed in this CSGM, line frequencies are allowed to be time-dependent. AFC data is used to generate the set of selected route alternatives. Ideally, for each observation, the data should contain information regarding the sequence of stops (i.e., origin, transfers, and destination), the lines used between each stop, and boarding times. It should be noted, however, that for data from nearly all AFC systems, at the least, transfer inference will be required.

In addition to the above data sources, rules regarding which routes are logical, and which aspects and thresholds are considered by the calibration process are also inputs to the methodology. However, in the current implementation, these inputs are defined as part of the methodology.

#### B. Network Representation

Proper network representation is key to the computational efficiency of route generation. For the topological constrained enumeration methodology used in this study, the public transport network is represented in the

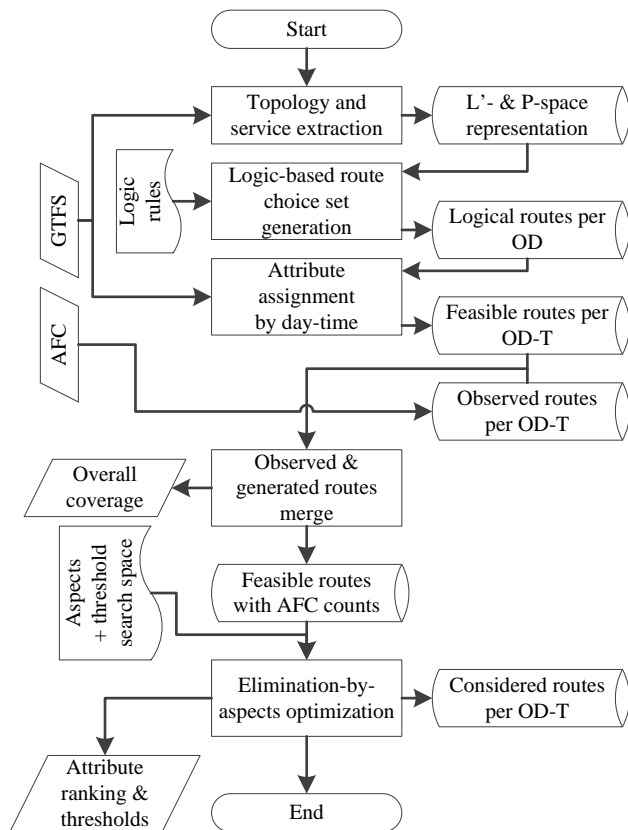


Fig. 1. Overall methodology

P-space which explicitly represents the service layer: nodes are stops while links are (groups of) public transport lines that provide direct (transfer-less) connections to other stops. Although urban public transport networks may use schedules, this study assumes a frequency-based system. Thus, time is not included in the graph representing the network.

Stops in the representation are defined by the ‘parent station’ field in the GTFS data. Moreover, different lines are grouped together as one connection in P-space if they pass through the same sequence of stops, that is, they are common lines. Each cell in the P-space adjacency matrix contains information about the connections between the origin and destination stops. For each connection, this information consists of the common lines and the stops they traverse (pass without alighting) through for this connection.

Since the generation methodology also considers transfers which require walking to another stop, walkable links are stored as an binary adjacency matrix of all stops. To avoid generating too many irrelevant alternatives, a conservative threshold of 200 Euclidean metres is set as the maximum acceptable walking distance.

### C. Constrained Enumeration of Logical Routes

#### 1) Enumeration

The enumeration methodology applied here uses a one-to-all, breadth-first search algorithm, similar to that used in [17]. The methodology is applied to the P-space graph representation of the network as defined above.

First, a stop is selected as the origin and the vertex root of the search tree. The stops it is directly connected to in the P-space graph of the network become destinations; this is the first level of the search tree. The information contained in the connections (lines and traversed stops) are stored for the respective OD pairs. For the next level of the search tree, the following become intermediate origins: transfer stops (stops connected by more than one line) amongst the neighbouring stops and stops accessible by walking from the neighbouring stops. Then the stops directly connected to these intermediate origins become destinations and connection information is stored, retrospectively from the origin stop (vertex root), for the respective OD pairs. For the next level, intermediate origins are selected in the same way as above, and the process is repeated up to a desired depth of the search tree. This way, all route alternatives between the origin stop and others are generated and stored. The procedure can then be repeated with another stop as the origin.

