Journal of Choice Modelling 38 (2021) 100257

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: http://www.elsevier.com/locate/jocm

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

Giancarlos Parady [a,*], David Ory [b], Joan Walker [c]

[a] Department of Urban Engineering, The University of Tokyo, Japan
[b] WSP, USA
[c] Department of Civil and Environmental Engineering, University of California, Berkeley, USA

ARTICLE INFO

Keywords:
Validation
Generalizability
Transferability
Policy inference
Transportation
Discrete choice models

ABSTRACT

An examination of model validation practices in the peer-reviewed transportation literature published between 2014 and 2018 reveals that 92% of studies reported goodness-of-fit statistics, and 64.6% reported some sort of policy-relevant inference analysis. However, only 18.1% reported validation performance measures, out of which 78% (14.2% of all studies) consisted of internal validation and 22% (4% of all studies) consisted of external validation. The proposition put forward in this paper is that the reliance on goodness-of-fit measures rather than validation performance is unwise, especially given the dependence of the transportation research field on observational (non-experimental) studies. Model validation should be a non-negotiable part of presenting a model for peer-review in academic journals. For that purpose, we propose a simple heuristic to select validation methods given the resources available to the researcher.

1. Introduction

Ioannidis (2005) brought to light the issue of lack of demonstrated reproducibility in the natural sciences and, in subsequent years, the so-called "reproducibility crisis" in science has made headlines in the popular media (Baker and Penny, 2016). The transportation domain can benefit from thinking about the underlying concerns about reproducibility.[1] Though the transportation field generally does not rely on experiments that can be readily retried, to be useful, the observational studies that we generally do rely on should generalize across time and space within reasonable boundaries. To do so, we argue that more attention needs to be paid to the validity of the models used to extract information from observational studies.

The proposition put forward by this paper is that the transportation community over-relies on statistical goodness-of-fit when assessing a model's performance and policy relevance. For example, it is common in the transportation space for models to be estimated from an observational study of travel behaviour. An analyst, to continue the example, may use such a data set revealing travel mode choice decisions and strive to quantify the relative onerousness of waiting for a bus relative to riding on a bus. Coefficients from a statistical model, say linear regression or discrete choice, may suggest that waiting is three times as unpleasant as riding. A typical analyst would take comfort in robust goodness-of-fit statistics, e.g., an R-squared or pseudo-rho-squared in line with similar studies, as

* Corresponding author. The University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, Japan.
E-mail address: gtroncoso@ut.t.u-tokyo.ac.jp (G. Parady).

[1] Note than in experimental studies the concept of reproducibility differs from the one commonly used in observational studies, and in this article.

https://doi.org/10.1016/j.jocm.2020.100257
Received 19 February 2020; Received in revised form 24 October 2020; Accepted 31 October 2020
Available online 5 November 2020

well as statistics showing the statistical significance of the coefficients for the key variables, e.g., a t-statistic suggesting the coefficients are likely not equal to zero. It is rare for an analyst to try and replicate these findings on a new sample or even against a "hold-out" from the same sample.

To understand how goodness-of-fit and validation is approached in the transportation literature, we review their use in the peer-reviewed transportation academic literature published between 2014 and 2018, focusing on models that use discrete choice. The balance of the paper is organized as follows: In Section 2, we discuss the reproducibility crisis introduced above and its applicability to the study of transportation. In Section 3, we put forth an operational definition of validation and related concepts, to organize the myriad of related terms in the field. In Section 4 and 5 we discuss the most commonly used validation methods, and performance indicators, respectively. The results from the meta-analysis we conduct of the academic literature are presented in Section 6. In Section 7 we provide recommendations for improving validation practices in the field and end with concluding thoughts in Section 8.

2. A credibility crisis in science and engineering?

In 2016, the journal Nature published the results of a survey of its readers regarding reproducibility (Baker and Penny, 2016). A key finding was that *"more than 70 percent of researchers tried and failed to reproduce another scientist's experiment, and more than half have failed to reproduce their own experiments."* (pp. 452) Furthermore, 52 percent of respondents stated that there is a significant reproducibility crisis.

This perception echoes the argument of Ioannidis (2005), who estimated the positive predictive value of research – that is, the probability that a reported finding is true – under different criteria. Ioannidis suggested that most published research findings are likely to be false, due to factors such as lack of statistical power of the study, small effect sizes, and great flexibility in research design, definitions, outcomes, and methods.

While the study by Ioannidis focused on experimental studies in natural science, the underlying concern is of relevance to the transportation research field.

Generally speaking, the purpose of academic transportation research is to better understand (and ultimately forecast) transport-related human behaviour to better inform transportation policy design and implementation. However, conducting experiments in the transportation field is often expensive and/or disruptive to transport users. Consider, for example, running an experiment in which, a new subway line is constructed just to evaluate modal shift from automobile to transit. Although there are some examples of experimental economics applications (e.g. Holguín-Veras et al., 2004) the vast majority of transportation research relies on cross-sectional observational studies, which makes hypothesis testing in the classical way difficult.

Lack of validation of models increase the risk of model overfitting, where the model fits the estimation data well but performs poorly outside this estimation dataset, in other words, the model is fitted to the noise instead of the signal in the data.

Despite the limitations imposed by observational studies, impact evaluation of policies drawn based on model-based academic research is rarely conducted, meaning there is little feedback, if any, in terms of how right or how wrong are these models and the policy recommendations derived from them. Altogether, these issues strongly underscore the need to incorporate in the analysis means to evaluate the validity of results. However, we will show that the academic literature has over-relied on statistical goodness of fit and widely disregarded model validation.

3. Defining validation

The meaning of the term validation differs across fields, and even within fields, and certainly in the case of transportation, there is no agreed-upon definition. As such, to organize the myriad of validation-related terms, after reviewing the validation literature in different research fields, we propose a set of definitions that are adequate for the transportation field. Fig. 1 summarizes these concepts and illustrate the relationship among them.

We will start by defining six key terms adapted from Justice et al. (1999):

- *Predictive accuracy*: the degree to which predicted outcomes match observed outcomes. Predictive accuracy is a function of:
  - o *Calibration*[2] *ability:* the ability of a model or system of models to make predictions that match observed outcomes.
  - o *Discrimination ability:* the ability of a model or system of models to discriminate between those instances with and without a particular outcome.
- *Generalizability:* the ability of a model, or system of models to maintain its predictive accuracy in a different sample. The generalizability of a model is a function of:
  - o *Reproducibility*: the ability of a model, or system of models to maintain its predictive ability in different samples from the same population.
  - o *Transferability:* the ability of a model, or system of models to maintain its predictive ability in samples from different but plausibly related populations or in samples collected with different methodologies. In other fields it is also called transportability.

Given that the purpose of academic transportation research is to better understand (and ultimately forecast) transport-related

[2] Note that in the transportation field the term calibration is commonly used to refer to the adjustment of constants or other parameters to match observed outcomes.

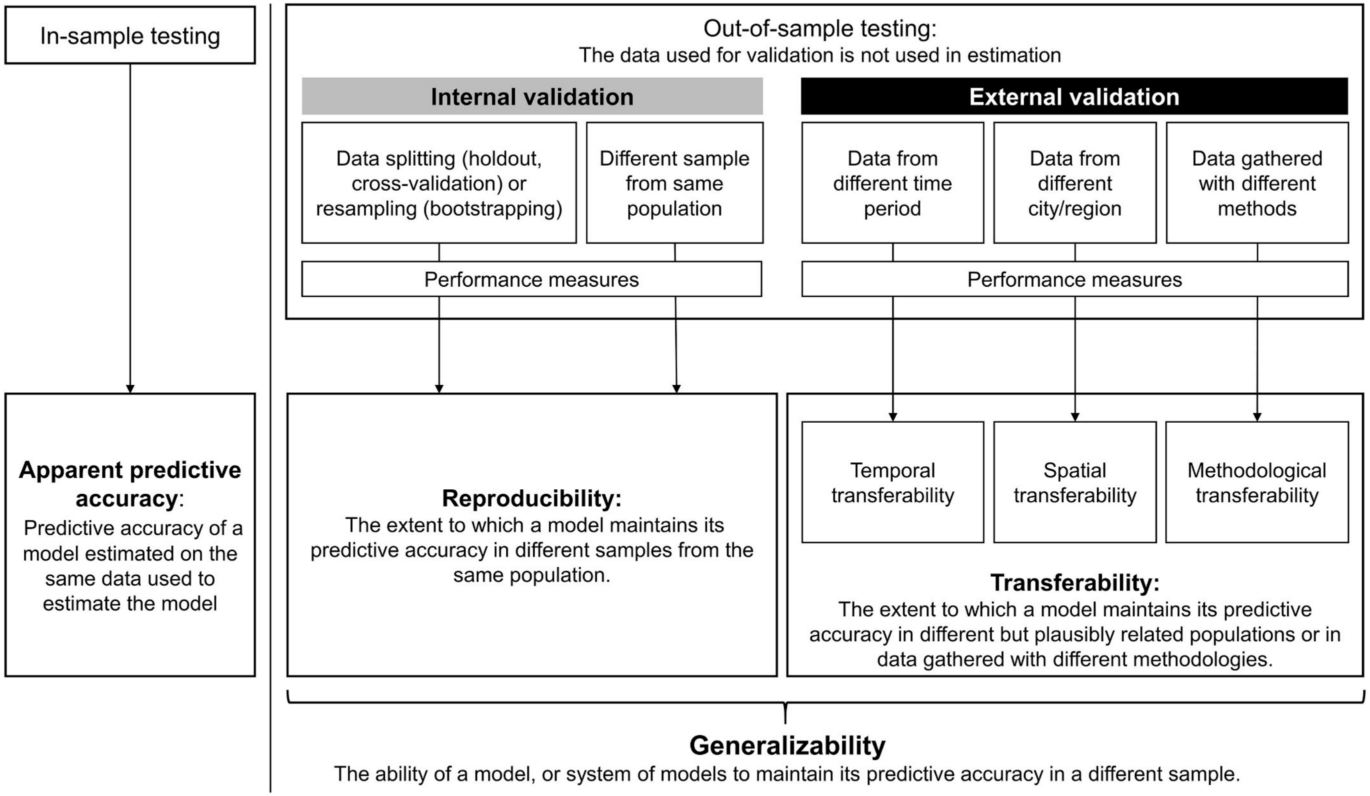

Fig. 1. Model validation is the evaluation of the generalizability of a statistical model. The evaluation of reproducibility is called internal validation, while the evaluation of transferability is called external validation. In-sample testing is not validation.

human behaviour to better inform transportation policy design and implementation, the usefulness of a statistical model is a function of its predictive accuracy outside the sample used to estimate it. We can thus define model validation as the evaluation of the generalizability of a statistical model. The evaluation of reproducibility is called internal validation, while the evaluation of transferability is called external validation. Note that internal validity is very often a precondition for external validity (Justice et al., 1999).

This definition outright excludes the classification of in-sample testing such as goodness-of-fit statistics from being considered a form of model validation. The "optimism" of in-sample estimates of predictive accuracy or apparent predictive accuracy is a well-documented phenomena (Efron, 1983; Steyerberg et al., 2001), which stems from the fact that the sample data is used for both estimation and testing. Correcting for this is the very reason internal and external validation methods have been developed.

Irrespective of the type of validation, the process itself is conceptually simple. It consists of (i) estimating a model with an estimation sample e (sometimes called training sample), (ii) applying the estimated model to a validation sample v (sometimes called testing sample), and (iii) evaluating the generalizability of the estimated model given the defined performance measure(s) of predictive accuracy (See Section 5). Predictive accuracy is usually quantified as a function of the discrepancy between predicted and observed outcomes (i.e. prediction error).

In the case of internal validation, while using independent samples is ideal, in many cases, producing such data is expensive, so data-splitting methods such as holdout and cross-validation, or resampling methods such as bootstrapping are often used. These methods will be discussed in Section 4.

Regarding the evaluation of transferability (external validation), in the transportation field, a significant number of studies were conducted during the 1980s on the subject (Ortúzar and Willumsen, 2011), even though it has *"dropped off the radar"* in recent years (Fox et al., 2014). In transportation, transferability has been defined as the *"usefulness of a model, information or theory in a new context"* (Koppelman and Wilmot, 1982) and as *"the ability of a model developed in one context to explain behaviour in another under the assumption that the underlying theory is equally applicable in the two contexts"* (Fox et al., 2014), definitions that are consistent with the definition presented above. Past studies on transferability have not only focused on predictive accuracy but also on the stability of parameters across contexts, especially in earlier studies (Koppelman and Wilmot, 1982; Ortúzar and Willumsen, 2011; Fox et al., 2014).

In Transportation, the two key dimensions of interest are temporal transferability and spatial transferability which refer to the potential to transfer a model to different points in time, and to different spatial contexts (i.e. cities, regions), respectively.

There are, of course, limits to how much a model can be transfered, and it is obvious that in transportation research, models are context dependent (Ortúzar and Willumsen, 2011), as such, rather than a pass/fail test, transferability is a matter of degree (Koppelman and Wilmot, 1982) and should be bounded by reasonable limits. No one would expect, for example, a discretionary activity destination choice model estimated for Tokyo to generalize well to Los Angeles. It would be hard to argue that these populations are "plausibly related." A similar argument can be made in terms of temporal transferability, which should be evaluated in general, considering the time frame of the forecast of interest. Consider, for example, a 20-year forecast, a timeframe typical of transportation forecast models. In such case, two independent samples from the same city and collected six months apart, are more likely different (yet contemporary) samples from the same population than temporally different samples. That is, they are too similar to constitute a proper external validation dataset. As such, the validation effort would in fact be an internal validation test.

Significant policy interventions (i.e. the construction of a new expressway, or rail line), or high-impact external events (i.e. natural disasters, economic shocks, pandemics) can also be thresholds to make a population temporally different. As such, samples from before and after such an event can be considered temporally different samples.

As an aside, Justice et al. (1999) have also discussed, in the context of prognostic assessments, the issue of methodological transferability, which refers to model performance on data collected with different methodologies (i.e. different variable definitions, survey methods, etc.) Although potentially relevant to the field, this is a largely understudied issue, but included in this article for completeness.

