TU Delft

Delft University of Technology

Validation of a multi-modal transit route choice model using smartcard data

Dixit, Malvika; Cats, Oded; van Oort, Niels; Brands, Ties; Hoogendoorn, Serge

DOI
10.1007/s11116-023-10387-z

Publication date
2023

Document Version
Final published version

Published in
Transportation

Citation (APA)
Dixit, M., Cats, O., van Oort, N., Brands, T., & Hoogendoorn, S. (2023). Validation of a multi-modal transit route choice model using smartcard data. *Transportation*, *51*(5), 1809-1829. https://doi.org/10.1007/s11116-023-10387-z

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.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Transportation
https://doi.org/10.1007/s11116-023-10387-z

Validation of a multi-modal transit route choice model using smartcard data

Malvika Dixit[1] · **Oded Cats**[1] · **Niels van Oort**[1] · **Ties Brands**[1] · **Serge Hoogendoorn**[1]

Accepted: 23 March 2023
© The Author(s) 2023

Abstract
Validation of travel demand models, although recognised as important, is seldom undertaken. This study adds to the scarce literature in this field by undertaking an external validation of a multi-modal transit route choice model. The model was estimated using smart card data for the urban transit network of Amsterdam before the introduction of a new metro line and is used to predict changes in travel behaviour after the network change. To validate, the model was checked for changes in estimated parameters between the two time periods, and predictive ability was evaluated at different aggregation levels. Although most model parameters were found to be unstable between the two contexts, the predictive performance at all levels was similar to the locally estimated model. Moreover, individual choices and transit mode-share predictions were found to be close to the observed ones. The errors were relatively larger for the link and route-level predictions, some of which could be attributed to the assumptions made regarding consideration choice set given as input to the model. On comparing alternative model specifications, using generic instead of mode-specific travel attributes lead to a strong degradation in predictive performance. Conversely, a model incorporating overlap between routes, with a better model fit in the base period, did not offer a clear improvement in prediction performance. The study highlights the need to validate transit route choice models before using them for deriving policy recommendations, especially in this data-rich age in which it can often be undertaken at a relatively low additional cost.

Keywords Route choice · Automated data · Public transport · Model transferability · Ex-post model evaluation

Introduction

The last few decades have seen substantial research into discrete choice models of transit route choice (Bovy and Hoogendoorn-Lanser 2005; Guo and Wilson 2011; Liu et al. 2010). These models aid in understanding transit riders' preferences by revealing the

✉ Malvika Dixit
M.Dixit-1@tudelft.nl

[1] Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628CN Delft, The Netherlands

relative valuation of various travel attributes, often specifically focusing on service quality characteristics such as those related to transfers (Garcia-Martinez et al. 2018; Guo and Wilson 2011; Nielsen et al. 2021), crowding (Hörcher et al. 2017; Kim et al. 2015; Yap et al. 2020) or reliability (Swierstra et al. 2017). The relative valuations obtained from these models can be used for predicting passenger flows in response to changes in policy, enabling the comparison of alternative policy scenarios. When the selected model is close to the true representation of reality, the estimated parameters are expected to be stable for a reasonable range of temporal and spatial conditions, and the model forecasts are expected to resemble the observed demand. However, the process of model validation is only seldom undertaken, and model selection is typically made based on goodness-of-fit statistics such as log-likelihood and rho-squared (Parady et al. 2021). Although useful in their own right, models with high goodness-of-fit may not necessarily be well-specified and hence may not be transferable (Koppelman and Wilmot 1982). Issues like overfitting, endogeniety, omission of variables, measurement errors, incorrect model structure or incorrect theoretical assumptions about the travel behaviour could lead to a misspecified model, which may still have an acceptable goodness-of-fit statistic.

Model validation can be defined as "the evaluation of generalizability of a statistical model" (Parady et al. 2021) and includes both internal validation or reproducibility and external validation or transferability. External validation can be further divided into spatial transferability, temporal transferability, and methodological transferability (Parady et al. 2021), with the latter referring to the model performance on data collected using different methodologies. Although recognised as important, external validation is rarely undertaken in the case of travel demand models, probably due to the lack of suitable data. Parady et al. (2021) highlight that only 4% of transport academic literature published between 2014 and 2018 conducted an external validation.

In recent years, revealed preference data in general, and smart card data, in particular, has become increasingly available for inferring route choices of transit travellers (see for exampleHörcher et al. 2017; Jánošíkova et al. 2014; Kim et al. 2019; Yap et al. 2020). Depending on the penetration rate amongst transit riders, smart card data can provide information on almost all journeys made in the network at a highly disaggregated level. However, no information is available on the intention of the travellers, their origin location, and in many cases the time of arrival at the origin stop. Due to these limitations, several assumptions need to be made along the modelling process, specifically regarding the travellers' consideration choice set and the perceived level of service values. However, to the best of our knowledge, none of the studies that elicits route choice preferences from smart card data has attempted to validate their performance. This study aims to address this gap in the literature by undertaking an external validation of a transit route choice model using smart card data and thereby provide valuable insights on how transferable such models are and how we can facilitate their transferability.

A model of transit mode-route choice was developed for the urban transit network of Amsterdam, where a new North–South metro line was added to the existing bus, tram, and metro network in July 2018. Along with the addition of the new line, significant changes were made to the rest of the network (see Brands et al. (2020) for details). This major network change provides an opportunity to perform an ex-post evaluation of the route choice model developed based on data before the network change. Two types of validation tests are undertaken. First, we compare the model parameters estimated for the transferred ('before') model with the locally estimated model developed based on the data 'after' the network change. The two data sets used are ~3 months apart. Second, the demand changes estimated using the transferred model are compared with the observed demand after the

network change. The results aim to establish the validity of route choice models estimated using smart card data for predicting the change in travel behaviour because of a major network change.

The rest of the paper is structured as follows: we start with reviewing the literature on "External validation of travel demand models". Section "Methods" describes the study setting and the various statistical tests used for model validation, along with the model specifications. Section "Results and Discussion" presents the results of the validation tests undertaken on our data, and Section "Conclusion" discusses the main conclusions.

External validation of travel demand models

External validation of models, or transferability, implies the ability of a model developed in one context to be useful in another context. Transferability is implicitly assumed when models are used to predict change in demand in response to a policy change. Some of the earliest literature on (external) model validation dates back to Atherton and Ben-Akiva (1976) and Train (1978). Since then, most work in this area has focused on the temporal transferability of models over long time horizons (often more than 10 years) and/or their spatial transferability. The primary motivation for such studies was to reduce costs of data collection and model development by using an existing model for a comparable region or during a different time period for the same region. Parady et al. (2021) provide a comprehensive review of the recent literature on the validation of discrete choice models in transportation. Here we narrow our focus to external validation studies, and discuss the main issues and corresponding learnings from these studies.

A fundamental theoretical assumption behind any model transferability is the consistency of underlying behavioural theory in both contexts. Koppelman and Wilmot (1982) highlight that model transferability is a "property of the estimation and application contexts, as well as the specification of the model". Naturally, a model with highly context-specific variables will not be transferable to a new context. Sometimes the Alternative Specific Constants (ASCs) are updated based on the application context to account for average changes in unobserved variables between the two contexts (Atherton and Ben-Akiva 1976; Badoe and Miller 1995; Sanko and Morikawa 2010). While the updated ASCs capture the mean contribution of the unobserved terms, there could also be differences in the variance of these unobserved terms. Hence, before transferring a model, the scale for the transferred model needs to be updated to match the scaling differences between the two contexts (Swait and Louviere 1993). In cases where the estimation and application contexts are widely different, one could implement a partial model transfer with varying transfer scales for different sub-groups of variables (Gunn et al. 1985). This is especially applicable when some parameters are more transferable than others. For example, Fox et al. (2014) found the level of service parameters to be more transferable than cost parameters in their study of mode-destination choice models.