#### 2) Constraints

Obviously, simply enumerating this universal set of routes would be unending. To prune the search tree, the depth is constrained by assuming that travellers accept a maximum of two transfers. This behavioural assumption should be reasonable for most urban public transport networks. Additionally, to ensure that only logical routes are produced, two rules are used as breadth-wise constraints to the enumeration. (i) No loops – traversing through, alighting at, or walking to stops previously traversed through or boarded from is not allowed. (ii) No transferring between common lines – alighting at a stop which is connected by the same set of lines as the previous connection is not allowed. Since travellers may want to shift time spent waiting for a particular line downstream, transferring to stops with a subset of the previous connection’s lines is permitted. In this

case, this subset of lines is removed from the previous connection to ensure that transfers do not occur between the same lines. In the current implementation, it is assumed that travellers do not shift their waiting times by walking to another stop, hence, walking to a stop connected by a subset of the lines as the previous connection is not allowed. These logical constraints are only applied after the first level of the search tree.

### D. Attribute Assignment for Feasible Routes

In this step, route alternatives, that were generated from an unweighted graph, are assigned attribute values. This is required to remove infeasible routes as well as for the consideration set formation in section III.F.

#### 1) Attribute Values

The following route attributes are observable from the GTFS data, and therefore included in the study: (i) waiting time, (ii) in-vehicle time, and (iii) number of transfers. Currently, the waiting and in-vehicle times over different legs of the route are not considered separately and only the total values are used.

Expected waiting time for connections between two stops is calculated as the inverse of the sum of frequencies of the connecting lines, implicitly assuming them to be evenly spaced as well as assuming uniform arrivals of travellers at stops. The time-dependent nature of public transport line frequencies are taken into account by assigning them separately for each hour of the day in weekdays and weekends, respectively. Routes that become infeasible at a certain time (because a link has zero frequency) are eliminated from the feasible choice set for the respective time period. Values of the other attributes are time-independent. Although there might be small time-dependent fluctuations in the planned in-vehicle times, they are ignored for the sake of computational efficiency.

As discussed in section II, the consideration set formation model employed here uses the EBA behavioural model, which requires setting threshold constraints to different attributes. These thresholds are some factors of the attribute values of the alternatives (in the same time period) that perform best on the respective attributes. In preparation for the calibration step, these factors are calculated for each attribute in all the alternatives. Since waiting and in-vehicle times are more continuous in nature, multiplicative factors are employed, whereas for number of transfers an additive factor is used.

#### 2) Dominated Alternatives

Once attribute values have been assigned, alternatives that are state-wise dominated, that is, perform worse on all attributes, by others (in the same time period) are removed. It is rarely disputed that choosing such a dominated alternative is irrational. Although the existence of dominated alternatives in the choice set may have a decoy effect (see [18]), such effects are rarely modelled in the route choice context.

### E. Merge Generated and Observed Routes

The calibration uses generated-feasible routes as well as those observed from AFC data. This step merges these two route sets on the basis of the sequence of stops boarded, lines used, the hour of the first boarding, and the final destination stop. For the calibration, only those observed routes that were also generated are considered. Given that the

constraints assumed during route enumeration are not very restrictive, discarding observations that are not generated should not affect the final calibration too much. In case, the overall coverage does turn out to be low, it may make sense to check the AFC data for issues such as improper transfer inference.

#### F. Calibration using Elimination-by-Aspects

In EBA, travellers are assumed to rank and set threshold cut-offs for attributes. In order to deduce these preferences, the generated feasible route alternatives may be compared with the observed ones. For such a comparison, two indicators are commonly used in literature, albeit for validation purposes rather than calibration: (i) coverage – the proportion of observed routes that have been generated, and (ii) efficiency – the proportion of generated routes that are observed.