While it can be argued that as an analysis of model sensitivity, the calculation of elasticities and marginal effects are a part of the model validation process (Cambridge Systematics, 2010), as measures of effect size, these are key policy-related values. As such, although the distinction might seem trivial, we classify these values as part of the policy-relevant inference analysis rather than part of the validation process, which focuses, as per the definition provided above, on predictive ability.

4. Data splitting and resampling methods for validation

Data splitting and resampling methods have become common methods to conduct validation, largely due to the high costs associated with producing independent datasets to test models against. While these methods are largely used for internal validation, it is possible to adapt these methods for external validation, although not very common (e.g. Austin et al. (2016) in epidemiology, and Sanko (2017) in transport). We will now briefly introduce these methods.

4.1. Holdout validation

Holdout validation (HOV) is the simplest data splitting method. In holdout validation, the dataset is randomly split into an estimation dataset and a validation dataset (holdout). For illustration purposes, let us define $\mathrm{Q}[y_n, \widehat{y}_\mathrm{n}]$ as a measure of prediction correctness for the nth instance, for the binary choice case as:

$$Q\left[y_n, \widehat{y}_{\mathrm{n}}\right] = \begin{cases} 0 \; if \; y_n = \widehat{y}_{\mathrm{n}} \\ 1 \; if \; y_n \neq \widehat{y}_{\mathrm{n}} \end{cases} \tag{1}$$

where y_n is the observed outcome, and $\widehat{y}_{\mathrm{n}}$ is the predicted outcome for instance n. The holdout estimator is

$$HOV = \frac{1}{N_v} \sum_{n_v=1}^{N_v} Q\left[y_{n_v}, \widehat{y}_{n_v}^{e}\right] \tag{2}$$

where $\widehat{y}_{n_v}^{e}$ is the predicted outcome for instance n in sample v, using the model estimated with sample e, and N_v is the validation sample size. The performance measure in equation (2) is called the misclassification rate, but as will be discussed in Section 5, there is myriad of performance measures that can be used to evaluate predictive accuracy.

4.2. Cross-validation methods

When the holdout process is repeated multiple times, thus generating a set of randomly split estimation-validation data pairs, we refer to the validation procedure as cross-validation (CV). The basic cross-validation estimator is defined as

$$CV = \frac{1}{\mathrm{B}} \sum_{b=1}^{B} HOV_b \tag{3}$$

where B is the number of estimation-validation data pairs generated, andHOV_b is the holdout estimator for set b. Cross-validation methods differ from one another in the way the data is split. When the data splitting considers all possible estimation sets of size $\mathrm{N_c}$, the splitting is exhaustive, otherwise the splitting is partial. Partial data splitting methods have the advantage of being less computationally intensive than exhaustive splitting. Here we will briefly define common cross-validation methods. We refer readers to the work of Arlot and Celisse (2009) for a more comprehensive review of cross-validation practices in general.

Regarding exhaustive splitting, in the leave-one-out (LOO) approach (Stone, 1974) sometimes referred to as the jacknife approach, the estimation set size is $\mathrm{N_c} = N - 1$, and $B = N$. That is, the model is fitted leaving out one instance per iteration, and the outcome of that single instance is predicted based on the estimated model. In the leave-p-out (LPO) approach (Shao, 1993) the estimation set size is $\mathrm{N_c} = N - p$.

Regarding partial data splitting, in the K-fold cross-validation (K-CV) approach (Geisser, 1975), data is partitioned into K mutually-exclusive subsets of roughly equal size, and $B = K$. It can be seen that the particular case of N-fold cross validation is equivalent to the leave-one-out approach. In the case of repeated K-fold cross-validation, the process is repeated R times and the CV performance measures averaged over R. In the repeated learning-testing (RLT) approach (Burman, 1989) a B number of randomly-split estimation-validation pairs are generated. This method is also called repeated holdout validation.

As an aside, it is important to note that in-sample statistics have been proposed to aid on model selection, such as the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC). Stone (1977) showed the asymptotic equivalence between cross validation (specifically, the leave-one-out method) and the AIC statistic for model selection. However, the strength of validation relies on the quasi-universality of its applicability and its robustness to violations of assumptions necessary for these statistics to be correct (Efron and Tibshirani, 1993; Arlot and Celisse, 2009).

4.3. Bootstrapping methods

Bootstrapping validation methods were proposed by Bradley Efron to address some of the limitations of cross-validation methods. Although (leave-one-out) cross-validation gives a nearly unbiased estimate of predictive accuracy, it often exhibits unacceptably high variability, particularly when sample size is small, whereas bootstrapping methods have been shown to be more efficient (Efron, 1983; Efron and Tibshirani, 1995). We will briefly summarize some basic bootstrapping estimators borrowing from Efron and Tibshirani (1993, 1997) to whom we refer the reader for a more extensive treatment of bootstrapping for validation.

The idea of bootstrapping is conceptually simple. In the simplest case, given a dataset of size n, a bootstrap sample is generated by randomly sampling (with replacement) n instances from the original dataset and estimating the model on this sample. A prediction error estimate for this sample can be obtained by applying the model to the original sample. This process is repeated B times, and the prediction error averaged over B to obtain the simple bootstrap prediction error estimate,

$$BS_{simple} = \frac{1}{\mathrm{B}} \sum_{b=1}^{B} \sum_{n=1}^{N} Q\left[y_n, \widehat{y}_n^{b}\right] \Big/ N \tag{4}$$

Another estimator is the leave-one-out bootstrap estimator, defined as

$$BS_{loo} = \frac{1}{\mathrm{N}} \sum_{n=1}^{N} \sum_{b \varepsilon C_n} Q\left[y_n, \widehat{y}_n^{b}\right] \Big/ B_n \tag{5}$$

where, given a set of B bootstrap samples, C_n is the subset of bootstrap samples not containing instance n, and B_n is the size of C_n. This estimator can be considered a smoothed version of the leave-one-out cross-validation estimator, and while being more efficient, it has been shown to be upward-biased (Efron and Tibshirani, 1995).

A more refined approach is to get bootstrap estimates of the optimism of the apparent prediction error (the prediction error estimated on the same sample used to estimate the model) and correct for it. The bootstrap estimator can then be defined as

$$BS = \overline{err} + \widehat{\omega} \tag{6}$$

where $\overline{err}$ is the apparent prediction error (of the original sample) defined as

$$\overline{err} = \frac{1}{N}\sum_{n=1}^{N} Q\left[y_n, \widehat{y}_n\right] \tag{7}$$

and $\widehat{\omega}$ is a measure of optimism. One way to estimate $\widehat{\omega}$ is as the difference between the simple bootstrap estimate (Eq. (4)) and the apparent prediction error of the bootstrap samples. Another way is using the 0.632 optimism estimator defined as

$$\widehat{\omega}_{.632} = .632[BS_{loo} - \overline{err}] \tag{8}$$

Plugging Equation (8) into equation (6) results in the 0.632 bootstrap estimator (Efron, 1983)

$$BS_{.362} = .368\ \overline{err} + .632 \cdot BS_{loo} \tag{9}$$

This estimator mitigates the upward bias of the leave-one-out bootstrap estimate. The weights for this estimator come from the fact that the probability of any instance to be present in a bootstrap sample is 0.632. Although Efron discussed the 0.632 estimator in terms of the leave-one-out case, it can be generalized to non-exhaustive cases (Steyerberg et al., 2001).

Although bootstrapping is rather common in transportation for standard error estimation, it is rarely used for model validation. And while it has been suggested that bootstrapping is superior to cross-validation (Efron, 1983; Steyerberg et al., 2001), other studies have suggested that repeated K-fold cross-validation (Kim, 2009) or stratified cross-validation (Kohavi, 1995) are superior. The fact of the matter is that the performance of these methods is dependent on data characteristics such as sample size, and the type of model used. As such, we will refrain from making a recommendation, noting the need for an empirical study on the performance of different validation methods using typical data from transportation studies such as household travel surveys, and typical models.

5. Performance measures

In the transportation literature several articles have been published discussing in detail validation methods for discrete choice models from Train (1978, 1979) to Gunn and Bates (1982) and Koppelman and Wilmot (1982), and more recently de Luca and Cantarella, 2009). Using these studies as a departing point, in this section we briefly review the key performance measures reported in the transportation literature. Although the list of measures discussed here is not exhaustive, it is comprehensive in that it covers the vast majority of measures reported for internal and/or external validation in the reviewed studies.

One of the simplest ways to evaluate the predictive ability of a model is to compare the predicted market shares against the observed shares. While very simple and easy to understand, this approach does not provide a quantitative measure to evaluate the level of agreement between predictions and observations.

A quantitative measure often used is the percentage of correct predictions of a model (sometimes called the First Preference Recovery) where the alternative with the highest probability is defined as the predicted choice. Although this measure is widely reported, its use in models with more than two alternatives has been criticized, since it cannot differentiate between different probabilities assigned to a chosen alternative (de Luca and Cantarella, 2009). de Luca & Cantarella illustrate this point with a choice exercise with three alternatives a, b, c where the chosen alternative is a. With the percentage of correct predictions measure, a model that predicts choice ratios of 0.34, 0.33 and 0.33, for alternatives a, b, c, respectively, is equivalent to a model that predicts 0.90, 0.05, 0.05, even though the latter model assigns a considerably higher probability to the first alternative and the former is very close to a random prediction. To overcome this limitation, they proposed additional measures to evaluate "clearness of predictions," a concept that bears some resemblance to the concept of discriminative ability. They proposed the percentage of clearly right choices, the percentage of clearly wrong choices and the percentage of unclear choices.

They defined the percentage of clearly right choices as *"the percentage of users in the sample whose observed choices are given a probability greater than threshold t by the model"*. The idea here is that a model that predicts with higher probability a chosen alternative performs better than one that does so with a lower probability. Conversely, they defined the percentage of clearly wrong choices as the "*the percentage of users in the sample for whom the model gives a probability greater than threshold t to a choice alternative differing form the observed one.*" Finally, the percentage of unclear choices is the *"percentage of users for whom the model does not give a probability greater than threshold t to any choice."* To be meaningful, the threshold t must be "considerably larger" than c^{-1}, where c is the choice set size. While there is no agreed-upon definition of what qualifies as "considerably larger," if the threshold is just marginally larger to c, the test becomes useless. As a reference, when reporting the percentage of clearly right choices, de Luca and Di Pace (2015) used a threshold of 0.9 for binary choice models, while Glerum, Atasoy and Bierlaire (2014) used a threshold of 0.5 for choice models with three alternatives. When in doubt, we recommend reporting results for different threshold values. For example, de Luca and Cantarella (2009) reported values for 0.50, 0.66, 0.90 in tabular format for a pair of models with four alternatives, as well as a plot for all threshold values above 0.5.

Concordance probability (c-statistic) is a measure of discriminative ability, in the case of binary choices. This measure estimates the probability that an individual who was observed choosing alternative *a* is assigned a higher probability than an individual who did not. This probability is calculated as the ratio between the number of concordant pairs and comparable pairs. A comparable pair is a pair of individuals where one individual chose alternative *a*, and one did not. Such a pair is concordant if the individual who chose alternative *a* was assigned a higher probability of doing so than the individual who did not choose it. If the model has no discriminative ability the c-statistic equals 0.5 and equals 1 for perfect discriminant ability.

Several extensions have been proposed for the multinomial choice cases, see Calster et al. (2012) for a review of existing measures.

The fitting factor is defined as the ratio between the sum of predicted choice probabilities of the chosen alternatives and the number of individuals. A fitting factor of 1 indicates a perfect forecast of all choices with predicted probability of 1 (de Luca and Cantarella, 2009).

The correlation between prediction and outcomes can be used to evaluate predictive performance when dealing with continuous outcome such ridership levels, traffic flow, etc.

Other commonly used measures include the sum of square error (SSE), mean square error (MSE), root sum square error (RSSE), root mean square error (RMSE), mean absolute error (MAE), absolute percentage error (APE), and mean absolute percentage error (MAPE). Although there are some differences between these measures, there is no consensus regarding which measure is better. While quadratic measures like MSE, RSSE and RMSE tend to weight heavier on efficiency (Troncoso Parady et al., 2017) it has been argued that absolute measures like MAE are more natural and less ambiguous error measures than quadratic indicators (Willmott and Matsuura, 2005). It has also been suggested that RMSE is more appropriate when errors are expected to be Gaussian distributed (Chai and Draxler, 2014)

While in the transportation field it is common to use aggregate outcome measures such as market shares (Train, 1979) or choice frequencies (Koppelman and Wilmot, 1982) when using such measures, for most cases, individual outcomes can also be used. For example, the Brier score is calculated as the mean square difference between predicted individual choice probabilities and actual choices across all choices. To calculate this difference, the observed outcome is assigned a value of 1. This score has a minimum score of 0 for perfect forecasting, and a maximum score of 2 for the worst possible forecast (Brier, 1950). The Brier score can be thought of as a disaggregate form of the MSE for discrete outcomes. Disaggregate forms of MAE (de Luca and Cantarella, 2009) and RMSE can be calculated in a similar manner.

When using choice frequencies instead of market shares, it is possible to use the χ^2 test as a test of consistency between predictions and observations. The null hypothesis of this test is that observed and expected frequency outcome distributions are the same. Different to the other measures reviewed so far, this is a pass/fail test. Although the χ^2 statistic is sometimes used with market shares, it must be noted that this test is only valid for frequencies.

Since the likelihood is proportional to the product of individual probabilities, likelihood-based loss functions are also frequently used. The log-likelihood is a natural measure given that maximum likelihood estimators are widely used in discrete choice models. The cross-entropy measure, which is essentially the negative of the log-likelihood function, is also a commonly used loss function in machine learning.

The transferability test statistic (TTS) is a likelihood ratio test between the base model applied to the transfer data and the model estimated in the transfer data. This statistic is χ^2distributed with degrees of freedom equal to the number of parameters in the model. As with the regular χ^2 test, strictly speaking this is a pass/fail test. It tests whether the model parameters are equal across contexts. However, we strongly agree with Koppelman and Wilmot (1982) in that while such tests are useful to alert the analyst to differences between models, these differences should be interpreted against the acceptable error in each application context. This means focusing not on whether there is a difference or not, but how big is this difference.