Multiple studies have noted that, generally, an improved model specification improves transferability (Badoe and Miller 1995; Fox et al. 2014; Rossi and Bhat 2014). However, some others also highlight the risk of overfitting which may reduce the transferability of models. For example, Fox (2015) found that although incorporating taste heterogeneity in time and cost parameters improved model fit for their base data, it did not necessarily result in enhanced transferability. Badoe and Miller (1995) also report a similar finding where

over-specification led to reduced transferability. Overall, it is noted that a good fit in the estimation context may not be sufficient.

Another issue of concern is the ability of a model to capture causal relationships. As clearly highlighted by Atherton and Ben-Akiva (1976): "To be transferable, then, it is not enough that the model merely fit existing data; it must also explain why travel behaviour changes as conditions change. Rather than simply correlating existing travel behaviour with socioeconomic characteristics and transportation level of service, the model specification must represent the causal relationships between these variables. Thus, the causal specification of a model is a precondition to its consideration for transferability." For example, Chorus and Kroesen (2014) argue against the transferability of hybrid choice models for predicting policy outcomes, as these models (theoretically) cannot capture the causal relationship between the latent variable and the travel choice.

The only way to empirically establish whether a model is under/over specified or if it captures the causal relations required for transferability is to undertake a posterior analysis of transferability. Nonetheless, as Koppelman and Wilmot (1982) note, such posterior analyses of transferability are undertaken with the intent to provide insights that can be helpful for (future) prior transferability studies. This study aims to get such insights for the case of transit route choice models, specifically the ones estimated based on smart card data.

So far, most validation studies in the literature have been for mode or mode-destination choice models. In the case of route choice models, some studies undertake an internal validation (see for examples Lai and Bierlaire (2015), Mai (2016)), but very few an external validation. Bekhor and Prato (2009) were the first to consider the issue of transferability of route choice models. They undertook a spatial transferability assessment of traffic route choice models based on two independent revealed preference survey data sets, one each for Boston and Turin networks. In addition to assessing the transferability of the route choice models, they also evaluated the transferability of path generation techniques. In their case, the transferability of route choice model parameters could not be verified, partly due to the dissimilarity in characteristics between the two networks.

To the best of our knowledge, none of the studies so far have undertaken an external validation of a transit route choice model. This study addresses this gap by undertaking a transferability analysis across two closely spaced time periods for the same urban area, which allows for many exogenous factors to be controlled for, including any major changes in the underlying population. Specifically, the following issues are investigated using the smart card data from before and after a major network change:

(i) How transferable are models of transit route choice estimated using smart card data, and can they be used for forecasting the changes in demand because of network changes?
(ii) How does omitting/adding relevant variables (determined based on improved goodness of fit measures in the base context) impact models' prediction performance?

Method

Case study context and data preparation

In July 2018, a new metro line (the north–south line) was introduced in the urban transit network of Amsterdam, the Netherlands, adding significant capacity to the existing

network of metro, bus and tram lines. The new metro line runs through the dense historical city centre, and connects the northern part of the city with the centre—a connection which was made earlier via buses with highly circuitous routes. The opening of the new metro line was accompanied by a re-structure of the existing bus and tram network, including the addition of new feeder routes and re-routing or removal of duplicate routes. The new metro line differs from the existing ones in a few aspects—some of the stations (especially the ones in the city centre) are deeper than the existing metro stations implying a longer access time to the metro. In addition, the frequency for the new line is higher than the frequencies offered on the other metro lines (see Brands et al. (2020) for details).

This significant change in public transport supply provides an opportunity to undertake a transferability analysis for the transit route choice models developed for the network using the two time periods corresponding to before and after the opening of the new metro, as shown in Fig. 1. We use 5 weeks of data in the time period before and 6 weeks in the time period after the opening of the new metro line. Although the two time periods used in this study are very close apart, the major changes to the transit network supply cause significant changes to the flow patterns (as shown in Brands et al. (2020)), making this case study ideal for undertaking a model transferability analysis.

We use a combination of smart card and Automated Vehicle Location (AVL) data for the route choice model estimation and validation analysis (see van Oort et al. (2015) for an overview of the Dutch smart card system). The smart card data used includes all the journeys made in the network, including those by tourists that could use an unlimited travel ticket for one or more days valid for all modes of public transport. These tickets also need to be validated for each public transport trip, and are hence recorded in the data. There are no mode-specific season passes in the network, and the fare is based on the (network) distance travelled irrespective of the mode used. It is also important to note that the same individual may be recorded multiple times, but owing to privacy concerns, we cannot track them across days and hence consider them as independent observations.

The raw smart card data is processed by undertaking cleaning, destination inference and transfer inference to form a journey database (see Dixit et al. (2019) for more details on these steps). For undertaking route choice analysis, we use only the morning peak period for our model estimation, as it is expected to have a higher share of commuters during this time, which are typically more regular travellers making their travel choices more conscious (Fox and Hess 2010). The choice set is derived based on the observed routes used by all the travellers in the data set. Transit stops in close proximity are clustered together to

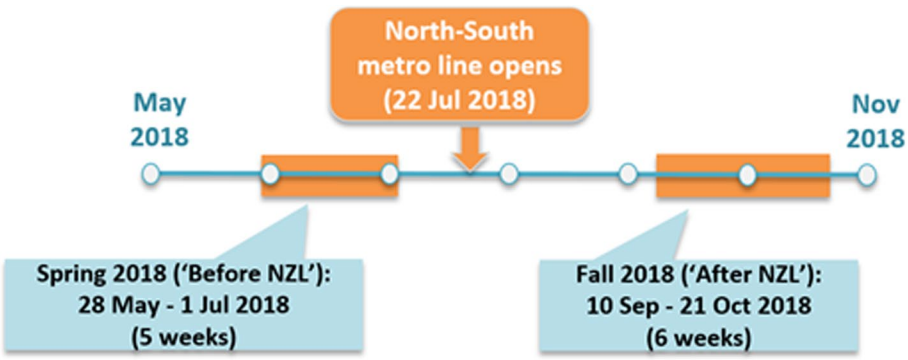

Fig. 1 The time period for validation analysis

form a more realistic consideration choice set, and a threshold of minimum 20 journeys for each route in the before period (and 24 in the after period) is applied to ensure only reasonable routes are included (see Dixit et al. (2021) for more details on this). After applying all filters, a dataset of 382,295 observations for the before period and 563,210 for the after period is obtained which is used for estimation and validation of the model, respectively. This corresponds to 582 OD pairs in the before period and 593 in the after period.

Model specification

We specify and test three models of transit route choice. We start with an MNL model with mode-specific travel attributes, with the deterministic component of utilities as specified in Eq. 1.

$$\begin{aligned} V^{MNL\_specific} &= \beta_{ivt_{bus}} * IVT_{bus} + \beta_{ivt_{tram}} * IVT_{tram} + \beta_{wait_{bt}} * WT_{bt} + \beta_{tt_{metro}} * TT_{metro} \\ &+ \beta_{trans_{bt}} * Trans_{bt} + \beta_{trans_{bt}} * Trans_{btm} + \beta_{trans_{m}} * Trans_{m} \\ &+ \beta_{TrT} * TrT + \beta_{Circ} * Circ + MSC_{Bus} + MSC_{Tram} + MSC_{Metro} \end{aligned} \tag{1}$$

where IVT_{bus}andIVT_{tram} are the in-vehicle times by bus and tram in minutes, respectively, WT_{bt} is the expected initial waiting time for bus and tram modes, TT_{metro} is the travel time by metro including the initial waiting time at the platform, $Trans_{bt}$, $Trans_{btm}$ and $Trans_{m}$ are the numbers of transfers made within the bus/tram network (which includes bus-bus, tram-tram and bus-tram transfers); transfers between metro and bus/tram; and transfers within the metro network, respectively, TrT is the transfer time in minutes, $Circ$ is the circuity of the route measured as the ratio of network to Euclidean distance, and MSC_{bus}, MSC_{tram} and MSC_{metro} are the mode-specific constants for bus, tram and metro, respectively.