With respect to calibration, clearly, the likeliest combination of choice set formation preferences is one that maximize both coverage and efficiency; that is, reproduces as many observed routes as possible while not generating too many irrelevant alternatives. Thus, to derive EBA preferences, an optimization problem that maximizes these indicators is setup. First, however, small modifications to the above indicators are proposed.

##### 1) Indicators

In their simplest form, coverage and efficiency do not take into account demand across OD pairs and weigh each route alternative as the same. For example, if there is an OD pair with only one trip, it would still have an effect on the choice set calibration even though there is little behaviour to be observed. To this end, the coverage indicator is modified by simply adding demand weights per route. Efficiency is changed more fundamentally by making it a proportion of routes not observed (but in the generated feasible choice set), rather than a proportion of generated routes, to avoid asymmetric demand weighting in the definition. These indicators are defined below.

Let  $N$  be the set of stops in the network under consideration; and  $R_{ij}^f$  the set of generated-feasible between OD pairs  $i, j \in N$ ,  $R_{ij}^o$  the set of observed routes therein ( $R_{ij}^o \subseteq R_{ij}^f$ ), and  $R_{ij}^c$  be the calibrated choice set, such that  $R_{ij}^c \subseteq R_{ij}^f$ , for a given combination of EBA preferences. Then, Fig. 2 gives the four possible sets (and notations) of route alternatives that result when comparing the observed and calibrated choice sets. Finally, let  $q_{ij}$  be the total of all demand on routes  $R_{ij}^o$ , and  $q_{ij}^{TP}, q_{ij}^{FN}$  be the total demand for route sets  $R_{ij}^{TP}, R_{ij}^{FN}$ , respectively. Then, coverage and efficiency are defined in this study as:

$$coverage = \frac{\sum_{i,j} q_{ij}^{TP}}{\sum_{i,j} q_{ij}^{TP} + q_{ij}^{FN}}, \text{ and} \quad (1)$$

$$efficiency = \frac{\sum_{i,j} |R_{ij}^{TN}| q_{ij}}{\sum_{i,j} (|R_{ij}^{FP}| + |R_{ij}^{TN}|) q_{ij}}, \quad (2)$$

where  $|\cdot|$  denotes set size. Then, to achieve a balance between coverage and efficiency, the following optimization indicator is minimized for each attribute:

$$x_a = \text{abs}(coverage_a - efficiency_a). \quad (3)$$

#### 2) Algorithm

The EBA based analysis conducted here considers only a few aspects (i.e., attributes). Moreover, it is reasonable to expect that the potential thresholds are close to the respective smallest values (i.e., 1 for waiting time and in-vehicle time ratios, and 0 for difference in number of transfers). Therefore, to deduce EBA preferences, a brute force algorithm may be feasibly employed. The algorithm to calculate indicator values for different attribute rankings (Fig. 3) works as follows: all possible attribute permutations are listed; for a given permutation, different thresholds from the pre-defined search space are tried to find the minimum indicator value for the first attribute; before repeating this for the next attribute, routes that do not comply with the previously found threshold(s) are eliminated; this is repeated until all attribute thresholds (and indicator values) for the permutation have been found; and the process is repeated for the next permutation.

It should be noted that a key difference from other threshold based CSGMs is the sequential elimination of routes. Thus, for each permutation we have a number of optimization indicator values associated with each attribute

		In calibrated choice set?	
		Yes	No
In observed choice set?	Yes	True positives $R_{ij}^{TP} = R_{ij}^c \cap R_{ij}^o$	False negatives $R_{ij}^{FN} = \{R_{ij}^o \cup R_{ij}^c\} - R_{ij}^c$
	No	False positives $R_{ij}^{FP} = \{R_{ij}^c \cup R_{ij}^o\} - R_{ij}^o$	True negatives $R_{ij}^{TN} = R_{ij}^f - \{R_{ij}^c \cup R_{ij}^o\}$