Other likelihood-based indices are the transfer index (TI) and the transfer rho-square proposed by Koppelman and Wilmot (1982). The transfer index (TI) measures the degree to which the transferred model (a model estimated on sample *e* and transferred to transfer sample *v*) exceeds a local reference model (e.g. a market share model estimated on transfer sample *v*) relative to a model estimated directly on the transfer sample *v*. This measure has an upper-bound of one, in the case where the transferred model is as accurate as the local model, and takes negative values when the model is worse than the local reference model. While the transfer index is a relative measure of transferability, the transfer rho-square can be used as an absolute measure. This is the usual likelihood ratio index but used to evaluate performance of the transferred model against the local reference model. This index is upper-bounded by the local rho-square and has no lower bound. Negative values are interpreted in a similar manner to the transfer index.

Table 1 summarizes the performance measures described above and their respective equations. Measures are classified into absolute measures, relative measures, and pass/fail measures. While relative measures are useful for model selection, they do not give an absolute measure of predictive accuracy. Absolute measures, on the other hand, can be used to generate benchmark values against which researchers can evaluate, to a certain extent, the performance of their models against similar studies in the literature. It must be noted, however, that even when using absolute measures, model performance is relative to factors such as choice set size and the base market share of alternatives. As such, comparisons across different studies must be interpreted with a clear understanding of these limitations, and not as an absolute indicator of model superiority.

6. Validation and reporting practices in the transportation academic literature

Having summarized the basic validation process and the most used performance indicators in the literature, this section will review

Table 1
Definition of model validation performance measures reported in the literature.

Measure	Abbrv.	Type	Equation	Lower bound	Upper bound	Notes
Predicted vs observed outcomes	PVO	–	–	-[a]	-[a]	Simple comparison of predicted and observed outcomes or market shares. Usually in the form of a table or plot. No prediction accuracy statistics are calculated.
Percentage of correct predictions or First Preference Recovery	FPR	Absolute	$\frac{100}{N_v}\sum_{n_v=1}^{N_v} \widehat{y}^e_{n_v} = y_{n_v}$	0	100	y_{n_v} is the observed choice made by individual n in validation sample v, and $\widehat{y}^e_{n_v}$ is the choice with the highest predicted probability, predicted from model estimated on sample e.
% $_{\text{clearly right}}$ (t)	%CR	Absolute	$\frac{100}{N_v}\sum_{n_v=1}^{N_v} CR_{n_v}$ *where,* $CR_{n_v} = \begin{cases} 1 \text{ if } \widehat{P}(y^e_{n_v}) > t \\ 0 \quad \text{otherwise} \end{cases}$	0	100	$\widehat{P}(y^e_{n_v})$ is the estimated choice probability of the chosen alternative for individual n in validation sample v, predicted from model estimated on sample e. $\widehat{P}(!y^e_{n_v})$ is the estimated choice probability of an alternative other than the chosen one. t is an arbitrary threshold.
% $_{\text{clearly wrong}}$ (t)	%CW	Absolute	$\frac{100}{N_v}\sum_{n_v=1}^{N_v} CW_{n_v}$ *where,* $CW_{n_v} = \begin{cases} 1 \text{ if } \widehat{P}(!y^e_{n_v}) > t \\ 0 \quad \text{otherwise} \end{cases}$	0	100	
% $_{\text{unclear}}$ (t)	%U	Absolute	$100 - (\%_{\text{clearly right}}\,(t) + \%_{\text{clearly wrong}}\,(t))$	0	100	
Fitting factor	FF	Absolute	$\frac{1}{N_v}\sum_{n_v=1}^{N_v} \widehat{P}(y^e_{n_v})$	0	1	
Concordance statistic	C	Absolute	$\frac{1}{N_{c1}N_{c0}}\sum_{n_{c1}=1}^{N_{c1}}\sum_{n_{c0}=1}^{N_{c0}} C(\widehat{P}(y^e_{n_1,c1}), \widehat{P}(y^e_{n_0,c0}))$ *where,* $C(\widehat{P}(y^e_{n_1,c1}), \widehat{P}(y^e_{n_0,c0})) = \begin{cases} 1 \text{ if } \widehat{P}(y^e_{n_1,c1}) > \widehat{P}(y^e_{n_0,c0}) \\ 0.5 \text{ if } \widehat{P}(y^e_{n_1,c1}) = \widehat{P}(y^e_{n_0,c0}) \\ 0 \text{ if } \widehat{P}(y^e_{n_1,c1}) < \widehat{P}(y^e_{n_0,c0}) \end{cases}$	0	1	Given a binary choice situation: N_{c1} is the subset of the sample that chose alternative c. N_{c1} is the subset of the sample that did not choose alternative c. $\widehat{P}(y^e_{n_1,c1})$is the probability of choosing alternative c for individuals that chose it. $\widehat{P}(y^e_{n_0,c0})$ is the probability of choosing alternative c for individuals that did not choose it. The v subscript is omitted for simplicity.
Correlation	CORR	Absolute	$corr(s_v, \widehat{s}_v{}^e)$	−1	1	Correlation between predicted and observed outcomes. s_v is a continuous aggregate outcome measure in sample v (i.e. train ridership, etc.) $\widehat{s}^e_v$ is a continuous aggregate outcome measure predicted from model estimated on sample e.
Error	E	Relative	$\widehat{s}^e_{v,m} - s_{v,m}$	–	–	$s_{v,m}$ is an aggregate outcome measure in sample v, such as the market share of alternative m (i.e. modal market share), choice frequency, etc. $\widehat{s}^e_{v,m}$ is an aggregate outcome measure in sample v, such as the market share of alternative m, predicted from model estimated on sample e. M is the number of alternatives in the choice set.
Percentage error	PE	Absolute	$100 \cdot \frac{\widehat{s}^e_{v,m} - s_{v,m}}{s_{v,m}}$	-	-	
Absolute percentage error	APE	Absolute	$100 \cdot \left\lvert\frac{\widehat{s}^e_{v,m} - s_{v,m}}{s_{v,m}}\right\rvert$	0	-	
Mean absolute percentage error	MAPE	Absolute	$\frac{100}{M}\sum_{m=1}^{M}\left\lvert\frac{\widehat{s}^e_{v,m} - s_{v,m}}{s_{v,m}}\right\rvert$	0	-	
Sum of square error	SSE	Relative	$\sum_{m=1}^{M}(\widehat{s}^e_{v,m} - s_{v,m})^2$	0	-[b]	
Root sum of square error	RSSE	Relative	$\sqrt{\sum_{m=1}^{M}(\widehat{s}^e_{v,m} - s_{v,m})^2}$	0	-[b]	

(*continued on next page*)

Table 1 (*continued*)

Measure	Abbrv.	Type	Equation	Lower bound	Upper bound	Notes
Mean absolute error	MAE	Aggregate: Relative Disaggregate: Absolute	$\frac{1}{M}\sum_{m=1}^{M}\left\|\widehat{s}_{v,m}^{e} - s_{v,m}\right\|$	0	-[b, c]	
Mean squared error	MSE	Aggregate: Relative Disaggregate: Absolute	$\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{s}_{v,m}^{e} - s_{v,m}\right)^2$	0	-[b, c]	
Root mean square error	RMSE	Aggregate: Relative Disaggregate: Absolute	$\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{s}_{v,m}^{e} - s_{mv,}\right)^2}$	0	-[b, c]	
Brier Score	BS	Absolute	$\frac{1}{N_v}\sum_{n_v=1}^{N_v}\sum_{m=1}^{M}\left(\widehat{P}(y_{n_v,m}^{e}) - y_{n_v,m}\right)^2$	0	2[d]	$\widehat{P}(y_{n_v,m}^{e})$ is the predicted probability that individual n chooses alternative m, predicted from model estimated on sample e. y_{nm} is the actual outcome variable valued 0 or 1.
Log-likelihood	LL	Relative	$LL_v(\widehat{\boldsymbol{\beta}}^{e})$	–	0	$LL_{v,r}(\widehat{\boldsymbol{\beta}}^{e})$ is the log-likelihood of the model estimated on data e applied to the validation data v_r. $N_{v,r}$ is the size of the validation (holdout) sample r, and R is number of validation samples generated.
Log-likelihood loss	LLL[e]	Relative	$\frac{1}{\mathrm{R}}\sum_{r} - \frac{1}{N_{v,r}}\sum_{n_{v,r}} LL_{v,r}(\widehat{\boldsymbol{\beta}}^{e})$ $\forall\, 1 \le r \le R$	0	–	
Rho-square (Likelihood ratio index)	RHOSQ	Absolute	$\rho^2 = 1 - \frac{LL_v(\widehat{\boldsymbol{\beta}}^{e})}{LL_v(0)}$	0	1	$LL_v(\widehat{\boldsymbol{\beta}}^{e})$ is the log-likelihood of the model estimated on data e applied to the validation data v. $LL_v(0)$ is log-likelihood of the model when all parameters are zero for data v.
Transfer rho-square (Likelihood ratio index)	T-RHOSQ	Absolute	$\rho_{transfer}^2 = 1 - \frac{LL_v(\widehat{\boldsymbol{\beta}}^{e})}{LL_v(\boldsymbol{MS}^{v})}$	–	ρ_{local}^2	$LL_v(\boldsymbol{MS}^{v})$ is a base model estimated on validation data v (i.e. market share model.) ρ_{local}^2 is the local rho-square of the model.
Transfer index	TI	Relative	$\frac{LL_v(\widehat{\boldsymbol{\beta}}^{e}) - LL_v(\boldsymbol{MS}^{v})}{LL_v(\widehat{\boldsymbol{\beta}}^{v}) - LL_v(\boldsymbol{MS}^{v})}$	–	1	$LL_v(\widehat{\boldsymbol{\beta}}^{v})$ is the likelihood of the model estimated on the validation data v.
Transferability test statistic	TTS	Pass/Fail	$-2(LL_v(\widehat{\boldsymbol{\beta}}^{v}) - LL_v(\widehat{\boldsymbol{\beta}}^{e}))$	0	–	
χ^2 test	CHISQ	Pass/Fail	$\sum_{m=1}^{M}\frac{(f_m - E(f_{v,m}^{e}))^2}{E(f_{v,m}^{e})}$	0	–	f_m is the observed choice frequency of alternative m in sample v, and $E(f_{v,m}^{e})$ is the expected choice frequency predicted from model estimated on sample e.

[a] Bounded for market shares.
[b] In the case aggregate outcomes are market shares upper bounds are dependent on the choice set size.
[c] Upper bounds exist for the disaggregate case.
[d] For the specific case of binary choices, it is common to drop the second summation sign for simplicity. In this case the upper bound is 1.
[e] A simple average over R, or a moving average across all validation sets can be calculated.

the validation and reporting practices in the peer-reviewed literature and show that the transportation academic literature has over-relied on statistical goodness-of-fit and disregarded model validation to a very large extent.

Using the Web of Science Core Collection maintained by Clarivate Analytics we reviewed validation and reporting practices in the transportation academic literature published between 2014 and 2018. Articles were selected based on the following criteria:

(1) Peer-reviewed journal articles published between 2014 and 2018
(2) Analysis uses discrete choice models

(3) Target choice dimensions are
  ➢ Destination choice
  ➢ Mode choice
  ➢ Route choice
(4) Articles that analyse other choice dimensions are considered if and only if the article includes at least one of the three target choice dimensions defined in (3).
(5) Web of Science Database search keywords are:
  ➢ Destination choice model
  ➢ Mode choice model
  ➢ Route choice model
(6) Web of Science Database fields are:
  ➢ Transportation
  ➢ Transportation science and technology
  ➢ Economics
  ➢ Civil engineering
(7) Research scope is limited to human land transport and daily travel behaviour (tourism, evacuation behaviour, and freight transport articles are excluded)
(8) Analysis uses empirical data from revealed preference (RP) or joint revealed/stated preferences (SP-RP) studies.[3] (Studies using numerical simulations only are excluded)
(9) Methodological papers only included if they use empirical data

A total of 226 articles met the above criteria. Note that although the choice dimensions are destination, mode and route choice, the definition of choice settings or choice sets differ by study. For example, in the category of mode choice, in addition to the traditional way to define choices (i.e. car, rail, bus, walk, etc.) changes in mode choices are also included in the review. Similarly, in route choice analysis, in addition to link and path choices, simpler choice settings are also included such as, riding on a bike lane, or on the walkway, taking the stairs or the escalator, using a particular detour or not, etc. In the case of route choice models, Stochastic user equilibrium (SUE) models are excluded, as discrete choice models are just a subcomponent of a larger model.

Of the 226 articles reviewed, and consistent with standard practice in the field, 92% of articles reported a goodness of fit statistic. 64.6% of articles reported some kind of policy-relevant inference analysis. This means going beyond simply discussing the statistical significance and direction of the estimated parameters, and focusing instead on effect magnitudes, and estimating values that are interpretable in a policy context such as marginal effects, elasticities, odds ratios, marginal rates of substitution, and/or policy scenario simulations (See Table 5 in the appendix section for the complete table summarizing the articles reviewed in this paper) This is a welcomed finding amid criticism that focusing exclusively on statistical significance is widespread in many sciences at the expense of policy-relevant inferences, discussion of effect magnitudes, and tests of power (Ziliak and McCloskey, 2007).

Only 18.1% of the reviewed articles reported model validation, out of which 78% (14.2% of all studies) consisted of internal validation and 22% (4% of all studies) consisted of external validation. Table 2 summarizes the details of these studies. Note that only studies that reported explicitly and in a clear way how model validation was conducted were considered.

In terms of internal validation, as illustrated in Table 3, of the studies that reported any validation practice, 56.3% used the holdout validation method, followed by repeated learning-testing (25%). These two methods add up to 81.3% of the studies reporting internal validation in the literature. 12.5% of studies used an independent sample for internal validation, with the remaining 87.5% relying on sample splitting approaches.

The bootstrap method was only used by one study (Sanko, 2017), which used bootstrapping to evaluate the effects of data newness and sample size on the temporal transferability of demand forecasting models.

Irrespective of type of validation, as shown in Table 4, the log-likelihood (34.1%) and the log-likelihood loss (12.2%) had jointly the highest reporting share, with 46.3% of studies reporting either one of them.[4] The second most frequently used measures were the predicted-vs-observed simple comparison and the percentage of correct predictions with a 24.4% share each.