It is important to note that the smart card data in Amsterdam contains different information for bus and tram versus metro. For buses and trams, the smart card is tapped inside the vehicle, whereas for metro, this happens at the station. Hence, the time measured by smart card data for metro includes the waiting time at the origin. For buses and trams the time of arrival of the passenger at the bus/tram stop is not known. Hence, the effective waiting time is derived based on the observed headway at each origin station using the corresponding vehicle arrival information available from AVL data. Hence, we use $IVT_{bus/tram}$ for denoting the in-vehicle time (without waiting time) for buses and trams and TT_{metro} for the travel time by metro, which includes waiting time.

Amsterdam public transport network follows a (network) distance-based fare system, with the same fare per distance travelled irrespective of the mode(s) used. This makes the travel cost directly proportional to circuity for the alternative routes between an origin–destination pair. Hence, travel cost has not been considered as an additional, separate, variable in our models.

Next, to analyse the impact of omitting variables, instead of mode-specific in-vehicle time and transfer penalties, we use generic ones. The deterministic utility function in this case is shown in Eq. 2.

$$\begin{aligned} V^{MNL\_generic} &= \beta_{ivt} * IVT + \beta_{wait_{bt}} * WT_{bt} + \beta_{trans} * Trans \\ &+ \beta_{TrT} * TrT + \beta_{Circ} * Circ + MSC_{Bus} + MSC_{Tram} + MSC_{Metro} \end{aligned} \tag{2}$$

where IVT corresponds to the in-vehicle time in minutes, and $Trans$ is the number of transfers made within or across modes.

Lastly, we test the model which incorporates the overlap between alternative routes. For this, we use a Path Size Correction Logit model which includes overlap of path and transfer nodes. The path size correction terms for journey legs (PSC_i^T) and transfer nodes (PSC_i^X) are as defined in Dixit et al. (2021), given by.

$$PSC_i^T = -\sum_{l\in\Gamma_i}\left(\frac{t_l}{T_i}\ln\sum_{j\in C}\delta_{lj}\right) \text{ and } PSC_i^X = -\sum_{n\in K_i}\left(\frac{1}{X_i}\ln\sum_{j\in C}\delta_{nj}\right) \quad (3)$$

where t_l = travel time for journey leg l in route i, T_i = total travel time for route i, Γ_i = set of all legs for route i, C = set of all routes between the chosen origin–destination pair, δ_{lj} = leg-route incidence between leg l belonging to alternative route j, X_i = Number of transfer nodes in route i, K_i = set of all nodes for route i, and δ_{nj} = node-route incidence between node n belonging to alternative route j.

The PSC terms decrease as the amount of overlap between alternative routes increases, with a maximum value of 0 for perfectly independent routes and a lower theoretical bound $-\infty$ for perfectly correlated alternatives. These path size correction terms are added to the deterministic utility function with mode-specific travel attributes as defined in Eq. 1.

Validation assessment

We divide the external validation metrics into two categories. The first category relates to the stability of estimated parameters, while the second one assesses the predictive ability of the model. The second category is further split into measures of disaggregate and aggregate predictions. Figure 2 shows this classification, and the sets of metrics used for each category. Subsequent sections elaborate on each of the metrics used.

While this study focuses on external validation of route choice models, it should be noted that the models used in this study have been validated internally using a cross-validation approach. Readers are referred to Dixit et al. (2021) for more details and results of the internal validation exercise.

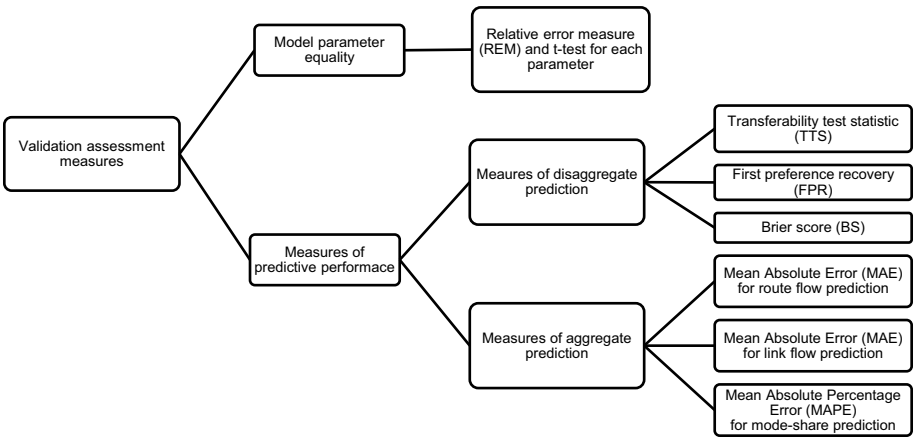

Fig. 2 Validation assessment metrics used

Model parameter equality

The first test of transferability consists of comparing the parameters estimated for the base and transfer contexts (the before and after situations in our case). This helps establish whether some parameters are more transferable than others. Since the two datasets are from different time points, we first check for differences in scale parameters between the two. For this, the two datasets are pooled together and the scale parameter is estimated relative to the 'before' dataset.

After adjusting for scale differences, the parameters estimated for the two cases are compared. For each estimated parameter, the relative error measure (REM) is calculated as:

$$REM_{\beta_k} = \frac{\mu\beta_k^{after} - \beta_k^{before}}{\beta_k^{before}} \tag{4}$$

where μ is the scale parameter to account for differences in error variance between the two cases, β_k^{after} is the parameter for attribute 'k' estimated for the after case, and β_k^{before} is the parameter for attribute 'k' estimated for the before case.

Next, we check for statistical significance of the differences in each of the model parameters by means of a t-test, as described in Fox (2015). The t-statistic, in this case, is given by,

$$t\left(\mu\beta_k^{after} - \beta_k^{before}\right) = \frac{\mu\beta_k^{after} - \beta_k^{before}}{\sigma\left(\mu\beta_k^{after} - \beta_k^{before}\right)} \tag{5}$$

The denominator σ corresponds to the standard error of the difference in parameters. In our study, although the two datasets were collected a few months apart, we do not link individual observations collected in different periods. When the covariance is assumed to be zero, and scale μ fixed, the standard error of difference is given by,

$$\sigma\left(\mu\beta_k^{after} - \beta_k^{before}\right) = \sqrt{\left(\mu\sigma\left[\beta_k^{after}\right]\right)^2 + \left(\sigma\left[\beta_k^{before}\right]\right)^2} \tag{6}$$

where $(\sigma\left|\beta_k^{after}\right|)$ is the standard error of β_k^{after}, and $(\sigma\left|\beta_k^{before}\right|)$ is the standard error of β_k^{before}.

Since model parameters cannot be interpreted directly, we also compare the model elasticities between the two time periods. Elasticities capture the sensitivity of a model to changes in key input variables such as travel times. The disaggregate point elasticity of alternative route 'r' for an individual 'i' with respect to a variable K is calculated as

$$E_K^{P_{ir}} = \frac{K}{P_{ir}} . \frac{\partial P_{ir}}{\partial k} \tag{7}$$

The disaggregate elasticity is aggregated by calculating a (probability) weighted average across all individuals and alternatives, to arrive at an average elasticity value which is compared between the locally estimated and transferred models. Being dimensionless, model elasticities can be compared across different models and time periods (Fox 2015).