Fig. 2. Comparison between calibrated and observed choice sets

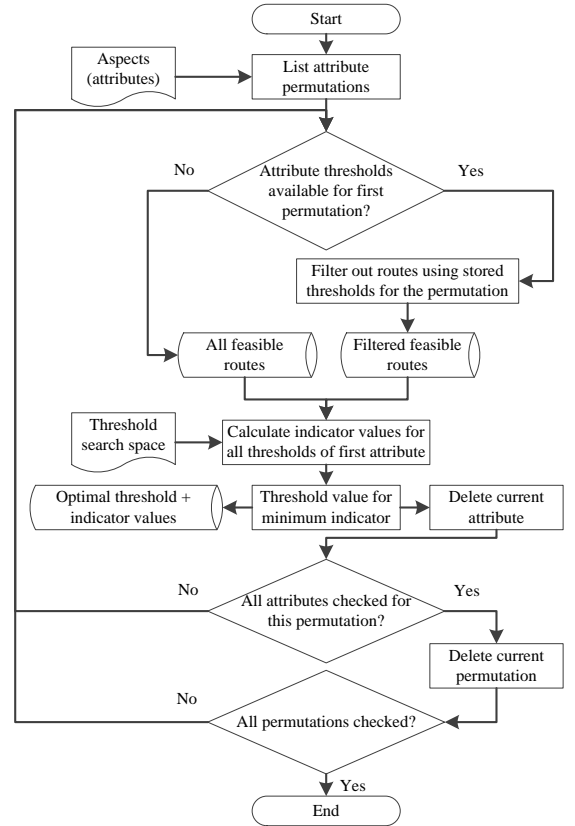


Fig. 3. Elimination-by-aspects brute force calculation of indicator values for different attribute rankings

in it. The performance of a permutation is assessed by calculating the natural logarithm of the product of attribute optimization indicator values in that permutation:

$$x^p = \ln\left(\prod_a x_a^p\right). \quad (4)$$

Since the optimization indicator has to be minimized, the permutation with the lowest value is considered optimal.

#### IV. CASE STUDY: THE HAGUE TRAM & BUS NETWORK

##### A. Setup

To demonstrate the above methodology and obtain attribute ordering and threshold preferences of travellers, the urban public transport network of The Hague, Netherlands (Fig. 4) is used as a case study. The network consists of both tram and bus lines which mainly serve The Hague but also connect to the neighbouring cities of Zoetermeer and Delft. The case study uses smart card data from March 2015 and the corresponding GTFS data for the analysis. The network then consisted of 12 tram and 8 bus bidirectional lines serving a total of 459 stations (as defined under ‘parent stations’ in the transit feed).

The AFC system on both trams and buses requires travellers to check-in and out with the *OV-chipkaart*, (the national public transport smart card; for more details see [19]) every time they board and alight a vehicle; thus, potentially allowing full observation of chosen routes. Moreover, since, a large number of travellers in the network use smart cards for fare payment a significant amount of data is available for analysis. The data, made available by the operator, is pre-processed such that individual smart card transactions (check-ins and outs) are already chained to approximately 5.8 million journeys from origin to destination stations. Out of these, the case study, which only includes trips in weekday extended morning peak hours (0600h to 1100h), makes use of about 1.5 million journeys.

##### B. Analysis

The pre-determined journeys used in the case study have been inferred using a simplistic rule based on maximum transfer time (35 minutes [20]). Such inference methods typically lead to an overestimation of routes with more transfers and can leave seemingly illogical trips in the data. A full and robust (against disruptions) transfer inference algorithm as given in [20] can solve these issues. However, this is not done and misidentified journeys are directly filtered out when they do not match with the generated feasible routes. This seems to have a relatively low impact for the time period selected for the analysis as the overall coverage of the generated-feasible routes is nearly 85 percent of the observed routes.