One limitation of likelihood-based measures such as LL and LLL is that they fail to provide an absolute measure of predictive accuracy. This is important because gains in predictive power of a “superior” model can be, in fact, very minimal. That is, the best model among a set of models can still be a very bad model prediction-wise. A similar argument can be made with relation to other relative performance measures. While we believe that model validation is not a pass/fail test, absolute measures of predictive ability such as the percentage of correct predictions, rho-square (7.3%), or the Brier Score (2.4%), among others, are useful as the can be used to generate benchmark values against which researchers can evaluate, to a certain extent, the performance of their models against similar studies in the literature. For example, in the reviewed articles, the percentage of correct predictions for destination (or location) choice models ranged between 13% and 22%, while for mode choices it ranged from 36% to 87%, with most studies reporting values above the 60% threshold. Similarly, for route choice, values ranged between 51% and 73%. However, note that since the number of

[3] Give that the error component variance for designed experiments will differ from the variance in the real world, the use of stated preference (SP) surveys for forecasting and calculation of elasticities is not recommended (Hess and Rose, 2009; de Jong, 2014). As such SP studies were excluded from the analysis. In the case of SP-RP studies, the SP error component can be calibrated against the RP data (de Jong, 2014).

[4] Note that as shown in Table 2, some articles reported more than one measure.

Table 2
Summary of articles reporting validation in the literature.

Article	Model	Dependent variables	Validation			Evaluation Measure																					
			Type	Method	Notes	PVO	FPR	% CR	% CW	E/ PE/ APE	SSE	MAE	MAPE	MSE	RMSE	BS	C	CHISQ	MSD	FF	LL	LLL	RHOSQ	CORR	TTS	TI	OTHER
Zimmermann et al. (2018)	MNL; MXL; MRL	AC-DC-DT-MC	Internal	RLT	54% estimation, 46% validation, 13 runs.																	●					
Danalet et al. (2016)	MNL;	DC	Internal	HOV	91.5% estimation, 9.5% validation (past choices used for estimation, most recent choice used for validation).	●					●																
Faghih-Imani and Eluru (2015)	MNL	DC	Internal	HOV	75% estimation, 25% validation.		●													●	●						
Assi et al. (2018)	MNL; NN*	MC	Internal	RLT	75% estimation, 25% validation, 3 runs.		●																				
Bohluli et al. (2014)	MNL; NL	MC	External	IS	Validation data is observed ridership data after introduction of new transit service.	●				●					●									●			
Bueno et al. (2017)	MNL	MC	Internal	R-K-CV	10-fold CV, 5 000 runs.		●																				
Glerum et al. (2014)	HCM	MC	Internal	HOV	80% estimation, 20% validation.			●													●		●				
Hasnine and Habib (2018)	HDDC	MC	Internal	HOV	80% estimation, 20% validation.	●																					
Kunhikrishnan and Srinivasan (2018)	MNL	MC	Internal	HOV	70% estimation, 30% validation.																●		●				
Ma et al. (2015)	MNL; MXL; LCM	MC	Internal	HOV	80% estimation, 20% validation.		●																				
Mahmoud et al. (2016)	MNL; PLC	MC	Internal	HOV	80% estimation, 20% validation.		●																				
Sanko (2014)	MNL	MC	External	IS	Validation against temporal transfer sample.							●									●						
Sanko (2016)	MNL	MC	External	IS																	●						

(continued on next page)

Table 2 (*continued*)

Article	Model	Dependent variables	Validation			Evaluation Measure																					
			Type	Method	Notes	PVO	FPR	% CR	% CW	E/ PE/ APE	SSE	MAE	MAPE	MSE	RMSE	BS	C	CHISQ	MSD	FF	LL	LLL	RHOSQ	CORR	TTS	TI	OTHER
					Validation against temporal transfer sample.																						
Sanko (2017)	MNL	MC	External	IS	Validation against temporal transfer sample.																●						
Sanko (2018)	MNL	MC	External	IS	Validation against temporal transfer sample.																●						
Vij and Krueger (2017)	MNL; MXL; LCM	MC	Internal	HOV	90% estimation, 10% validation.																●						
Vij and Walker (2014)	LCM	MC	Internal	HOV	90% estimation, 10% validation.																●						
Wang et al. (2014)	NL	MC	Internal	IS	Validation data is observed mode shares collected by transit agencies.	●																					
Weng et al. (2018)	MNL	MC	Internal	IS	Validation data is smart cart data.							●															
Gokasar and Gunay (2017)	MNL	MC	Internal	HOV	75% estimation, 25% validation.	●	●																				
Habib et al. (2014)	hTEV	MC; △MC	External	IS	Validation against temporal transfer sample.	●												●									●
Chikaraishi and Nakayama (2016)	MNL; WM; QL	MC; RC	Internal	RLT	50% estimation, 50% validation, 100 runs.		●														●						
Idris et al. (2015)	MNL; NL; HCM	MC; △MC	Internal	HOV	Validation group is subset of observations with an observed mode shift (in SP experiment).	●				●																	●
Golshani et al. (2018)	MNL; ANN*; JDC	MC-DT	Internal	HOV	80% estimation, 20% validation.		●						●														
Suel and Polak (2017)	NL	O -MC	Internal	HOV	Validation sample is observations from same sample, for the one-month										●												

(*continued on next page*)

Table 2 (*continued*)

Article	Model	Dependent variables	Validation Type	Validation Method	Validation Notes	PVO	FPR	% CR	% CW	E/ PE/ APE	SSE	MAE	MAPE	MSE	RMSE	BS	C	CHISQ	MSD	FF	LL	LLL	RHOSQ	CORR	TTS	TI	OTHER
					period prior to survey week.																						
Faghih-Imani and Eluru (2017)	LMM; MNL	DC; O	Internal	HOV	For AT/DT: Approx. 96% estimation, 4% validation (1week/ 24weeks). For DC: holdout of 5 000 trips (estimation was conducted with sample sizes ranging from 1 000 to 20 000).		●					●			●						●						
Alizadeh et al. (2018)	PSL; EPSL; IAL; LK; NL	RC	Internal	HOV	80% estimation, 20% validation, 1 run.		●																				
Kazagli et al. (2016)	MNL	RC	Internal	RLT	80% estimation, 20% validation, 100 runs.																						●
Lai and Bierlaire (2015)	MNL; PSL; CNL	RC	Internal	HOV	50% estimation, 50% validation. 3 runs, performance measures not averaged.																●						
Li et al. (2016)	MXL (PSL; GNL; LK; CL, etc)	RC	Internal	SS-O	Two non-overlapping samples were generated from original dataset for validation.																●						
Mai (2016)	NRL; RCNL	RC	Internal	RLT	80% estimation, 20% validation, 20 runs.																	●					
Mai et al. (2017)	RL; NRL; ML; RRM	RC	Internal	RLT	80% estimation, 20% validation, 40 runs.																	●					
Mai et al. (2015)	RL; NRL	RC	Internal	RLT	80% estimation, 20% validation, 40 runs.																	●					
Papola (2016)	CoRUM	MC	Internal	HOV	70% estimation, 30% validation.																●		●				

(*continued on next page*)

Table 2 (*continued*)

Article	Model	Dependent variables	Validation			Evaluation Measure																					
			Type	Method	Notes	PVO	FPR	% CR	% CW	E/ PE/ APE	SSE	MAE	MAPE	MSE	RMSE	BS	C	CHISQ	MSD	FF	LL	LLL	RHOSQ	CORR	TTS	TI	OTHER
Zhang et al. (2015)	MNL	RC	Internal	IS	Validation data is an independent observation of 3 sites plus a new site.	●																					
Zhang et al. (2018)	NL	RC	Internal	HOV	85% estimation, 15% validation.	●						●															
Zimmermann et al. (2017)	NRL	RC	Internal	RLT	80% estimation, 20% validation, 20 runs.																	●					
Xie et al. (2016)	MXL	RC	Internal	IS	New observations from same site.							●															
Duc-Nghiem et al. (2018)	MNL	RC	External	IS	Validation data is pre-post data from site not used in calibration.											●	●										
Fox et al. (2014)	NL	MC-DC	External	IS	Validation against temporal transfer sample.	●				●					●											●	
Forsey et al. (2014)	O	MC	External	IS	Validation against temporal transfer sample.							●													●	●	

Model abbreviations:

CNL: Cross-nested logit
CoRUM: Combination of RUM models
EPSL: Extended path size logit
HCM: Hybrid choice model
hTEV: Heteroscedastic tee extreme value
IAL: Independent availability logit
IBL: Instance-based learning
LCM: Latent class model
LK: Logit kernel
MNL: Binary or multinomial logit
MNLp: Binary or multinomial logit with panel data corrections
MNP: Binary or multinomial probit
MRL: Mixed recursive logit
MXL: Mixed logit
NL: nested logit
NN*: Neural network
O: Other
PLC: Parameterized logit captivity
PSL: Path size logit
QL: q-logit
RCNL; Recursive cross nested logit
RRM: Random-regret maximization
WB: Weibit model
*Machine learning models

Dependent variable abbreviations:

AC: Activity choice
AT: Arrival time
DT: Departure time choice/Time of day choice
DC: Destination choice
MC: Mode choice
△MC: Mode choice change
RC: Route choice
O: Other choice dimension

Validation method abbreviations

HOV: Holdout validation
IS: Validation against an independent sample
R-K-CV: Repeated K-fold cross-validation
RLT: Repeated learning-testing
SS-O: Other sample splitting method

Table 3
Internal validation methods reported in the literature by frequency.

Method	Abbvr.	Frequency	Percentage
Holdout validation	HOV	18	56.3%
Repeated learning-testing	RLT	8	25.0%
Validation against an independent sample	IS	4	12.5%
Repeated K-fold cross-validation	R-K-CV	1	3.1%
Other sample splitting methods	SS-O	1	3.1%

studies is not large, a myriad of studies with different definition of choice variables and choice set sizes were combined to calculate these values. As such, they need to be interpreted with caution. Furthermore, as pointed out by de Luca and Cantarella (2009), this index fails to discriminate between different degrees of predicted probabilities of correct predictions. Unfortunately, "clearness of prediction" measures that do account for this, are still not very widely used. The percentage of clearly right index had a share of 2.4%, while the percentage of clearly wrong index had a 0% share. Similarly, the share of studies reporting measures of model discriminatory ability was 2.4%.

Due to the fact that model performance is relative to factors such as choice set size and the base market share of alternatives, comparisons across models in different studies and contexts are complicated. Such comparisons must be interpreted with clear understanding of these limitations, and not as absolute measures of model superiority.

Although ideally all studies would be externally validated, in most cases the types of validation that a researcher can conduct are limited by the availability of data and computational resources. But there is certainly room for improvement. Researchers should strive to conduct the best validation tests possible given the resources at hand and carefully report the details of how these tests were conducted, so that other researchers can clearly understand to what extent the results presented are generalizable. In Fig. 2 we illustrate as simple heuristic to determine the recommended validation practices given available resources, as well as what measures to report. We start by acknowledging that if a randomized controlled trial is possible, it would be the best alternative. That being said, as we discussed earlier, such an experiment is extremely difficult in the field.

The existence of data from either a different but plausibly related population, or from the same population is one of the key factors defining what kind of validation is possible. The bottom line is conducting internal validation either via data/splitting methods or bootstrapping, in the absence of independent datasets. Unless the model is too computationally intensive, we recommend avoiding the holdout validation as it a pessimistic estimator that makes inefficient use of the data (Kohavi, 1995). Regarding the choice between cross-validation and bootstrapping, as discussed earlier, the performance of these methods is dependent on data characteristics such as sample size, and the type of model used. In the absence of an empirical study comparing the performance of these methods using typical data from transportation studies and commonly used models, we will refrain from making a recommendation.

In terms of what performance measures to report, rather than relying on a single indicator we recommend reporting several indicators as shown in Fig. 2, which include measures comparable across studies, and measures of discriminative ability and clearness of prediction. When using indicators for model selection, ideally, the best model will excel in all performance measures, but there is no guarantee this is so. As such, the analyst should strive to clearly explain the criteria used for model selection.

Table 4
Predictive accuracy performance measures reported in the literature by frequency.

Performance measure	Abbrv.	Frequency	Percentage
Log-likelihood/log-likelihood loss	LL/LLL	19	46.3%
Percentage of correct predictions or First Preference Recovery	FPR	10	24.4%
Predicted vs observed market outcomes	PVO	10	24.4%
Mean absolute error	MAE	6	14.6%
Root mean square error	RMSE	4	9.8%
Error/Percentage error/Absolute percentage error	E/PE/APE	3	7.3%
Rho-Square	RHOSQ	3	7.3%
Transfer index	TI	2	4.9%
% clearly right (t)	%CR	1	2.4%
Brier Score	BS	1	2.4%
Chi-square	CHISQ	1	2.4%
Concordance index	C	1	2.4%
Correlation	CORR	1	2.4%
Fitting factor	FF	1	2.4%
Mean absolute percentage error	MAPE	1	2.4%
Sum of square error	SSE	1	2.4%
Transferability test statistic	TTS	1	2.4%
All other measures specified in Table 1	–	0	0%
Other measures not specified in Table 1	–	3	7.3%

Very similar measures are reported jointly.