Disaggregate measures of predictive ability

Next, we assess how well the transferred model can predict the outcome of the network change. For this, the model parameters estimated based on the 'before' data are used to estimate the probabilities for the 'after' situation. The outcomes obtained using the transferred model are then compared to those from the locally estimated model (i.e. model estimated with same the specification but using the 'after' data). In this section, we discuss the methods for comparing the performance for individual-level predictions. As there is no agreement in the literature on the best metric for this, we use multiple metrics, each providing a different perspective on it, as described below:

- **Transferability Test Statistics (TTS):** The TTS statistic is similar to a likelihood ratio test undertaken between transferred and locally estimated model, both applied to the 'after' data. It is a strict pass/fail test and is chi-squared distributed with degrees of freedom equal to the number of model parameters. This has been used by Atherton and Ben-Akiva (1976) and Koppelman and Wilmot (1982) among others to test model transferability.

$$TTS_{after}(\beta_{before}) = -2 * (LL_{after}(\beta_{before}) - LL_{after}(\beta_{after})) \tag{8}$$

where $LL_{after}(\beta_{before})$ is the log of the likelihood that the observed 'after' data were generated by the transferred model, and $LL_{before}(\beta_{before})$ is the log-likelihood of the locally estimated model on the 'after' data.

Although commonly noted, it has been observed that almost all models fail this strict test of transferability (Badoe and Miller 1995; Fox 2015).

- **First preference recovery (FPR)**: Also referred to as 'percentage of correct predictions', this shows the percentage of choices correctly estimated by the model, given by Eq. 8:

$$FPR = \frac{100}{N}\sum_{i=1}^{N}(y_i^p = y_i^o) \tag{9}$$

where y_i^p is the predicted choice (route) for an individual 'i', and y_i^o is the observed choice (route) for an individual 'i', and$N$ is the number of individuals (observations) in the data.

As opposed to TTS, the FPR provides an indication on the degree of transferability and can be used to compare alternative models in terms of how well they can predict individual choices. It can also be useful when comparing the results with similar studies in the literature. However, a major limitation of this measure is its inability to differentiate between the range of probabilities assigned to the chosen alternatives (de Luca and Cantarella 2016). Hence, we look at another measure—Brier score—of disaggregate predictive performance that considers the probabilities assigned to the chosen and non-chosen alternatives.

- **Brier score (BS):** Brier score (Brier 1950) is an absolute measure used to quantify the accuracy of probabilistic predictions. For each alternative route in each observation, the predicted probability of choosing it is subtracted by the actual outcome. The square of this value is summed across all alternatives for each observation, and averaged across all observations. Mathematically, it is given by:

$$BS = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R_i}(P_{ir} - y_{ir})^2 \tag{10}$$

where P_{ir} is the predicted probability an individual 'i' chooses alternative route 'r'.y_{ir} is equal to 1 is alternative route 'r' is chosen by individual 'i' and 0 otherwise,R_i is the number of alternative routes available to individual 'i', and N is the number of individuals (observations) in the data.

The Brier Score has a minimum value of 0 for perfect predictions and a maximum value of 2 for the worst possible prediction.

Aggregate measures of predictive ability

From a policy perspective, one is often more interested in aggregate level shares as opposed to individual level predictions. Depending on the application under consideration and the requirements of the decision maker, different levels of forecasting may be relevant. For example, for understanding capacity and infrastructure issues as well as for operational planning, link-level forecasts are highly relevant to identify bottlenecks and related capacity needs. On the other hand, for service planning aspects such as fare scheme policies, network design, and frequency setting, mode or route level forecasts may be most relevant for assessing overall trends. Hence, in this study, we compare the shares estimated by the models at each of these levels of aggregation. To do this, individual probabilities are summed to calculate the market shares for each alternative route for each OD pair. The aggregate predictions are assessed using three metrics addressing the different levels of aggregation, as discussed below:

- **Predictions per route—Mean Absolute Error (MAE):** The predicted shares (passenger flows) are compared to the observed ones, and the Mean Absolute Error (MAE) for each origin-destination (OD) pair is calculated as:

$$MAE_{od} = \frac{1}{R}\sum_{\mathrm{r}=1}^{R}\left|S_r^p - S_r^o\right| \tag{11}$$

where S_r^p is the predicted flows for alternative route 'r' for origin–destination pair 'o-d',S_r^o is the observed flows for alternative route 'r' for origin–destination pair 'o-d', and R is the number of available routes for origin–destination pair 'o-d'.

The $\mathrm{MAE_{od}}$ is then averaged across all ODs to get an average MAE per route.

- **Predictions per link—Mean Absolute Error (MAE):** The passenger flows per route are aggregated to calculate flows for each link. A link here refers to the path connecting two consecutive transit stops, which may be used by multiple transit routes. Similar to the route level, MAE is calculated for each link, and a mean MAE over all links is reported. The percentage error in predicting the flow on each link is also visualized to identify patterns.
- **Predicted modal shares—Mean Absolute Percentage Error (MAPE):** The predicted passenger flows on each route are further aggregated to calculate the market share for each mode combination. This is specifically relevant in our case as we would like to know how well the model performs when estimating the impact of network change on public transport mode-shares. The observed and predicted mode shares are compared, and the Mean Absolute Percentage Error (MAPE) is calculated for each mode as:

$$MAPE_{modes} = \frac{1}{M}\sum_{m=1}^{M}\left|P_m^p - P_m^o\right| \tag{12}$$

where P_m^p is the predicted mode-share for mode (combination) 'm', P_m^o is the observed mode-share for mode (combination) 'm', and$M$ is the number of mode (combinations).

Results and discussion

We first evaluate the validity of the MNL model with mode-specific travel attributes as described in Eq. 1. Then, the impact of variable omission is examined by testing the validity of the two alternate model specifications.

Model parameter equality

We start with examining the stability of the estimated model parameters across the before and after time periods. Before comparing the parameters, the two models were checked for differences in scale parameters. The scale difference was found to be significant with a value of 0.92 for the after model relative to the before model, implying a lower variance in the unobserved parameters for the after case.

Table 1 shows the parameters estimated from the two models after scaling, and the corresponding REM and t-test statistic for each. The relative error measure is the highest for the mode-specific constants, implying a significant difference in the average effect of unobserved (excluded) variables specific to each mode between the two contexts. This could include attributes like comfort, safety, cleanliness, reliability, weather protection at stations, availability of information, ease-of-navigation or any other inherent (dis)preference for a particular (public transport) mode. Some of these attributes are expected to change after the introduction of the new line. For example, deeper stations of the new metro line may reduce the attractiveness of the mode. On the other hand, the higher frequency and more options for travel may lead to it being more attractive. Amongst the rest of the parameters, circuity is found to have the highest change (an increase of 45% in magnitude), followed by the number of transfers within the metro which is found to decrease in magnitude by 19%. Although the REM values for all other parameters are approximately 10% or less, the null hypothesis of the parameters being identical across the two cases is rejected for most of them (with a 95% confidence interval). Only the travel time by metro, number of transfers between bus and tram, and the transfer time are found to be stable across the two time periods as per the t-statistic. It should be noted though that for both transferred and locally estimated models, the parameters are precisely estimated with relatively low standard errors (t-ratios of >20), which is often the case for models estimated with large scale data sources such as the smart card. Because of this, the null hypothesis of the parameters being identical across the two contexts is more likely to be rejected.

Next, we compare the model elasticities. Table 2 shows the elasticities for the transferred and locally estimated models. For both contexts, the (absolute) elasticity values are the highest for bus in-vehicle time, followed closely by the waiting time for bus and trams. The (absolute) elasticity values are found to be higher for the locally estimated model compared to the transferred model, implying the transferred model estimates the demand to be more inelastic that the locally estimated model.