Fig. 5 compares the logical, feasible, and identified route choice set size distributions. As one would expect, logical choice sets are typically large (median size: 58 routes). A sharp decline in the sizes for the feasible set (median size: 9 routes) is brought about mainly by the state-wise dominance elimination rule, although some routes are also removed due to service unavailability in certain time periods.

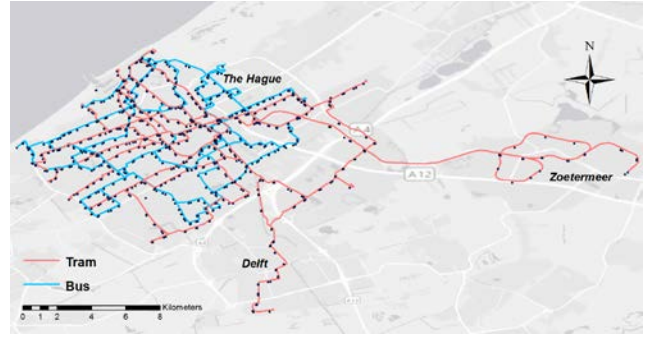


Fig. 4. The Hague tram and bus networks

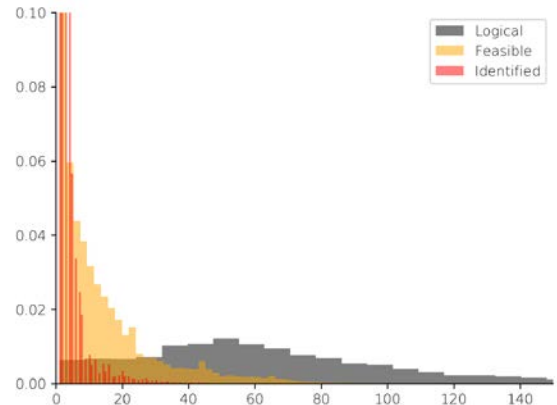


Fig. 5. Comparison of choice set size distributions (normalized) of logical, feasible, and identified route choice sets (axes top- and right-censored for better focus)

For the EBA calibration, three attributes, waiting time, in-vehicle time, and number of transfers, are considered. Based on experience and with an eye on computational efficiency, the threshold search space for the former two is defined between 1 and 2 with a step size of 0.025, while all possible values (0 to 2) are tried for the latter. Note that, if an intuition for these values is not available, one could simply try a larger search space.

##### C. Results

Performance of the six permutations (Fig. 6) indicate a clear preference in attribute ranking. Similar to findings for fully compensatory route choice models in literature, people rank number of transfers as the most important parameter followed by waiting and in-vehicle time, respectively.

For all permutations, constraints on individual attributes are quite restrictive: for waiting and in-vehicle time most multiplicative thresholds lie between 1 and 1.1 (meaning that only a 10 percent increase is acceptable), while for transfers, routes with even a single extra transfer are unacceptable in the choice set. These thresholds are lower than those assumed in CSGM studies assuming a conjunctive model for consideration set formation. For instance, for road traffic, the threshold used for travel time is 1.5 in [10]. Moreover, because of low thresholds, the choice sets sizes are also small (Fig. 5) with a median size of only 2 routes. Table 1 gives the threshold values obtained for individual attributes.



TABLE I. OPTIMAL ATTRIBUTE RANKING AND THRESHOLDS

Rank	Attribute	Threshold <sup>a</sup>	Sequential Coverage
1	Number of transfers	0	99.3%
2	Waiting time	1.1	82.0%
3	In-vehicle time	1.1	78.4%

<sup>a</sup> Threshold accuracy for waiting and in-vehicle time = 0.025

To assess the performance of the calibration the overall coverage of the EBA model can be calculated as the product of the coverage values obtained sequentially for each attribute (Table 1). The overall coverage for this case study is 63.9%.