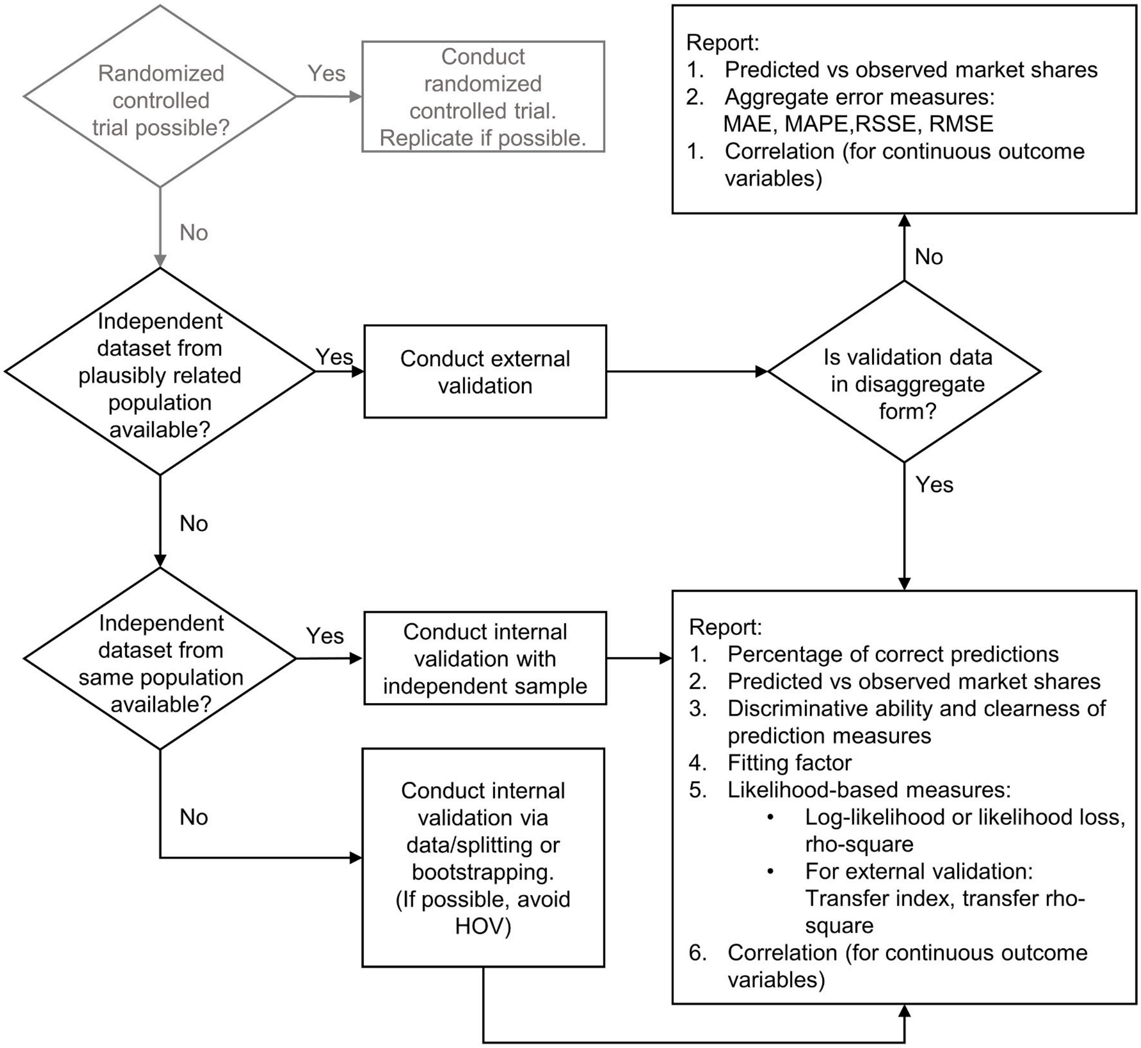

Fig. 2. Heuristic to select validation method given available resources and recommended performance measures to report.

7. Towards better validation practices in the field

While most researchers would agree that the purpose of travel demand analysis is to make valid predictions to aid effective policy evaluation (evident in that most studies reported some sort of policy implication,) we showed that there is an evident disconnect between the objective of transportation research and the current practice of research, partially evident in the low levels of validation reporting in the literature. Strong pressure to publish among academics means little incentive to conduct proper validation. While producing independent samples for validation might be prohibitively expensive, whenever possible researchers should strive to produce such data or use existing periodical surveys such as household travel surveys. In the absence of such data, data-splitting and bootstrapping methods are inexpensive and should not be very different from drawing policy-relevant inferences, in terms of efforts required.

Stronger criteria are needed to evaluate models in academia to increase the reliability of results and the credibility of inferences based on statistical models. We put forth a set of recommendations aimed at improving validation practices in the field:

(1) *Make validation mandatory:* model validation should be a non-negotiable part of model reporting and peer-review in academic journals for any study that provides policy recommendations. At the very least, internal validation results should be reported. Conducting internal validation is a norm in machine learning studies, and there is no reason why similar standards cannot be implemented in the field. This will provide better incentives to perform validation.
(2) *Sharing of benchmark datasets:* as pointed out by an anonymous reviewer at the conference the previous version of this article was presented, a fundamental limitation in the field is the lack of benchmark datasets and a general culture of sharing code and data. Certainly, privacy concerns as well as institutional limitations impede the free sharing of most relevant and larger datasets (i.e.

household travel surveys, etc.), and in most cases it is out of the control of individual researchers to decide, but efforts should be made towards the collection and opening of relevant benchmark data.

(3) *Incentivize validation studies:* most prestigious journals put a lot of emphasis on theoretically innovative models. While we recognize its importance, submissions that focus on proper validation of existing models and/or theories should be encouraged by journal editors.

(4) *Draw and enforce clear reporting guidelines:* efforts to improve reporting are well documented in other fields. For example, in the medical field, following a consensus among researchers that reporting of published observational studies is inadequate, and hinders the assessment of quality and generalizability of these studies, guidelines have been developed to strengthen the reporting of observational studies (e.g. STROBE statement (von Elm et al., 2007)). Although results are mixed regarding the impact of these guidelines (as in many journals the use of the guideline is recommended but not mandatory), it is a step in the right direction. A transportation field specific guideline could be developed and properly enforced, where, in addition to detailed information of survey characteristics such as sampling method and representativeness of the data, validation reporting is required. It is worth noting that guidelines for travel model validation do exist for practitioners (i.e. Cambridge Systematics, 2010), so validation and reporting guidelines specific to transportation researchers are a welcomed contribution. Proper reporting should also include policy-relevant values such as elasticities, marginal effects and other inferences that properly convey the magnitude of the effects of interest. As mentioned earlier, 64.6% of the studies reviewed reported some sort of policy-relevant inferences, a welcomed finding. Models with high-predictive power are not very useful if no policy implications can be drawn from it.

Finally, there is one argument and two question that we would like to address that are commonly heard in academic circles related to validation practices.

The argument is that "*I'm not validating my model because I'm not trying to build a predictive framework. I'm trying to learn about travel behaviour*". In response, we argue that the more exploratory the subject is, the less the onus of validation until some critical mass of research has been conducted. On the other hand, the more orthodox the type of analysis conducted (such as the dimensions of travel behaviour covered in this study), the stronger the onus of validation.

This response motivates the following questions: "*Should every study using a discrete choice model be conducting validation?*" and *"Is what we learn about travel behaviour from coefficient estimation less valuable if validation not conducted?"*

Regarding the first question, in short, yes. At the very least, any article that makes policy recommendations should be subject to proper validation given the issues discussed in Section 2 about policy impacts and the lack of a feedback loop in academia. There is a myriad of reasons why some scepticism is warranted against any particular model outcome, the most obvious one being model overfitting. So proper validation should be used to strengthen presented results, especially given the limitations discussed in Section 2 regarding the dependence on cross-sectional studies, and difficulties associated with scientific hypothesis testing.

Regarding the second question, while coefficients are useful for policy-relevant inferences, coefficients by themselves do not inform us about model predictive ability. Policy inference analysis should be complemented with evidence on the generalizability of such inference.

Finally, while we recognize that better validation practices will not solve the credibility crisis in the field, it is certainly a step in the right direction. More specifically, model validation is no solution to the causality problem in the field (see Brathwaite and Walker (2018) for an in-depth discussion of causality in transportation studies), but we want to underscore that the reliance on observational studies inherent to the field demands more stringent controls to improve the validity of results. Although out of the scope of the present study, such controls should also include aspects such as sample representativeness, proper model specification, and statistical power of effects of interest, which are also critical to the validity of results.

8. Conclusion

In this article we reviewed validation practices from the transportation field in the peer-reviewed literature published between 2004 and 2008. We found that although 92% of studies reported goodness of fit statistics, and 64.6% reported some sort of policy-relevant inference analysis, the percentage of validation reporting stood at 18.1%.

We argued that model validation should be a non-negotiable part of model reporting and peer-review in academic journals and proposed a simple heuristic to choose validation methods given available resources. At the same time, we recognize the need for proper incentives to promote better validation practices and providing tools and knowledge to do so, such as reporting guidelines, and encouragement by journals of submissions that focus on validation of existing models and theories, and not only new theoretically innovative models.

Author statement

Giancarlos Parady conceived the study, conducted the literature review and wrote the first draft.

Giancarlos Parady, David Ory and Joan Walker reviewed and revised the manuscript, confirmed the analysis results, and approved the final version.

Funding

This work was supported by JSPS KAKENHI Grants Number 17K14737 and 20H02266.

Declaration of competing interest

The authors report no conflict of interest.

Acknowledgements

An earlier version of this work was presented at the 6th International choice modelling conference, Kobe, Japan, August 19–21, 2019

Appendix

Table 5
Summary of reviewed articles

No	Article	Model(s)	Dependent variable(s)	Data characteristics					Goodness of fit reported	Policy inference reported	Validation reported
				Data structure				Data type			
				CS	RCS	PP	TP	RP, RP-SP			
1	Manoj and Verma (2015b)	MNL	AC; MC; DT	●				RP	●	●	
2	Sadhu and Tiwari (2016)	NL	AC-DC	●				RP	●	●	
3	Arman et al. (2018)	MXNL	AC-MC	●				RP	●		
4	He and Giuliano (2018)	MNL	DC	●				RP	●	●	
5	Mahpour et al. (2018)	HCM	DC	●				RP	●		
6	Basu et al. (2018)	NL	DC	●				RP	●		
7	Clifton et al. (2016)	MNL	DC	●				RP	●	●	
8	Danalet et al. (2016)	MNLp;	DC				●	RP	●	●	●
9	Deutsch-Burgner et al. (2014)	MNL	DC	●				RP	●		
10	Faghih-Imani and Eluru (2015)	MNL	DC				●	RP	●	●	●
11	Ho and Hensher (2016)	MNL	DC	●				RP	●		
12	Huang and Levinson (2015)	MXL	DC				●	RP	●	●	
13	Huang and Levinson (2017)	MXL	DC				●	RP	●	●	
14	Shao et al. (2015)	MNL	DC	●				RP	●	●	
15	Wang et al. (2017)	SCL; MNL	DC				●	RP	●		
16	Faghih-Imani and Eluru (2017)	LMM; MNL	DC; AT; DT				●	RP	●	●	●
17	Khan et al. (2014)	MNL	DC; MC	●				RP		●	
18	Paleti et al. (2017)	MXL; JDC	DC-DT-O	●				RP	●	●	
19	González et al. (2016)	NL(PSL)	DC-RC	●				RP	●	●	
20	Ma et al. (2018)	CNL	DT-MC	●				RP		●	
21	Abasahl et al. (2018)	NL	MC	●				RP	●		
22	Anta et al. (2016)	MNL; NL	MC	●				RP-SP	●	●	
23	Assi et al. (2018)	MNL	MC	●				RP	●		●
24	Aziz et al. (2018)	MXL	MC	●				RP	●	●	
25	Böcker et al. (2017)	MNL	MC				●	RP	●		
26	Bohluli et al. (2014)	O	MC	●				RP	●	●	●
27	Braun et al. (2016)	MNL	MC	●				RP-SP	●		
28	Bridgelall (2014)	MNL	MC	●				RP		●	
29	Bueno et al. (2017)	MNL	MC	●				RP	●	●	●
30	Castillo-Manzano et al. (2015)	MNL	MC	●				RP	●	●	
31	Chakour and Eluru (2014)	LCM	MC	●				RP	●	●	
32	Cherchi and Cirillo (2014)	MXL	MC				●	RP	●	●	
33	Cherchi et al. (2017)	MXL	MC				●	RP	●		
34	Chica-Olmo et al. (2018)	MNL	MC	●				RP	●	●	
35	Clark et al. (2014)	MXL	MC				●	RP	●	●	
36	Cole-Hunter et al. (2015)	MNL	MC	●				RP	●	●	
37	Collins and MacFarlane (2018)	MNL	MC				●	RP	●	●	

(continued on next page)

Table 5 (*continued*)

No	Article	Model(s)	Dependent variable(s)	Data characteristics: Data structure CS	RCS	PP	TP	Data type RP, RP-SP	Goodness of fit reported	Policy inference reported	Validation reported
38	Danaf et al. (2014)	NL	MC	●				RP	●	●	
39	de Grange et al. (2015)	O	MC	●				RP	●	●	
40	Di Ciommo et al. (2014)	MXL	MC				●	RP	●		
41	Ding et al. (2017)	HCM	MC	●				RP	●		
42	Ding et al. (2018)	HCM	MC-MTO	●				RP	●		
43	Ding et al. (2014b)	O	MC	●				RP	●		
44	Dong et al. (2016)	NL	MC	●				RP	●		
45	Nguyen-Phuoc (2018)	MNL	MC	●				RP	●	●	
46	Nguyen-Phuoc (2018)	MNL	MC	●				RP	●	●	
47	Efthymiou and Antoniou (2017)	HCM	MC		●			RP			
48	Ermagun and Levinson (2017)	CNL	MC	●				RP	●	●	
49	Ermagun and Samimi (2015)	NL	MC	●				RP	●	●	
50	Ermagun et al. (2015)	NL; O	MC	●				RP	●	●	
51	Fernández-Antolín et al. (2016)	O	MC	●				RP	●	●	
52	Flügel et al. (2015)	CNL	MC		●			RP-SP	●		
53	Forsey et al. (2014)	O	MC		●			RP	●	●	●
54	Fu and Juan (2017)	LCM	MC	●				RP	●		
55	Gao et al. (2016)	MNL	MC	●				RP	●		
56	Gao et al. (2017)	MNL	MC	●				RP	●	●	
57	Gerber et al. (2017)	MNL; MXL	MC			●		RP	●		
58	Glerum et al. (2014)	HCM	MC	●				RP	●	●	●
59	Goel and Tiwari (2016)	MNL	MC	●				RP	●		
60	Guan and Xu (2018)	MNLp	MC	●				RP	●	●	
61	Guerra et al. (2018)	MNL	MC	●				RP	●	●	
62	Guo et al. (2018)	MXL	MC	●				RP	●		
63	Habib and Sasic (2014)	O	MC	●				RP	●	●	
64	Habib and Weiss (2014)	O	MC		●			RP	●	●	
65	Habib et al. (2014)	HCM	MC	●				RP	●		
66	Halldórsdóttir et al. (2017)	JMXL	MC	●				RP	●	●	
67	Hasnine and Habib (2018)	O	MC	●				RP	●	●	●
68	Hasnine et al. (2018)	CNL	MC	●				RP	●	●	
69	He and Giuliano (2017)	MNL	MC	●				RP	●	●	
70	Helbich (2017)	MXL	MC	●				RP	●		
71	Helbich et al. (2014)	O	MC				●	RP	●		
72	Hensher and Ho (2016)	MXL	MC	●				RP	●	●	
73	Hsu and Saphores (2014)	MNL	MC	●				RP	●	●	
74	Hurtubia et al. (2014)	LCM	MC	●				RP	●	●	
75	Irfan et al. (2018)	MNL	MC	●				RP-SP	●	●	
76	Lin et al. (2018)	LCM	MC	●				RP	●	●	
77	Ji et al. (2017)	NL	MC	●				RP	●	●	
78	Kamargianni et al. (2014)	HCM	MC	●				RP			
79	Kamruzzaman et al. (2015)	MNL	MC	●				RP	●	●	
80	Keyes and Crawford-Brown (2018)	MNL	MC	●				RP	●	●	
81	Khan et al. (2016)	MXL	MC	●				RP	●	●	
82	Kunhikrishnan and Srinivasan (2018)	MNL	MC	●				RP	●	●	●
83	Yang et al. (2017)	MNL	MC	●				RP	●		
84	Larsen et al. (2018)	MNL	MC	●				RP	●	●	
85	Lee (2015)	MNL	MC	●				RP	●		
86	Lee et al. (2017)	MNL	MC	●				RP	●	●	
87	Lee et al. (2014)	MNL	MC	●				RP	●	●	
88	Li and Kamargianni (2018)	MXNL	MC	●				RP-SP	●	●	
89	Liu et al. (2018)	MNL	MC	●				RP	●	●	
90	Liu et al. (2015)	MNL	MC		●			RP	●	●	
91	Lorenzo Varela et al. (2018)	ML; NL	MC	●				RP	●	●	