Table 1 Model parameter comparison between models estimated on 'before' and 'after' datasets

Parameter*	Before	After**	REM	t-statistic	Significantly different?
Mode-specific constant for bus < fixed >	0.00	0.00	–	–	–
Mode-specific constant for tram	0.49	0.25	−48.5%	−13.90	Yes
Mode-specific constant for metro	0.84	0.37	−56.0%	15.04	Yes
Bus in-vehicle time (mins)	−0.11	−0.12	10.4%	4.97	Yes
Tram in-vehicle time (mins)	−0.09	−0.10	11.1%	6.54	Yes
Effective wait time bus/trams (mins)	−0.19	−0.20	5.6%	4.48	Yes
Metro time[a] (mins)	−0.09	−0.10	3.6%	1.22	No
Number of transfers between bus & tram[b]	−1.24	−1.21	−3.1%	0.80	No
Number of transfers between metro and bus/tram	−2.38	−2.19	−9.0%	6.83	Yes
Number of transfers within metro	−1.50	−1.22	−19.3%	6.30	Yes
Transfer time (mins)	−0.25	−0.25	0.2%	0.07	No
Circuity	−0.43	−0.63	45.1%	9.17	Yes

*$p < 0.01$ for all estimates

**the reported estimates are after adjusting for scale differences

[a]includes in-vehicle time and origin waiting times

[b]includes bus-bus, tram-tram and bus-tram transfers

Table 2 Direct model elasticities for locally estimated and transferred model

Elasticity	Locally estimated model	Transferred model	Transferred/ Locally estimated
Bus in-vehicle time (mins)	−0.49	−0.43	0.88
Tram in-vehicle time (mins)	−0.38	−0.32	0.85
Effective wait time bus/trams (mins)	−0.48	−0.43	0.90
Metro time (mins)	−0.31	−0.28	0.92
Transfer time (mins)	−0.35	−0.33	0.95
Circuity	−0.33	−0.22	0.66

There could be several reasons for the differences in estimated parameters and elasticities between the two contexts. Firstly, there could be contextual factors that are not captured by the observed variables that could differ between the two contexts. Secondly, the underlying population may have changed—some travellers may have stopped using transit after the network change while other new travellers may have been added. Also, some existing travellers may have reduced/increased their travel frequencies. A related point is the possible presence of endogeniety, especially since our study is based on observational data. The models assume the explanatory variables to be exogenous, which may not be true. This could be due to multiple reasons, including omitted variables. For example, the new metro line has newer, cleaner and more aesthetically pleasing trains and stations—contributing toward comfort, which is not included in our model(s). Concurrently, the new metro line provides direct routes with lower circuity values compared to the rest, making

the circuity correlated with the unobserved attribute of comfort. When using a model to predict the demand in response to changes in policy, it is important to have a model capturing the causal relationship between them. The smart card data used for this study does not provide the origin (home) location of the travellers. The missing attributes (such as access/egress distance or time, comfort levels, reliability, and accessibility of modes among others) and/or endogeniety may hamper the establishment of a causal relationship. Lastly, one cannot theoretically rule out that our model may have been misspecified (wrong model structure or non-linear relationships between variables), or that the behavioural theory is altogether inconsistent with the observed behaviour. Irrespective of the reasons behind the instability of model parameter values, our results imply that one should be cautious when making inferences on the relative valuation of travel time or service quality attributes from such models, specifically if they have not been thoroughly validated. We also note that since we do not track the individuals across days, we are not able to capture the impact of panel structure of the data in our model. This may have resulted in an underestimation of the standard errors of parameters for both transfer and locally estimated models. This further motivates us to analyse the predictive performance of our model to establish its usefulness for applications.

Predictive performance

Disaggregate measures

Next, we test the predictive ability of the model by forecasting the impact of the network change at an individual, route, link and mode level. The predictions are compared with the predictions from a locally estimated model (i.e. model estimated with same the specification but using the 'after' data) to benchmark the performance. We start with the measures of predictive performance at a disaggregate level, which are shown in Table 3. The TTS, compared against the chi-squared distribution for 11 degrees of freedom, strongly rejects the hypothesis that the two sets of parameters are equal. However, as many other studies note, most models fail this test of model transferability, but may still be good in their predictive abilities (see for example Badoe and Miller (1995); Forsey et al. (2014); Fox (2015)). In most cases, following the results from a strict pass/fail test such as the TTS blindly may not be a wise decision. While the TTS is useful to inform the modeler of the two models being different, as Parady et al. (2021) highlight, it is important to assess the extent to which they differ. To understand the extent of these differences in our case, we evaluate other disaggregate and aggregate level measures as discussed below.

Table 3 Disaggregate measures of predictive ability for locally estimated and transferred models

Statistic	Locally estimated model	Transferred model	Difference
Log-likelihood	−328,646	−330,299	0.5%
TTS	–	3,306	
First preference recovery	71.7%	71.5%	−0.3%
Brier score	0.370	0.372	0.7%

The FPR of over 70% is found to be rather high compared to values reported for other route choice models in the literature, where this percentage ranges between 51 and 73% for some of the recent studies (Parady et al. 2021). Moreover, both FPR and the Brier score of the transferred models are found to be very close to the locally estimated models (< 1% difference), with the FPR being marginally lower and the Brier score slightly higher in the case of transferred models. Hence, although many of the parameter estimates differ for the two periods, the predicted choice probabilities of the transferred model at an individual level are found to be close to the locally estimated model.

Aggregate measures

Disaggregate measures like FPR and Brier score are often used for assessing the models in terms of their ability to predict individual-level choices in the new context. However, in most applications, one is more interested in the predictions at the mode, route, or link levels. Hence, we analyse the performance of the transferred model to predict the market shares at each of these levels.

First, we use the MAE to compare the local and transferred models in terms of their predictions at the route level (Table 4). The MAE for the transferred model shows an average error of 45 journeys per route for the transferred model as compared to 43 for the locally estimated model. Examining the predicted flows at a link level, we observe an MAE of 328 passengers per link over the entire morning peak period when predicted using the transferred model, ~8.6% higher than that those obtained by the locally estimated model.

Figures 3 and 4 show the error in flow prediction at the link level by the local and transferred models, respectively. The width of the lines corresponds to the observed flow on the link. A positive error implies that the model overestimated the flow on the link, while a negative error means an underestimate of observed flow. The maps show that the link-level flow predictions using the two models are similar overall, implying that using the 'before' data set for estimation of the model is not a problem per-se, compared to the inherent estimation errors when using such a model. The maps can give an indication of the possible causes of such errors. For example, close to the central station, there are two parallel tram routes with one showing an underestimation while the other an overestimation of flows. These parallel tram lines are highlighted with a red circle in the maps and enlarged in the top right corner of both the figures. The errors in estimation could be attributed to the assumptions made regarding the consideration choice set for the model. The smart card data does not provide information on the origin (home) location of the travellers. Hence, stops within a maximum distance of 500 m were clustered together to form the consideration choice set for travellers (for more details on the clustering process see Dixit et al. (2021)). In the absence of the actual origin location, all routes between the origin–destination stop-clusters are assumed to be equally accessible for the travellers. However, for the origin–destination pairs such

Table 4 Aggregate measures of predictive ability for locally estimated and transferred models

Statistic	Locally estimated model	Transferred model	Difference
MAE per route	43.2	45.4	5.0%
MAE per link (AM peak)	302.2	328.2	8.6%