D. Discussion

Fig. 7 takes a deeper look into the values of individual thresholds for the optimal permutation. It can be seen that for the first attribute – number of transfers – at 99.3 percent, coverage is already extremely high with no extra transfers; a clear reflection of travellers’ dislike for transferring. Thus, any increase in the transfer threshold only decreases efficiency thereby increasing the indicator value. For waiting time too, the initial coverage is quite high, meaning that improvements in coverage tend to be quite slow. Accepting twice the least possible waiting time only increases coverage from 79.9 to 93.3 percent. On the other hand, efficiency quickly decreases by approximately 40 percentage points. Although the initial slope for coverage is slightly higher than efficiency, the overall change in the latter is higher for in-vehicle time too. The fact that the initial value of coverage is more moderated for this attribute could be because the values are calculated after the feasible choice set has been filtered based on the thresholds of the previous two attributes. Finally, it should be noted that the for all three attributes, the optimal indicator values are unambiguous.

A possible explanation for the restrictive constraints may be a combination of the following statistical observations and hypothesis. The statistical observations are (Fig. 8) (i) OD pairs with a high demand tend to be nearby (in terms of in-vehicle time) and (ii) OD pairs that are farther away tend to have more feasible routes generated by the CSGM. The hypothesis is that (iii) travellers are either able to evaluate alternatives better or have a lower threshold acceptance for OD pairs that are nearby. From statistical observation (i), the hypothesis in (iii), and the definition given in Eq. 1, it can be seen why the coverage values are already quite high at low thresholds. This increase in coverage is mainly due to the highly-used routes between OD pairs that are close to one another. On the other hand, statistical observation (ii) and the definition in Eq. 2 explain why non-selected alternatives from farther away OD pairs might play a larger role in the value of efficiency. This potential disconnect might cause a decrease in efficiency that is not sufficiently balanced by the increase in coverage, leading to smaller thresholds. The larger slopes of efficiency in comparison to coverage (Fig. 7) seem to indicate that this is indeed the case here.

Finally, while the coverage (63.9%) on its own is a somewhat moderate performance, it may be expected because, in an effort to be more transferable, the model trades-off coverage for an increase in efficiency.

V. CONCLUSIONS

Route choice set identification for public transport networks is a vital but complicated task. Identifying the correct route choice sets are crucial for both, estimation and application, of route choice models. However, approaches

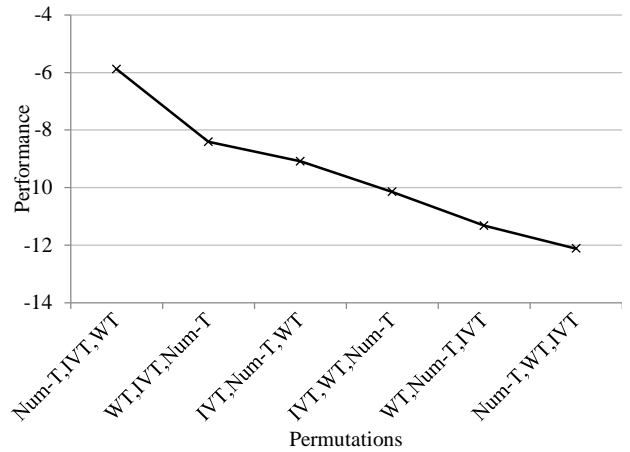


Fig. 6. Performance for different attribute rankings (Num-T: number of transfers, WT: waiting time, IVT: in-vehicle time) (lower is better)

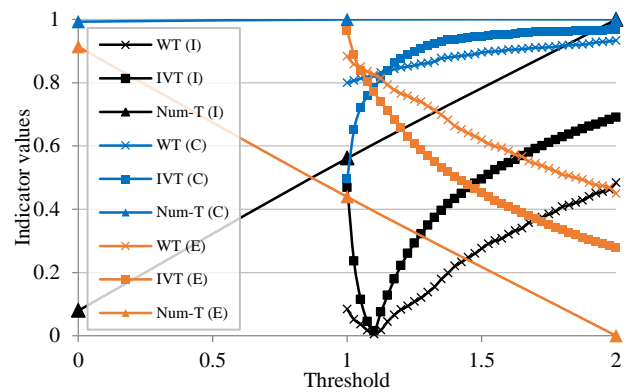


Fig. 7. Coverage (C), efficiency (E), and optimization indicator (I) values (y-axes) by threshold values (x-axis) of different attributes for the optimal permutation