(*continued on next page*)

Table 5 (*continued*)

No	Article	Model(s)	Dependent variable(s)	Data characteristics: Data structure: CS	RCS	PP	TP	Data type: RP, RP-SP	Goodness of fit reported	Policy inference reported	Validation reported
92	Ma et al. (2015)	LCM	MC	●				RP	●	●	●
93	Mahmoud et al. (2016)	O	MC	●				RP	●	●	●
94	Mattisson et al. (2018)	MNL	MC	●				RP	●		
95	Mehdizadeh et al. (2018)	MXL	MC	●				RP	●		
96	Meng et al. (2016)	MNL	MC	●				RP	●	●	
97	Minal and Ravi (2016)	MXL	MC	●				RP	●	●	
98	Mitra and Buliung (2014)	MNL	MC	●				RP	●		
99	Mitra and Buliung (2015)	MNL	MC	●				RP	●	●	
100	Mitra et al. (2015)	MNL	MC	●	●			RP	●	●	
101	Moniruzzaman and Farber (2018)	MNL	MC	●				RP	●	●	
102	Myung-Jin et al. (2018)	MNL	MC	●				RP	●	●	
103	Noland et al. (2014)	MXL	MC	●				RP	●		
104	Owen and Levinson (2015)	MNL	MC	●				RP	●		
105	Park et al. (2015)	MNL	MC	●				RP			
106	Park et al. (2014)	MNL	MC	●				RP			
107	Paulssen et al. (2014)	HCM	MC	●				RP		●	
108	Pike and Lubell (2018)	O	MC	●				RP	●	●	
109	Pnevmatikou et al. (2015)	NL	MC	●				RP-SP	●	●	
110	Prato et al. (2017)	LCM	MC	●				RP	●	●	
111	Ramezani, Pizzo and Deakin (2018b)	MNL	MC	●				RP	●		
112	Ramezani, Pizzo and Deakin (2018a)	MNL	MC	●				RP	●	●	
113	Rashedi et al. (2017)	O	MC	●				RP-SP	●	●	
114	Rubin et al. (2014)	MNLp	MC				●	RP	●		
115	Rybarczyk and Gallagher (2014)	MNL	MC	●				RP		●	
116	Sanko (2014)	MNL	MC		●			RP	●		●
117	Sanko (2016)	MNL	MC		●			RP	●		●
118	Sanko (2017)	MNL	MC		●			RP	●		●
119	Sanko (2018)	MNL	MC		●			RP	●		●
120	Sarkar and Chunchu (2016)	MNL	MC	●				RP	●	●	
121	Sarkar and Mallikarjuna (2018)	HCM	MC	●				RP	●	●	
122	Scheepers et al. (2016)	MNL	MC	●				RP		●	
123	Shaheen et al. (2016)	MNL	MC	●				RP	●		
124	Sharmeen and Timmermans (2014)	MNL	MC	●				RP	●		
125	Shirgaokar and Nurul Habib (2018)	HCM	MC	●				RP	●	●	
126	Singh and Vasudevan (2018)	MNL	MC	●				RP	●	●	
127	Soltani and Shams (2017)	NL	MC	●				RP	●		
128	Stone et al. (2014)	MNL	MC	●				RP	●	●	
129	Sun et al. (2018)	MNL	MC	●				RP	●	●	
130	Thigpen et al. (2015)	O	MC	●				RP	●	●	
131	Toşa et al. (2018)	NL	MC	●				RP-SP	●	●	
132	Venigalla and Faghri (2015)	MNL	MC	●				RP	●	●	
133	Verma et al. (2015)	MNL	MC	●				RP	●	●	
134	Vij and Krueger (2017)	MXL	MC	●				RP	●	●	●
135	Vij and Walker (2014)	LCM	MC	●				RP	●	●	●
136	Vij et al. (2017)	LCM	MC		●			RP	●	●	
137	Wang et al. (2014)	NL	MC	●				RP	●	●	●
138	Wang et al. (2015)	MNP; O	MC	●				RP	●	●	
139	Weiss and Habib (2018)	O	MC	●				RP	●	●	
140	Weng et al. (2018)	MNL; O	MC	●				RP-SP	●		●
141	Yang et al. (2014)	MNL	MC	●				RP	●	●	
142	Yang et al. (2018)	MXL	MC	●				RP	●	●	
143	Yang et al. (2016)	NL	MC	●				RP	●	●	
144	Yazdanpanah and Hadji Hosseinlou (2017)	HCM	MC	●				RP	●	●	

(*continued on next page*)

Table 5 (*continued*)

No	Article	Model(s)	Dependent variable(s)	Data characteristics					Goodness of fit reported	Policy inference reported	Validation reported
				Data structure				Data type			
				CS	RCS	PP	TP	RP, RP-SP			
145	Yen et al., 2018a	LCM	MC				●	RP	●	●	
146	Yen et al., 2018b	MNL	MC				●	RP	●	●	
147	Zhang et al. (2017)	MNL	MC	●				RP	●		
148	Zhao and Li (2017)	ML-MNL	MC	●				RP		●	
149	Zimmermann et al. (2018)	MRL	MC	●				RP	●	●	●
150	Zolnik (2015)	ML-MNL	MC		●			RP		●	
151	Gokasar and Gunay (2017)	MNL	MC	●				RP	●	●	●
152	Tilahun et al. (2016)	MNL	MC	●				RP	●	●	
153	Astegiano et al. (2017)	MXL; CNL	MC; MTO				●	RP	●		
154	Heinen (2016)	MNL	MC; O			●		RP		●	
155	Chikaraishi and Nakayama (2016)	O	MC; RC	●			●	RP	●	●	●
156	Habib et al. (2014)	O	MC; △MC		●			RP	●		●
157	Idris et al. (2015)	HCM	MC; △MC	●				RP-SP	●		●
158	Ahmad Termida et al. (2016)	MXL	□MC				●	RP		●	
159	Fatmi and Habib (2017)	MXL	□MC			●		RP	●		
160	Heinen and Ogilvie (2016)	MNL	□MC				●	RP	●	●	
161	Mitra et al. (2017)	MNL	□MC			●		RP	●	●	
162	Rahman and Baker (2018)	MNL	□MC	●				RP	●	●	
163	Standen et al. (2017)	MNL	□MC; △RC	●				RP	●	●	
164	Llorca et al. (2018)	MNL	MC; DC; TF			●		RP	●	●	
165	Manoj and Verma (2015a)	MNL	MC; O	●				RP	●		
166	Rahul and Verma (2018)	MNL; OLS	MC; TD	●				RP	●		
167	Popovich and Handy (2015)	MNL; OL	MC; TF	●				RP	●		
168	Kristoffersson et al. (2018)	NL	MC-DC	●				RP	●	●	
169	Fox et al. (2014)	NL	MC-DC		●			RP	●		●
170	Ding et al. (2014a)	CNL	MC-DT	●				RP	●	●	
171	Golshani et al. (2018)	NN; JDC	MC-DT	●				RP	●		●
172	Ermagun and Samimi (2018)	JDC	MC-O	●				RP	●	●	
173	Kaplan et al. (2016)	O	MC-O	●				RP	●	●	
174	Xiqun et al. (2015)	NL	MC-O	●				RP	●	●	
175	Habib (2014b)	O	MC-O-MTO	●				RP	●	●	
176	Habib (2014a)	JDC	MC-TD	●				RP	●	●	
177	Schoner et al. (2015)	O	MC-TF	●				RP	●		
178	Marquet and Miralles-Guasch (2016)	MNL	MTO; MC	●				RP		●	
179	Shen et al. (2016)	MNL; NL	MTO; MC	●				RP	●	●	
180	Khan et al. (2014)	O; MNL	MTO; TF; MC; O	●				RP	●	●	
181	Picard et al. (2018)	NL	MTO-MC	●				RP	●	●	
182	Suel and Polak (2017)	NL	O-MC				●	RP	●		●
183	Yang (2018)	MXL	O; DC	●				RP	●		
184	Daisy et al. (2018)	OP; MNL	O; MC				●	RP	●		
185	Ho and Mulley (2015)	NL	O-MC		●			RP	●	●	
186	Liu et al. (2017)	O; MDCEV	O-MS	●				RP	●	●	
187	Zhang et al. (2014)	NL	O-RC	●				RP	●	●	
188	Pang and Khani (2018)	MNL; MXL	DC	●				RP	●		
189	Alizadeh et al. (2018)	NL	RC				●	RP	●		●
190	Anderson et al. (2014)	MXL (PSCL)	RC	●				RP	●	●	
191	Baek and Sohn (2016)	MNL (PSL)	RC	●				RP	●		
192	Basheer et al. (2018)	MNL	RC	●				RP	●		
193	Chen and Wen (2013)	MNL	RC	●				RP	●	●	
194	Chen et al. (2018)	MXL(PSL)	RC				●	RP	●	●	
195	Li et al. (2016)	MXL(PSL; GNL; LK; CL, etc)	RC				●	RP	●		●

(*continued on next page*)

Table 5 (*continued*)

No	Article	Model(s)	Dependent variable(s)	Data characteristics: Data structure: CS	RCS	PP	TP	Data type: RP, RP-SP	Goodness of fit reported	Policy inference reported	Validation reported
196	Dalumpines and Scott (2017)	PSL	RC				●	RP	●		
197	Di et al. (2017)	MNP	RC				●	RP	●	●	
198	Garcia-Martinez et al. (2018)	MXL	RC	●				RP-SP	●	●	
199	Ghanayim and Bekhor (2018)	MXL(PSL; CL)	RC	●				RP	●	●	
200	Jánošíkova et al. (2014)	MNL	RC				●	RP	●	●	
201	Kazagli et al. (2016)	O	RC				●	RP	●	●	●
202	Zhang et al. (2018)	MNL	RC				●	RP	●		
203	Lai and Bierlaire (2015)	CNL	RC				●	RP	●		●
204	Mai (2016)	RCNL	RC				●	RP	●		●
205	Mai et al. (2017)	RL	RC				●	RP	●		●
206	Mai et al. (2015)	NRL	RC				●	RP	●		●
207	Moran et al. (2018)	MNL (PSL)	RC	●				RP	●	●	
208	Oyama and Hato (2018)	O	RC				●	RP	●		
209	Papola (2016)	O	MC	●				RP	●	●	●
210	Prato (2014)	MNL (RRM)	RC	●				RP	●	●	
211	Prato et al. (2018)	GMXL (PSL)	RC				●	RP	●	●	
212	Raveau et al. (2014)	MNL(CL)	RC	●			●	RP	●	●	
213	Thomas and Tutert (2015)	MNL	RC	●				RP	●		
214	Zhang et al. (2018)	NL	RC	●			●	RP	●		●
215	Yamamoto et al. (2018)	NL(PSL)	RC				●	RP	●		
216	Zhang et al. (2015)	MNL	RC	●				RP			●
217	Zhuang et al. (2017)	MNL	RC		●			RP		●	
218	Zimmermann et al. (2017)	NRL	RC				●	RP	●		●
219	Jánošíkova et al. (2014)	MNL	RC				●	RP	●	●	
220	Xie et al. (2016)	MXL	RC				●	RP	●		●
221	Yang (2016)	NL	RC	●				RP	●	●	
222	Duc-Nghiem et al. (2018)	MNL	RC	●				RP	●	●	●
223	Katoshevski et al. (2015)	MNL; NL	RLC; DC; MC; AC	●				RP	●		
224	Bhat et al. (2016)	O	RLC-MTO-D-MC	●				RP	●	●	
225	Tran et al. (2016)	O	RLC-O-MC	●				RP	●	●	
226	Sener and Reeder (2014)	O	TF-MC	●				RP	●		

Model abbreviations

CL: C-logit
CoRUM: Combination of RUM models
EPSL: Extended path size logit
FMS: Flexible model structure
GMXL: Generalized mixed logit
GNL: Generalized nested logit
HCM: Hybrid choice model
HDDC: heteroskedastic dynamic discrete choice
hTEV: heteroscedastic tree extreme value
IAL: independent availability logit
IBL: instance-based learning
JDC: Joint discrete continuous
JMXL: Joint mixed logit
LCM: latent class model
LK: Logit kernel
MRL: Mixed recursive logit
MVMLL: Multivariate multilevel binary logit
MVPM: multivariate probit model
MXL: Mixed logit
MXNL: Mixed nested logit
NN*: Neural network
NL: Nested logit
NRL: Nested recursive logit
O: Other extensions/generalizations
OP: Ordered probit
PL: Polarized logit
PLC: parameterized logit captivity
PSCL: Path sized correction logit
PSL: Path size logit
QL: q-logit
RCNL: Recursive cross nested logit
RL: Recursive logit
RRM: Random-regret maximization
SCL: Spatially correlated logit
SVM*: support vector machine
WB: weibit
*Machine learning models
**When several models of the same

Dependent variable abbreviations

AC: Activity choice
AT: Arrival time
DC: Destination choice
DT: Departure time choice/Time of day choice
MC: Mode choice
MS: Modal split
MTO: Mobility tool ownership
O: Other
RC: Route choice
RLC: Residential location choice
TF: Trip frequency

Data characteristics abbreviations

CS: Cross sectional data
RCS: Repeated cross section, pooled data
PP: Pseudo-panel data
TP: True panel data*
SP: Stated preference
RP: Revealed preference
*In this classification, true panel data are defined as any survey that measures travel behaviour of the same subjects at two or more different points in time. The smallest time unit is a day. (for example, repeated observations in the same day is classified as cross-sectional data, while a travel behaviour survey of two or more days is considered true panel data). This classification is irrespective of the way the data was handled by the analyst.
Stated preference surveys with multiple choice scenarios are considered cross-sectional data.