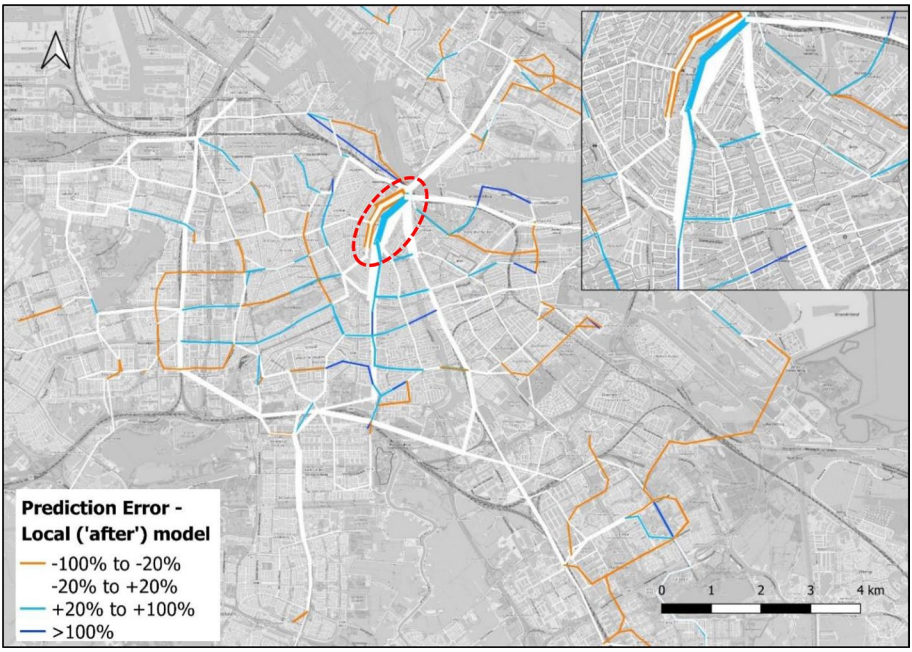

Fig. 3 Percentage error in prediction per link for the locally estimated model

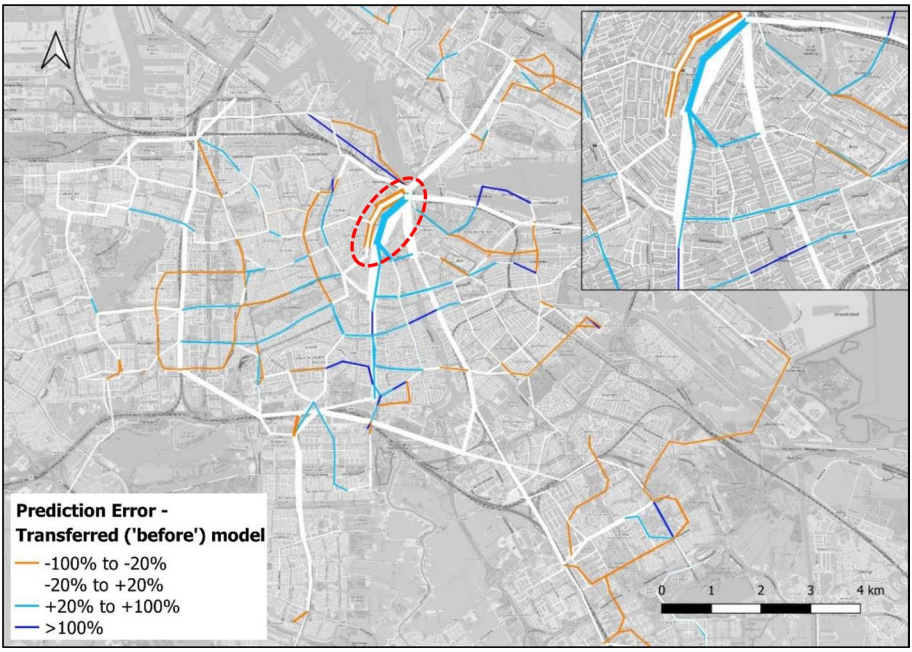

Fig. 4 Percentage error in prediction per link for the transferred model

Table 5 Observed and predicted mode shares

Mode combination	Observed share	Predicted share		Error	
		Local model	Transferred model	Local model	Transferred model
Bus only	18.4%	18.6%	18.1%	0.2%	−0.3%
Tram only	31.8%	31.6%	31.4%	−0.2%	−0.4%
Metro only	37.8%	37.8%	37.7%	0.0%	−0.1%
Bus + tram	0.4%	0.4%	0.4%	0.0%	0.0%
Bus + metro	3.3%	3.1%	3.1%	−0.2%	−0.2%
Tram + metro	8.2%	8.4%	9.3%	0.2%	1.1%
All modes	0.0%	0.0%	0.0%	0.0%	0.0%
MAPE				0.1%	0.3%

Table 6 Predictive measures of transferability for alternate transferred model specifications

Statistic	Generic MNL	Mode-specific MNL	Including overlap (PSCL)	Local model (Mode-specific MNL)
Number of parameters	7	11	13	11
Log-likelihood of estimation ('before' context)	−234,899	−233,892	−233,473	–
Log-likelihood of prediction ('after context')	−335,470	−330,299	−330,598	−328,646
First preference recovery	71.2%	71.5%	71.6%	71.7%
Brier score	0.379	0.372	0.372	0.370
MAE per route	48.2	45.4	45.2	43.2
MAPE (mode-share)	0.60%	0.30%	0.34%	0.14%

as the one highlighted where the distance travelled is very short, travellers are likely to choose the transit stop closest to them as opposed to the one with the shortest generalized cost that is predicted by the model. Hence, in such cases the link-level predictions can be erroneous, and should be used with caution.

Next, we compare the market shares of each transit mode combination in the data (Table 5). The predicted and observed shares are found to be close to each other with a difference of less than 1 percent for most mode combinations for both local and transferred models. As expected, the MAPE is found to be slightly higher for the transferred model than for the local model. Overall, the transferred model is found to perform close to the local model in terms of mode-share predictions as well.

Impact of omitted variables

In this section, we analyse the impact of omitting/adding one or more variables on the model's predictive performance (Table 6). We test two scenarios:

1. Generic MNL: Generic travel time and transfer parameters as opposed to mode-specific ones as specified in Eq. 2.
2. Including overlap: Path size correction logit (PSCL) model including path size correction terms as defined in Eq. 3 to incorporate the impact of overlap between alternate routes

Both in the estimation (before) and the prediction (after) contexts, the model with generic travel time and transfer parameters has the worst fit for the data, as shown by the respective log-likelihood values (even when adjusted for the number of parameters). The predictive performance is also found to suffer significantly when generic attributes are used. Conversely, when overlap is incorporated, the model fit is improved significantly in the estimation context (Likelihood ratio statistic of 838.6 exceeding the critical $\chi 2$ value of 9.2 at 1% significance level (df = 2)), but the log-likelihood for the prediction context is found to be lower than the mode-specific MNL model. In terms of predictions, there is a marginal improvement in the FPR and MAE for route-level prediction. However, the Brier score and the predictions at mode level are slightly worse.

When inferring route choice using revealed preference data sources in general, and smart card data in particular, the analyst does not have any 'direct' information on which attributes were considered by the decision-maker when making the choice. Hence, the selection of attributes to be included in the model depends heavily on the judgement of the analyst, and often data availability. It is known that the omission of a relevant variable can impact the model transferability (Koppelman and Wilmot 1986), especially when the missing variable is a confounding one. Conversely, if a simple model with fewer variables can perform just as well, then excluding variables can make the data collection as well as estimation easier. In our case, generic travel time and transfer parameters negatively impact the predictive ability of the transferred models. In contrast, including overlap does not offer a clear improvement in the predictive ability. In the end, the optimal selection of attributes depends on the purpose for which it is intended to be used. For predicting passenger flows in the regions where the tram was replaced by the new metro line, all attributes that distinguish a metro from a tram should ideally be included in the model. The mode-specific MNL shows that the travel time and transfer parameters are different for different modes. Hence, using generic travel time and transfer parameters impacts the predictive performance of the model significantly. On the other hand, correcting for route overlap typically leads to an improvement in model fit in the case of route choice models (like in our case for the estimation context). However, our results seem to suggest that it may not necessarily increase the transferability of the models, and the overlap term(s) may be context-specific and hence not as transferable.