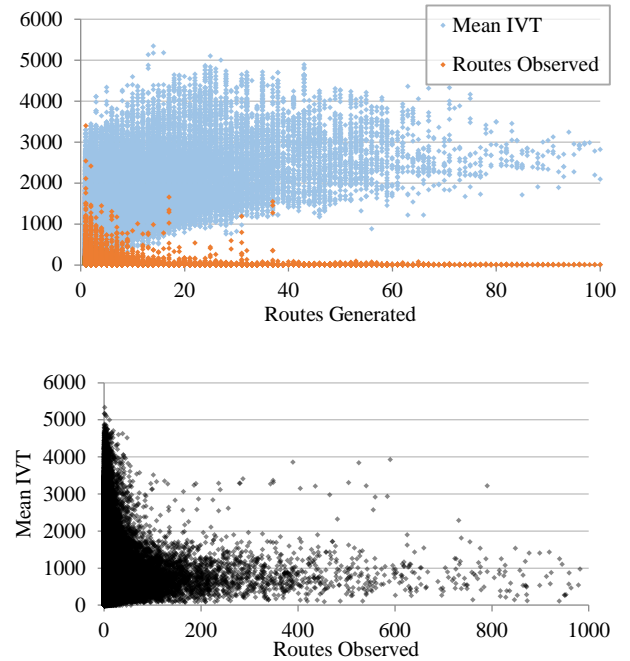


Fig. 8. Comparison between the number of generated-feasible routes, number of observed routes and average in-vehicle times (in seconds) per origin, destination, and time period (x-axis right-censored for better focus)



developed and commonly employed in literature either lack transferability (observation-driven methods) or require strong assumptions regarding traveller behaviour (uncalibrated CSGMs).

Given this key scientific gap, and the context of increasing availability of smart card data for public transport networks, this research makes two crucial contributions. First, a choice set generation methodology is proposed which uses elimination-by-aspects as the consideration set formation model. This model adds more behavioural dimensions over those used previously by taking into account attribute ranking as well as threshold constraints. Second, rather than making subjective assumptions regarding traveller preferences, the elimination-by-aspects model is calibrated using revealed behaviour observations from smart card data. The proposed methodology can be used to identify choice sets for estimating route choice models from revealed preferences as well as to predict alternative shares on the basis of available choice parameters.

Application of the proposed methodology to the urban public transport network of The Hague revealed that the number of transfers is the most important attribute for travellers, followed by waiting time, and in-vehicle time. Furthermore, the thresholds obtained for individual attributes are quite restrictive indicating that travellers make more optimal choices than previously assumed. Although the overall coverage for the EBA model is on the lower side, it makes up for this by being a more transferable model rather than a network-specific one.

An important limitation in the current implementation of the model is the assumption that the public transport services are frequency-based. Based on this, waiting times are calculated from the headways of individual lines under the assumption of evenly spaced arrivals of public transport lines and uniformly distributed traveller arrivals at stops. These assumptions may not hold outside rush hours or for non-urban networks where line frequencies are often lower, or when lines are explicitly synchronised to reduce transfer waiting time.

To overcome issues arising from the assumption of a frequency-based system, future implementations may consider using the following: (i) a schedule-based network which includes time in its representation and (ii) more complex traveller arrival models. Further improvements to the model could include taking into account that travellers behave differently for OD pairs that are relatively near, as hypothesised in the discussion of the case study results. Finally, future research could focus on using the calibration procedure proposed here, for the comparison of different behavioural models of route choice set formation.

#### ACKNOWLEDGMENT

The authors thank HTM, the urban public transport operator of The Hague, Netherlands, for their valuable cooperation and data provision.