(continued on next page)

Table 5 (*continued*)

Model abbreviations		Dependent variable abbreviations	Data characteristics abbreviations
LMM: Linear mixed model MDCEV: Multiple discrete continuous extreme value ML: Mother logit MLMXL: Multilevel mixed logit MNL: Binary or Multinomial logit MNLp: MNL with panel data corrections MPM: Multinomial probit model	dependent variable are compared, only the most general form model or best performing model is listed in the Models (s) column.		

References

Abasahl, F., Kelarestaghi, K.B., Ermagun, A., 2018. Gender gap generators for bicycle mode choice in Baltimore college campuses. Elsevier Travel Behaviour and Society 11, 78–85. https://doi.org/10.1016/j.tbs.2018.01.002. November 2017.

Ahmad Termida, N., Susilo, Y.O., Franklin, J.P., 2016. Observing dynamic behavioural responses due to the extension of a tram line by using panel survey. Elsevier Ltd Transport. Res. Pol. Pract. 86, 78–95. https://doi.org/10.1016/j.tra.2016.02.005.

Alizadeh, H., et al., 2018. On the role of bridges as anchor points in route choice modeling. Springer US Transportation 45 (5), 1181–1206. https://doi.org/10.1007/s11116-017-9761-7.

Anderson, M.K., Nielsen, O.A., Prato, C.G., 2014. Multimodal route choice models of public transport passengers in the Greater Copenhagen Area. Springer Berlin Heidelberg EURO Journal on Transportation and Logistics 6 (3), 221–245. https://doi.org/10.1007/s13676-014-0063-3.

Anta, J., et al., 2016. Influence of the weather on mode choice in corridors with time-varying congestion: a mixed data study. Transportation 43 (2), 337–355. https://doi.org/10.1007/s11116-015-9578-1.

Arlot, S., Celisse, A., 2009. A survey of cross-validation procedures for model selection, 4, 40–79. https://doi.org/10.1214/09-SS054.

Arman, M.A., Khademi, N., de Lapparent, M., 2018. Women's mode and trip structure choices in daily activity-travel: a developing country perspective. Taylor & Francis Transport. Plann. Technol. 41 (8), 845–877. https://doi.org/10.1080/03081060.2018.1526931.

Assi, K.J., et al., 2018. Mode choice behavior of high school goers: evaluating logistic regression and MLP neural networks. Elsevier Case Studies on Transport Policy 6 (2), 225–230. https://doi.org/10.1016/j.cstp.2018.04.006.

Astegiano, P., et al., 2017. Quantifying the explanatory power of mobility-related attributes in explaining vehicle ownership decisions. Res. Transport. Econ. 66, 2–11. https://doi.org/10.1016/j.retrec.2017.07.007.

Austin, P.C., et al., 2016. 'Geographic and temporal validity of prediction models : different approaches were useful to examine model performance'. Elsevier Inc J. Clin. Epidemiol. 79, 76–85. https://doi.org/10.1016/j.jclinepi.2016.05.007.

Aziz, H.M.A., et al., 2018. Exploring the impact of walk–bike infrastructure, safety perception, and built-environment on active transportation mode choice: a random parameter model using New York City commuter data. Springer US Transportation 45 (5), 1207–1229. https://doi.org/10.1007/s11116-017-9760-8.

Baek, J., Sohn, K., 2016. An investigation into passenger preference for express trains during peak hours. Springer US Transportation 43 (4), 623–641. https://doi.org/10.1007/s11116-015-9592-3.

Baker, M., Penny, D., 2016. Is there a reproducibility crisis? Nature 533 (7604), 452–454. https://doi.org/10.1038/533452A.

Basheer, S., Srinivasan, K.K., Sivanandan, R., 2018. Investigation of information quality and user response to real-time traffic information under heterogeneous traffic conditions. Transportation in Developing Economies 4 (2), 1–11. https://doi.org/10.1007/s40890-018-0061-5. Springer International Publishing.

Basu, D., et al., 2018. Modeling choice behavior of non-mandatory tour locations in California – an experience. Travel Behaviour and Society 12, 122–129. https://doi.org/10.1016/j.tbs.2017.04.008.

Bhat, C.R., et al., 2016. On accommodating spatial interactions in a Generalized Heterogeneous Data Model (GHDM) of mixed types of dependent variables. Transp. Res. Part B Methodol. 94, 240–263. https://doi.org/10.1016/j.trb.2016.09.002. Elsevier Ltd.

Böcker, L., van Amen, P., Helbich, M., 2017. Elderly travel frequencies and transport mode choices in Greater Rotterdam, The Netherlands. Transportation 44 (4), 831–852. https://doi.org/10.1007/s11116-016-9680-z.

Bohluli, S., Ardekani, S., Daneshgar, F., 2014. Development and validation of a direct mode choice model. Transport. Plann. Technol. 37 (7), 649–662. https://doi.org/10.1080/03081060.2014.935571. Taylor & Francis.

Brathwaite, T., Walker, J.L., 2018. Causal inference in travel demand modeling (and the lack thereof). Elsevier Ltd Journal of Choice Modelling 26, 1–18. https://doi.org/10.1016/j.jocm.2017.12.001. June 2017.

Braun, L.M., et al., 2016. Short-term planning and policy interventions to promote cycling in urban centers: findings from a commute mode choice analysis in Barcelona, Spain. Transport. Res. Pol. Pract. 89, 164–183. https://doi.org/10.1016/j.tra.2016.05.007. Elsevier Ltd.

Bridgelall, R., 2014. Campus parking supply impacts on transportation mode choice. Transport. Plann. Technol. 37 (8), 711–737. https://doi.org/10.1080/03081060.2014.959354. Taylor & Francis.

Brier, G., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78 (1), 2–4.

Bueno, P.C., et al., 2017. Understanding the effects of transit benefits on employees' travel behavior: evidence from the New York-New Jersey region. Transport. Res. Pol. Pract. 99, 1–13. https://doi.org/10.1016/j.tra.2017.02.009. Elsevier Ltd.

Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76 (3), 503–514.

Calster, B. Van, et al., 2012. Extending the c -statistic to nominal polytomous outcomes : the Polytomous Discrimination Index. Stat. Med. 31, 2610–2626. https://doi.org/10.1002/sim.5321.

Cambridge Systematics, 2010. Travel Model Validation and Reasonability Checking Manual, 2nd. Federal Highway Administration.

Castillo-Manzano, J.I., Castro-Nuño, M., López-Valpuesta, L., 2015. Analyzing the transition from a public bicycle system to bicycle ownership: a complex relationship. Transport. Res. Transport Environ. 38, 15–26. https://doi.org/10.1016/j.trd.2015.04.004.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. Geosci. Model Dev. (GMD) 7 (3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Chakour, V., Eluru, N., 2014. Analyzing commuter train user behavior: a decision framework for access mode and station choice. Transportation 41 (1), 211–228. https://doi.org/10.1007/s11116-013-9509-y.

Chen, D.J., Wen, Y.-H., 2013. 'Effects of freeway mileage-based toll scheme on the short-range driver's route choice behavior'. J. Urban Plann. Dev. 140 (2), 04013012 https://doi.org/10.1061/(asce)up.1943-5444.0000167.

Chen, P., Shen, Q., Childress, S., 2018. A GPS data-based analysis of built environment influences on bicyclist route preferences. International Journal of Sustainable Transportation 12 (3), 218–231. https://doi.org/10.1080/15568318.2017.1349222. Taylor & Francis.

Cherchi, E., Cirillo, C., 2014. Understanding variability, habit and the effect of long period activity plan in modal choices: a day to day, week to week analysis on panel data. Transportation 41 (6), 1245–1262. https://doi.org/10.1007/s11116-014-9549-y.

Cherchi, E., Cirillo, C., Ortúzar, J. de D., 2017. Modelling correlation patterns in mode choice models estimated on multiday travel data. Transport. Res. Pol. Pract. 96, 146–153. https://doi.org/10.1016/j.tra.2016.11.021. Elsevier Ltd.

Chica-Olmo, J., Rodríguez-López, C., Chillón, P., 2018. Effect of distance from home to school and spatial dependence between homes on mode of commuting to school. J. Transport Geogr. 72 (August), 1–12. https://doi.org/10.1016/j.jtrangeo.2018.07.013. Elsevier.

Chikaraishi, M., Nakayama, S., 2016. Discrete choice models with q-product random utilities. Transp. Res. Part B Methodol. 93, 576–595. https://doi.org/10.1016/j.trb.2016.08.013. Elsevier Ltd.

Clark, A.F., Scott, D.M., Yiannakoulias, N., 2014. Examining the relationship between active travel, weather, and the built environment: a multilevel approach using a GPS-enhanced dataset. Transportation 41 (2), 325–338. https://doi.org/10.1007/s11116-013-9476-3.

Clifton, K.J., et al., 2016. Development of destination choice models for pedestrian travel. Transport. Res. Pol. Pract. 94, 255–265. https://doi.org/10.1016/j.tra.2016.09.017. Elsevier Ltd.

Cole-Hunter, T., et al., 2015. Objective correlates and determinants of bicycle commuting propensity in an urban environment. Transport. Res. Transport Environ. 40 (2), 132–143. https://doi.org/10.1016/j.trd.2015.07.004. Elsevier Ltd.

Collins, P.A., MacFarlane, R., 2018. Evaluating the determinants of switching to public transit in an automobile-oriented mid-sized Canadian city: a longitudinal analysis. Transport. Res. Pol. Pract. 118 (September), 682–695. https://doi.org/10.1016/j.tra.2018.10.014. Elsevier.

Daisy, N.S., Millward, H., Liu, L., 2018. Trip chaining and tour mode choice of non-workers grouped by daily activity patterns. Elsevier J. Transport Geogr. 69, 150–162. https://doi.org/10.1016/j.jtrangeo.2018.04.016. November 2017.

Dalumpines, R., Scott, D.M., 2017. Determinants of route choice behavior: a comparison of shop versus work trips using the Potential Path Area - gateway (PPAG) algorithm and Path-Size Logit. J. Transport Geogr. 59, 59–68. https://doi.org/10.1016/j.jtrangeo.2017.01.003. Elsevier Ltd.

Danaf, M., Abou-Zeid, M., Kaysi, I., 2014. 'Modeling travel choices of students at a private, urban university: insights and policy implications', *Case Studies on Transport Policy*. World Conference on Transport Research Society 2 (3), 142–152. https://doi.org/10.1016/j.cstp.2014.08.006.

Danalet, A., et al., 2016. Location choice with longitudinal WiFi data. Journal of Choice Modelling 18, 1–17. https://doi.org/10.1016/j.jocm.2016.04.003. Elsevier.

de Grange, L., et al., 2015. A logit model with endogenous explanatory variables and network externalities. Network. Spatial Econ. 15 (1), 89–116. https://doi.org/10.1007/s11067-014-9271-5.

de Jong, G.C., 2014. Mode choice models. In: Tavasszy, L.A., de Jong, G.C. (Eds.), Modelling Freight Transport (Elsevier).

de Luca, S., Cantarella, G.E., 2009. Validation and comparison of choice models. In: Saleh, W., Sammer, G. (Eds.), Travel Demand Management and Road User Pricing: Success, Failure and Feasibility. Ashgate publications, pp. 37–58.

de Luca, S., Di Pace, R., 2015. 'Modelling users' behaviour in inter-urban carsharing program: a stated preference approach'. Transport. Res. Pol. Pract. 71, 59–76. https://doi.org/10.1016/j.tra.2014.11.001. Elsevier Ltd.

Deutsch-Burgner, K., Ravualaparthy, S., Goulias, K., 2014. Place happiness: its constituents and the influence of emotions and subjective importance on activity type and destination choice. Transportation 41 (6), 1323–1340. https://doi.org/10.1007/s11116-014-9553-2.

Di Ciommo, F., et al., 2014. 'Exploring the role of social capital influence variables on travel behaviour. Transport. Res. Pol. Pract. 68, 46–55. https://doi.org/10.1016/j.tra.2014.08.018. Elsevier Ltd.

Di, X., et al., 2017. Indifference bands for boundedly rational route switching. Transportation 44 (5), 1169–1194. https://doi.org/10.1007/s11116-016-9699-1. Springer US.

Ding, C., et al., 2014a. 'Cross-Nested joint model of travel mode and departure time choice for urban commuting trips: case study in Maryland–Washington, DC region'. J. Urban Plann. Dev. 141 (4), 04014036 https://doi.org/10.1061/(asce)up.1943-5444.0000238.

Ding, C., Lin, Y., Liu, C., 2014b. Exploring the influence of built environment on tour-based commuter mode choice: a cross-classified multilevel modeling approach. Transport. Res. Transport Environ. 32, 230–238. https://doi.org/10.1016/j.trd.2014.08.001. Elsevier Ltd.

Ding, C., et al., 2017. Exploring the influence of attitudes to walking and cycling on commute mode choice using a hybrid choice model. J. Adv. Transport. 2017, 1–8. https://doi.org/10.1155/2017/8749040.