Conclusion

This study adds to the scarce literature on the validation of travel demand models and is the first to undertake an external validation for a transit route choice model. The model was developed based on smart card data for the urban transit network of Amsterdam and was used to predict the impact of a significant network change (i.e. the introduction of a new metro line) on the route choice behaviour of travellers. Validation was conducted

by comparing the parameter values and a series of statistical performance indicators for the predictions with the observed behaviour after the network change.

Our results are overall in agreement with existing literature: the conclusion regarding model transferability depends on the (statistical) test used (Koppelman and Wilmot 1982). In our case, model parameter equality failed for most attributes, implying care should be taken in directly inferring behavioural insights from the parameter values from models such as those used in this study, specifically if they have not been thoroughly validated. However, the predictive performance of the transferred model was found to be close to the locally estimated model. When compared with the observed choices at an individual level, the model performed satisfactorily with a First Preference Recovery of 71.5%. Moreover, the predicted mode-shares were close to the observed ones, with a MAPE of 0.3%. When used for route and link level predictions, the errors were relatively larger, but the performance of the transferred model was similar to the local model (less than 10% error increase). We also investigated the impact of omitting relevant variables on predictive performance. When the mode-specific travel time and transfer parameters were replaced by generic ones, the performance suffered significantly. Conversely, including overlap in the model specification did not offer a clear improvement in model predictions, even though it had a better fit for the base data. This suggests that overlap definition may be context specific and could perhaps be excluded when using a route choice model for predictions in favour of a parsimonious model.

When using smart card data for travel demand modelling, several assumptions are made regarding travellers' consideration choice set and perceived travel attributes. Visualizing link-level prediction errors can help indicate potential causes of errors. In our case, the assumption regarding consideration choice set may be responsible for some of the prediction errors, which are consistent between local and transferred models.

"All models are wrong, but some are useful" (Box 1976). To establish how wrong a model needs to be to stop being useful, we need more studies undertaking validation analysis for different networks and policy scenarios. Guidelines and standards on what is considered acceptable in terms of the various transferability statistics remain yet to be defined. In the past, the cost of undertaking model validation was high primarily due to data collection costs. The abundance of passively collected data such as the smart card provides an opportunity to validate transit route choice and assignment models at relatively low additional costs. Hence, validation must become an integral part of the development process for such models, and should be considered non-negotiable when using them for deriving any policy recommendations.

Acknowledgements This research was funded by the municipality of Amsterdam, Vervoerregio Amsterdam and the AMS Institute. We thank GVB for providing data for this research.

Author contributions MD: Conceptualization, Methodology, Formal Analysis, Writing—Original draft preparation. OC: Conceptualization, Methodology, Supervision, Writing—Reviewing and Editing. NvO: Conceptualization, Supervision, Writing—Reviewing and Editing. TB: Data curation, Writing—Reviewing and Editing. SH: Writing—Reviewing and Editing.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons

licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Atherton, T., Ben-Akiva, M.: Transferability and updating disaggregate travel demand models Transp. Res. Rec. J. Transp. Res. Board. 610 (1976)

Badoe, D.A., Miller, E.J.: Analysis of temporal transferability of disaggregate work trip mode choice models Transp. Res. Rec. **1**, 11 (1995)

Bekhor, S., Prato, C.G.: Methodological transferability in route choice modeling. Transp. Res. Part B Methodol. **43**, 422–437 (2009). https://doi.org/10.1016/J.TRB.2008.08.003

Bovy, P.H.L., Hoogendoorn-Lanser, S.: Modelling route choice behaviour in multi-modal transport networks. Transportation **32**, 341–368 (2005). https://doi.org/10.1007/s11116-004-7963-2

Box, G.E.P.: Science and statistics. J. Am. Stat. Assoc. **71**, 791–799 (1976). https://doi.org/10.1080/01621459.1976.10480949

Brands, T., Dixit, M., van Oort, N.: Impact of a new metro line in amsterdam on ridership, travel times, reliability and societal costs and benefits. Eur. J. Transp. Infrastruct. Res. **20**, 335–353 (2020)

Brier, G.W.: Verification of forecasts expressed in terms of probability. Mon. Weather Rev. **78**, 1–3 (1950)

Chorus, C.G., Kroesen, M.: On the (im-) possibility of deriving transport policy implications from hybrid choice models. Transp. Policy. **36**, 217–222 (2014). https://doi.org/10.1016/J.TRANPOL.2014.09.001

de Luca, S., Cantarella, G.E.: Validation and Comparison of Choice Models. In: Sammer, G. and Saleh, W. (eds.) Travel Demand Management and Road User Pricing. pp. 57–78. Routledge (2016)

Dixit, M., Brands, T., van Oort, N., Cats, O., Hoogendoorn, S.: Passenger travel time reliability for multimodal public transport journeys. Transp. Res. Rec. J. Transp. Res. Board. **2673**, 149–160 (2019). https://doi.org/10.1177/0361198118825459

Dixit, M., Cats, O., Brands, T., van Oort, N., Hoogendoorn, S.: Perception of overlap in multi-modal urban transit route choice. Transp. A Transp. Sci. (2021). https://doi.org/10.1080/23249935.2021.2005180

Forsey, D., Nurul Habib, K., Miller, E.J., Shalaby, A.: Temporal transferability of work trip mode choice models in an expanding suburban area. Transp. A Transp. Sci. **10**, 469–482 (2014). https://doi.org/10.1080/23249935.2013.788100

Fox, J.B.: Temporal transferability of mode-destination models, (2015)

Fox, J., Hess, S.: Review of evidence for temporal transferability of mode-destination models. Transp. Res. Rec. (2010). https://doi.org/10.3141/2175-09

Fox, J., Daly, A., Hess, S., Miller, E.: Temporal transferability of models of mode-destination choice for the greater Toronto and Hamilton area. J. Transp. Land Use. (2014). https://doi.org/10.5198/jtlu.v7i2.701

Garcia-Martinez, A., Cascajo, R., Jara-Diaz, S.R., Chowdhury, S., Monzon, A.: Transfer penalties in multimodal public transport networks. Transp. Res. Part A Policy Pract. **114**, 52–66 (2018). https://doi.org/10.1016/J.TRA.2018.01.016

Gunn, H.F., Ben-Akiva, M.E., Bradley, M.A.: Tests of the scaling approach to transferring disaggregate travel demand models. Transp. Res. Rec. **1037**, 21–30 (1985)

Guo, Z., Wilson, N.H.M.: Assessing the cost of transfer inconvenience in public transport systems: a case study of the London underground. Transp. Res. Part A Policy Pract. **45**, 91–104 (2011). https://doi.org/10.1016/j.tra.2010.11.002

Hörcher, D., Graham, D.J., Anderson, R.J.: Crowding cost estimation with large scale smart card and vehicle location data. Transp. Res. Part B Methodol. **95**, 105–125 (2017). https://doi.org/10.1016/j.trb.2016.10.015

Jánošíkova, L., Slavík, J., Koháni, M.: Estimation of a route choice model for urban public transport using smart card data. Transp. Plan. Technol. **37**, 638–648 (2014). https://doi.org/10.1080/03081060.2014.935570

Kim, K.M., Hong, S.P., Ko, S.J., Kim, D.: Does crowding affect the path choice of metro passengers? Transp. Res. Part A Policy Pract. **77**, 292–304 (2015). https://doi.org/10.1016/j.tra.2015.04.023

Kim, I., Kim, H.-C., Seo, D.-J., Kim, J.I.: Calibration of a transit route choice model using revealed population data of smartcard in a multimodal transit network. Transportation (2019). https://doi.org/10.1007/s11116-019-10008-8