#### REFERENCES

- [1] J. R. Hauser, "Consideration-set heuristics," *Journal of Business Research*, vol. 67, no. 8, pp. 1688-1699, 2014/08/01/ 2014.
- [2] C. Prato and S. Bekhor, "Modeling Route Choice Behavior: How Relevant Is the Composition of Choice Set?," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2003, pp. 64-73, 2007.
- [3] S. Hoogendoorn-Lanser, "Modelling travel behaviour in multi-modal networks," PhD Thesis, Delft University of Technology, Delft, The Netherlands, 2005.
- [4] S. Raveau, "Modelling Travel Choices on Public Transport Systems with Smart Card Data," in *Public Transport Planning with Smart Card Data*, F. Kurauchi and J.-D. Schmocker, Eds. Boca Raton: CRC Press, 2017.
- [5] D. Ton, D. Duives, O. Cats, and S. Hoogendoorn, "Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam," *Travel Behaviour and Society*, vol. 13, pp. 105-117, 2018/10/01/ 2018.
- [6] P. H. L. Bovy, "On Modelling Route Choice Sets in Transportation Networks: A Synthesis," *Transport Reviews*, vol. 29, no. 1, pp. 43-68, 2009/01/01 2009.
- [7] C. G. Prato, "Route choice modeling: past, present and future research directions," *Journal of Choice Modelling*, vol. 2, no. 1, pp. 65-100, 2009/01/01/ 2009.
- [8] M. Ben-Akiva, M. Bergman, A. J. Daly, and R. Ramaswamy, "Modeling inter-urban route choice behaviour," in *Proceedings of the 9th international symposium on transportation and traffic theory*, 1984, pp. 299-330: VNU Science Press Utrecht, The Netherlands.
- [9] Z. Guo and N. H. M. Wilson, "Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 2, pp. 91-104, 2011/02/01/ 2011.
- [10] C. G. Prato and S. Bekhor, "Applying Branch-and-Bound Technique to Route Choice Set Generation," *Transportation Research Record*, vol. 1985, no. 1, pp. 19-28, 2006.
- [11] S. Bekhor and C. G. Prato, "Methodological transferability in route choice modeling," *Transportation Research Part B: Methodological*, vol. 43, no. 4, pp. 422-437, 2009/05/01/ 2009.
- [12] S. Hoogendoorn-Lanser, P. Bovy, and R. van Nes, "Application of Constrained Enumeration Approach to Multimodal Choice Set Generation," *Transportation Research Record*, vol. 2014, no. 1, pp. 50-57, 2007.
- [13] P. H. L. Bovy and S. Fiorenzo-Catalano, "Stochastic route choice set generation: behavioral and probabilistic foundations," *Transportmetrica*, vol. 3, no. 3, pp. 173-189, 2007/01/01 2007.
- [14] D. I. Hawkins and D. L. Mothersbaugh, *Consumer behavior: Building marketing strategy*. Boston: McGraw-Hill Irwin, 2010.
- [15] A. Tversky, "Elimination by aspects : A theory of choice," *Psychological Review*, vol. 79, no. 4, pp. 281-299, 1972.
- [16] P. H. L. Bovy and E. Stern, *Route Choice: Wayfinding in Transport Networks: Wayfinding in Transport Networks (Studies in Operational Regional Science)*. Springer Science & Business Media, 1994.
- [17] S. Shelat and O. Cats, "Measuring spill-over effects of disruptions in public transport networks," in *Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017 5th IEEE International Conference on*, 2017, pp. 756-761: IEEE.
- [18] C. Puto, J. Huber, and J. W. Payne, "Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis," *Journal of Consumer Research*, vol. 9, no. 1, pp. 90-98, 1982.
- [19] N. Van Oort, T. Brands, and E. de Romph, "Short term ridership prediction in public transport by processing smart card data," *Transportation Research Record*, no. 2015, 2015.
- [20] M. D. Yap, O. Cats, N. van Oort, and S. P. Hoogendoorn, "A robust transfer inference algorithm for public transport journeys during disruptions," *Transportation Research Procedia*, vol. 27, pp. 1042-1049, 2017/01/01/ 2017.