Koppelman, F.S., Wilmot, C.G.: Transferability analysis of disaggregate choice models. Transp. Res. Rec. J. Transp. Res. Board. **895**, 18–24 (1982)
Koppelman, F.S., Wilmot, C.G.: The effect of omission of variables on choice model transferability. Transp. Res. Part B Methodol. **20**, 205–213 (1986). https://doi.org/10.1016/0191-2615(86)90017-2
Lai, X., Bierlaire, M.: Specification of the cross-nested logit model with sampling of alternatives for route choice models. Transp. Res. Part B Methodol. **80**, 220–234 (2015). https://doi.org/10.1016/J.TRB.2015.07.005
Liu, Y., Bunker, J., Ferreira, L.: Transit users route-choice modelling in transit assignment: a review. Transp. Rev. (2010). https://doi.org/10.1080/01441641003744261
Mai, T.: A method of integrating correlation structures for a generalized recursive route choice model. Transp. Res. Part B Methodol. **93**, 146–161 (2016). https://doi.org/10.1016/J.TRB.2016.07.016
Nielsen, O.A., Eltved, M., Anderson, M.K., Prato, C.G.: Relevance of detailed transfer attributes in large-scale multimodal route choice models for metropolitan public transport passengers. Transp. Res. Part A Policy Pract. **147**, 76–92 (2021). https://doi.org/10.1016/J.TRA.2021.02.010
Parady, G., Ory, D., Walker, J.: The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: a review of validation practices in the transportation academic literature. J. Choice Model. (2021). https://doi.org/10.1016/j.jocm.2020.100257
Rossi, T.F., Bhat, C.R.: Guide for travel model transfer (No. FHWA-HEP-15–006). (2014)
Sanko, N., Morikawa, T.: Temporal transferability of updated alternative-specific constants in disaggregate mode choice models. Transportation **37**, 203–219 (2010). https://doi.org/10.1007/s11116-009-9252-6
Swait, J., Louviere, J.: The role of the scale parameter in the estimation and comparison of multinomial logit models. J. Mark. Res. **30**, 305–314 (1993)
Swierstra, A.B., van Nes, R., Molin, E.J.E.: Modelling travel time reliability in public transport route choice behaviour. Eur. J. Transp. Infrastruct. Res. **17**, 263–278 (2017)
Train, K.: A validation test of a disaggregate mode choice model. Transp. Res. **12**, 167–174 (1978). https://doi.org/10.1016/0041-1647(78)90120-X
van Oort, N., Brands, T., de Romph, E.: Short-term prediction of ridership on public transport with smart card data. Transp. Res. Rec. **2535**, 105–111 (2015). https://doi.org/10.3141/2535-12
Yap, M., Cats, O., van Arem, B.: Crowding valuation in urban tram and bus transportation based on smart card data. Transp. A Transp. Sci. **16**, 23–42 (2020). https://doi.org/10.1080/23249935.2018.1537319

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Malvika Dixit completed her PhD in 2022 at the Delft University of Technology. Her doctoral work focused on transit performance assessment and route choice modelling using smart card data. She won the Young Researcher of the Year 2022 award by the International Transport Forum (ITF) for her research on inclusive mobility conducted as part her PhD. She currently works at Amazon, Luxembourg, contributing to the long-term planning of their end-to-end transportation network for Europe.

Oded Cats is a Professor of Passenger Transport Systems at Delft University of Technology (TU Delft). He is the Deputy Head of the Department of Transport and Planning and a codirector of the Smart Public Transport Lab. He is a regular speaker at international conferences and has (co-)authored more than 150 peer-reviewed scientific journal papers. His research activities support public transport agencies and operators' decision making and has led to field implementations. His research is devoted to developing theories and models of multi-modal passenger transport networks by combining advancements in simulation and operations research, behavioural sciences and complex network theory and modelling. The domain of application for most of his work is metropolitan public transport systems where he focuses on network dynamics and robustness, service operations and control, passenger demand and flow distributions, travel information and service uncertainty. He has developed an agent-based dynamic public transport operations and assignment model called BusMezzo that has been used in a large range of research projects and real-world applications. He is the recipient of a European Research Council Starting Grant entitled CriticalMaaS. He is leading several European, national and industry research projects and is a member of several committees of the US Transportation Research Board. He has experience in managing European and national research projects and supervising PhD students. He is the Editor-in-Chief of the European Journal of Transport and Infrastructure Research. He teaches transport modelling and public transport networks and demand modelling at bachelor, master and PhD levels as well as part of training programs designed for young researchers

and professionals. Oded holds a dual- PhD from KTH Royal Institute of Technology, Sweden, and Technion—Israel Institute of Technology.

Niels van Oort works as an associate professor Public Transport at Delft University of Technology and is co-director of the Smart Public Transport Lab. He has been involved in public transport projects and research for over 15 years. His main fields of expertise are public transport planning and (data-driven) design, looking at the passenger perspective and societal impacts, related to for instance inclusiveness, land-use and sustainability. His application inspired work also involves shared mobility and first/last mile solutions. In addition to teaching, he is a frequently invited speaker and he published numerous articles in (international) journals and general media. He studied traffic and transport at Delft University of Technology and finished his master on Public Transport in 2003. He started to work as a researcher at HTM, the public transport company of The Hague. During this job, he started a part-time PhD research in Delft on service reliability, including joint work with Professor Wilson at MIT. In 2010, he changed jobs to work as a public transport consultant at Goudappel Coffeng mobility consultants. He finished his PhD in 2011 and started to work at Delft university of Technology in 2012 as a part-time assistant professor. Since 2018, after founding the Smart Public Transport Lab, he is full time employed. Since 2020, he has been involved in small projects via his company Van Oort Mobility Consultancy.

Ties Brands currently works as a Consultant Public Transport at Nederlandse Spoorwegen (NS). He holds a MSc degree at the University of Twente in Civil Engineering and Applied Mathematics. His MSc thesis was on optimisation of dynamic road pricing and was conducted at the Dutch consultancy company Goudappel Coffeng in Deventer. After graduation, he started working at Goudappel Coffeng as a public transport consultant, specialised in public transport modelling and data analysis. In 2010, he started, parallel to his work as a consultant, a PhD project at the University of Twente, Centre for Transport Studies. In 2015, he defended his thesis with the title 'Multi-objective Optimisation of Multimodal Passenger Transportation Networks'. In the thesis, the resolution of measures related to multimodal trip making and corresponding network designs are identified by solving a multi-objective optimisation problem. Methods to analyse the resulting Pareto set provide insight into how total travel time, CO2 emissions, urban space used by parking and costs relate. Between 2017- and 2021, he worked as a part-time postdoc at TU Delft working on the project studying the impact of the Noord/Zuidlijn: a metro connection in Amsterdam that started operation in the summer of 2018. It was mainly data-driven research, with aspects such as travel times, service reliability, ridership, passenger flows through the network and traveller's perception.

Serge Hoogendoorn was appointed Antonie van Leeuwenhoek professor Traffic Operations and Management in 2006. He has been the chair of the Department of Transport and Planning since 2018 and is currently (one of the four) Distinguished Professor of Smart Urban Mobility at Delft University of Technology. He has a part-time appointment at Monash University, and is an Honorary Professor in the School of Transportation at South East University in China. He is a distinguished research fellow position at RIOH (Beijing). He is the PI Mobility in the Institute of Advanced Metropolitan Solutions (www.ams-amsterdam.com), a staff member of the TRAIL Research School on Transport and Logistics at DUT, and he chairs the Network Management foundation. He completed his PhD at Delft University of Technology in 1999. His current research evolves around Smart Urban Mobility, with focal areas such as i) theory, modelling, and simulation of traffic and transportation networks, including cars, pedestrian, cyclists and novel public transport services (e.g. Demand Responsive Transit in combination with traditional PT); ii) development of methods for integrated management of these networks (regional network management, crowd and bicycle management; public transport operations); iii) impact of uncertainty of travel behaviour and network operations; iv) impact of ICT on network flow operations, robustness and resilience,; and v) urban data and their applications. In all these topics, his work considers both recurrent and non-recurrent (emergency) situations.