Ding, C., et al., 2018. Joint analysis of the spatial impacts of built environment on car ownership and travel mode choice. Transport. Res. Transport Environ. 60, 28–40. https://doi.org/10.1016/j.trd.2016.08.004. Elsevier Ltd.

Dong, H., Ma, L., Broach, J., 2016. Promoting sustainable travel modes for commute tours: a comparison of the effects of home and work locations and employer-provided incentives. International Journal of Sustainable Transportation 10 (6), 485–494. https://doi.org/10.1080/15568318.2014.1002027. Taylor & Francis.

Duc-Nghiem, N., et al., 2018. 'Modeling cyclists' facility choice and its application in bike lane usage forecasting', *IATSS Research*. International Association of Traffic and Safety Sciences 42 (2), 86–95. https://doi.org/10.1016/j.iatssr.2017.06.006.

Efron, B., 1983. 'Estimating the error rate of a prediction Rule : improvement on cross-validation'. J. Am. Stat. Assoc. 78 (382), 316–331.

Efron, B., Tibshirani, R., 1993. An Introduction to the Boostrap. Chapman & Hall.

Efron, B., Tibshirani, R., 1995. Cross-Validation and the Bootstrap : Estimating the Error Rate of a Prediction Rule. Division of Biostatistics, Stanford University.

Efron, B., Tibshirani, R., 1997. 'Improvements on cross-validation : the 632 + bootstrap method'. J. Am. Stat. Assoc. (438), 92. https://doi.org/10.1080/01621459.1997.10474007.

Efthymiou, D., Antoniou, C., 2017. 'Understanding the effects of economic crisis on public transport users' satisfaction and demand'. Transport Pol. 53 (August 2015), 89–97. https://doi.org/10.1016/j.tranpol.2016.09.007. Elsevier.

Ermagun, A., Levinson, D., 2017. Public transit, active travel, and the journey to school: a cross-nested logit analysis. Transportmetrica: Transport. Sci. 13 (1), 24–37. https://doi.org/10.1080/23249935.2016.1207723.

Ermagun, A., Samimi, A., 2015. Promoting active transportation modes in school trips. Transport Pol. 37, 203–211. https://doi.org/10.1016/j.tranpol.2014.10.013. Elsevier.

Ermagun, A., Samimi, A., 2018. 'Mode choice and travel distance joint models in school trips'. Transportation 45 (6), 1755–1781. https://doi.org/10.1007/s11116-017-9794-y. Springer US.

Ermagun, A., Hossein Rashidi, T., Samimi, A., 2015. A joint model for mode choice and escort decisions of school trips. Transportmetrica: Transport. Sci. 11 (3), 270–289. https://doi.org/10.1080/23249935.2014.968654.

Faghih-Imani, A., Eluru, N., 2015. 'Analysing bicycle-sharing system user destination choice preferences: chicago's Divvy system'. J. Transport Geogr. 44, 53–64. https://doi.org/10.1016/j.jtrangeo.2015.03.005. Elsevier Ltd.

Faghih-Imani, A., Eluru, N., 2017. Examining the impact of sample size in the analysis of bicycle-sharing systems. Transportmetrica: Transport. Sci. 13 (2), 139–161. https://doi.org/10.1080/23249935.2016.1223205.

Fatmi, M.R., Habib, M.A., 2017. Modelling mode switch associated with the change of residential location. Travel Behaviour and Society 9 (August), 21–28. https://doi.org/10.1016/j.tbs.2017.07.006. Elsevier.

Fernández-Antolín, A., et al., 2016. Correcting for endogeneity due to omitted attitudes: empirical assessment of a modified MIS method using RP mode choice data. Journal of Choice Modelling 20, 1–15. https://doi.org/10.1016/j.jocm.2016.09.001. Elsevier.

Flügel, S., et al., 2015. Methodological challenges in modelling the choice of mode for a new travel alternative using binary stated choice data - the case of high speed rail in Norway. Transport. Res. Pol. Pract. 78, 438–451. https://doi.org/10.1016/j.tra.2015.06.004.

Forsey, D., et al., 2014. Temporal transferability of work trip mode choice models in an expanding suburban area: the case of York Region, Ontario. Transportmetrica: Transport. Sci. 10 (6), 469–482. https://doi.org/10.1080/23249935.2013.788100.

Fox, J., et al., 2014. Temporal transferability of models of mode-destination choice for the Greater Toronto and Hamilton Area. Journal of Transport and Land Use 7 (2), 41. https://doi.org/10.5198/jtlu.v7i2.701.

Fu, X., Juan, Z., 2017. 'Accommodating preference heterogeneity in commuting mode choice: an empirical investigation in Shaoxing, China. Transport. Plann. Technol. 40 (4), 434–448. https://doi.org/10.1080/03081060.2017.1300240. Taylor & Francis.

Gao, J., et al., 2016. 'A study of traveler behavior under traffic information-provided conditions in the Beijing area'. Transport. Plann. Technol. 39 (8), 768–778. https://doi.org/10.1080/03081060.2016.1231896. Taylor & Francis.

Gao, Y., et al., 2017. 'Differences in pupils' school commute characteristics and mode choice based on the household registration system in China'. Case Studies on Transport Policy 5 (4), 656–661. https://doi.org/10.1016/j.cstp.2017.07.008. Elsevier.

Garcia-Martinez, A., et al., 2018. 'Transfer penalties in multimodal public transport networks'. Transport. Res. Pol. Pract. 114 (January), 52–66. https://doi.org/10.1016/j.tra.2018.01.016. Elsevier.

Geisser, S., 1975. The predictive sample reuse method with applications. J. Am. Stat. Assoc. 70 (350), 320–328.

Gerber, P., et al., 2017. 'Cross-border residential mobility, quality of life and modal shift: a Luxembourg case study'. Transport. Res. Pol. Pract. 104, 238–254. https://doi.org/10.1016/j.tra.2017.06.015. Elsevier Ltd.

Ghanayim, M., Bekhor, S., 2018. Modelling bicycle route choice using data from a GPS-assisted household survey. Eur. J. Transport Infrastruct. Res. 18 (2), 158–177.

Glerum, A., Atasoy, B., Bierlaire, M., 2014. Using semi-open questions to integrate perceptions in choice models. Journal of Choice Modelling 10 (1), 11–33. https://doi.org/10.1016/j.jocm.2013.12.001. Elsevier.

Goel, R., Tiwari, G., 2016. 'Access-egress and other travel characteristics of metro users in Delhi and its satellite cities'. *IATSS Research*. The Authors 39 (2), 164–172. https://doi.org/10.1016/j.iatssr.2015.10.001.

Gokasar, I., Gunay, G., 2017. Mode choice behavior modeling of ground access to airports: a case study in Istanbul, Turkey. J. Air Transport. Manag. 59, 1–7. https://doi.org/10.1016/j.jairtraman.2016.11.003. Elsevier Ltd.

Golshani, N., et al., 2018. Modeling travel mode and timing decisions: comparison of artificial neural networks and copula-based joint model. Elsevier Travel Behaviour and Society 10, 21–32. https://doi.org/10.1016/j.tbs.2017.09.003. October 2017.

González, F., Melo-Riquelme, C., de Grange, L., 2016. A combined destination and route choice model for a bicycle sharing system. Transportation 43 (3), 407–423. https://doi.org/10.1007/s11116-015-9581-6.

Guan, J., Xu, C., 2018. 'Are relocatees different from others? Relocatee's travel mode choice and travel equity analysis in large-scale residential areas on the periphery of megacity Shanghai, China'. Transport. Res. Pol. Pract. 111 (January), 162–173. https://doi.org/10.1016/j.tra.2018.03.011. Elsevier.

Guerra, E., et al., 2018. 'Urban form, transit supply, and travel behavior in Latin America: evidence from Mexico's 100 largest urban areas'. Transport Pol. 69 (June), 98–105. https://doi.org/10.1016/j.tranpol.2018.06.001. Elsevier Ltd.

Gunn, H., Bates, J., 1982. Statistical aspects of travel demand modelling. Transport. Res. Gen. 16 (5–6), 371–382. https://doi.org/10.1016/0191-2607(82)90065-6.

Guo, Y., et al., 2018. 'Impacts of internal migration, household registration system, and family planning policy on travel mode choice in China'. Travel Behaviour and Society 13 (June), 128–143. https://doi.org/10.1016/j.tbs.2018.07.003. Elsevier.

Habib, K.N., 2014a. An investigation on mode choice and travel distance demand of older people in the National Capital Region (NCR) of Canada: application of a utility theoretic joint econometric model. Transportation 42 (1), 143–161. https://doi.org/10.1007/s11116-014-9537-2.

Habib, K.N., 2014b. Household-level commuting mode choices, car allocation and car ownership level choices of two-worker households: the case of the city of Toronto. Transportation 41 (3), 651–672. https://doi.org/10.1007/s11116-014-9518-5.

Habib, K.M.N., Sasic, A., 2014. A GEV model with scale heterogeneity for investigating the role of mobility tool ownership in peak period non-work travel mode choices. Journal of Choice Modelling 10 (1), 46–59. https://doi.org/10.1016/j.jocm.2014.01.003. Elsevier.

Habib, K.N., Weiss, A., 2014. Evolution of latent modal captivity and mode choice patterns for commuting trips: a longitudinal analysis using repeated cross-sectional datasets. Transport. Res. Pol. Pract. 66 (1), 39–51. https://doi.org/10.1016/j.tra.2014.04.013. Elsevier Ltd.

Habib, K.M.N., Swait, J., Salem, S., 2014. Using repeated cross-sectional travel surveys to enhance forecasting robustness: accounting for changing mode preferences. Transport. Res. Pol. Pract. 67, 110–126. https://doi.org/10.1016/j.tra.2014.06.004. Elsevier Ltd.

Habib, K.N., et al., 2014. 'Synopsis of bicycle demand in the City of Toronto: investigating the effects of perception, consciousness and comfortability on the purpose of biking and bike ownership'. Transport. Res. Pol. Pract. 70, 67–80. https://doi.org/10.1016/j.tra.2014.09.012. Elsevier Ltd.

Halldórsdóttir, K., Nielsen, O.A., Prato, C.G., 2017. Home-end and activity-end preferences for access to and egress from train stations in the Copenhagen region. International Journal of Sustainable Transportation 11 (10), 776–786. https://doi.org/10.1080/15568318.2017.1317888. Taylor & Francis.

Hasnine, M.S., Habib, K.N., 2018. What about the dynamics in daily travel mode choices? A dynamic discrete choice approach for tour-based mode choice modelling'. Transport Pol. 71 (August), 70–80. https://doi.org/10.1016/j.tranpol.2018.07.011. Elsevier Ltd.

Hasnine, M.S., et al., 2018. Determinants of travel mode choices of post-secondary students in a large metropolitan area: the case of the city of Toronto. J. Transport Geogr. 70 (June), 161–171. https://doi.org/10.1016/j.jtrangeo.2018.06.003. Elsevier.

He, S.Y., Giuliano, G., 2017. 'Factors affecting children's journeys to school: a joint escort-mode choice model'. Transportation 44 (1), 199–224. https://doi.org/10.1007/s11116-015-9634-x. Springer US.

He, S.Y., Giuliano, G., 2018. 'School choice: understanding the trade-off between travel distance and school quality'. Transportation 45 (5), 1475–1498. https://doi.org/10.1007/s11116-017-9773-3. Springer US.

Heinen, E., 2016. 'Identity and travel behaviour: a cross-sectional study on commute mode choice and intention to change'. Transport. Res. F Traffic Psychol. Behav. 43, 238–253. https://doi.org/10.1016/j.trf.2016.10.016. Elsevier Ltd.

Heinen, E., Ogilvie, D., 2016. Variability in baseline travel behaviour as a predictor of changes in commuting by active travel, car and public transport: a natural experimental study. Journal of Transport and Health 3 (1), 77–85. https://doi.org/10.1016/j.jth.2015.11.002. Elsevier.

Helbich, M., 2017. 'Children's school commuting in The Netherlands: does it matter how urban form is incorporated in mode choice models?'. International Journal of Sustainable Transportation 11 (7), 507–517. https://doi.org/10.1080/15568318.2016.1275892. Taylor & Francis.

Helbich, M., Böcker, L., Dijst, M., 2014. Geographic heterogeneity in cycling under various weather conditions: evidence from Greater Rotterdam. J. Transport Geogr. 38, 38–47. https://doi.org/10.1016/j.jtrangeo.2014.05.009. Elsevier Ltd.

Hensher, D.A., Ho, C.Q., 2016. 'Experience conditioning in commuter modal choice modelling – does it make a difference?'. Transport. Res. E Logist. Transport. Rev. 95, 164–176. https://doi.org/10.1016/j.tre.2016.09.010. Elsevier Ltd.

Hess, S., Rose, J.M., 2009. Some lessons in stated choice survey design. European Transport Conference. Association for European Transport, Noordwijkerhout, The Netherlands.

Ho, C.Q., Hensher, D.A., 2016. A workplace choice model accounting for spatial competition and agglomeration effects. J. Transport Geogr. 51, 193–203. https://doi.org/10.1016/j.jtrangeo.2016.01.005. Elsevier Ltd.

Ho, C., Mulley, C., 2015. Intra-household Interactions in tour-based mode choice: the role of social, temporal, spatial and resource constraints. Transport Pol. 38, 52–63. https://doi.org/10.1016/j.tranpol.2014.12.001. Elsevier.

Holguín-Veras, J., Thorson, E., Ozbay, K., 2004. 'Preliminary results of experimental economics application to urban goods modeling research', *Transportation Research Record*. Journal of the Transportation Research Board 1873 (1), 9–16.

Hsu, H.P., Saphores, J.D., 2014. 'Impacts of parental gender and attitudes on children's school travel mode and parental chauffeuring behavior: results for California based on the 2009 National Household Travel Survey'. Transportation 41 (3), 543–565. https://doi.org/10.1007/s11116-013-9500-7.

Huang, A., Levinson, D., 2015. 'Axis of travel: modeling non-work destination choice with GPS data'. Transport. Res. C Emerg. Technol. 58, 208–223. https://doi.org/10.1016/j.trc.2015.03.022. Elsevier Ltd.

Huang, A., Levinson, D., 2017. A model of two-destination choice in trip chains with GPS data. Journal of Choice Modelling 24, 51–62. https://doi.org/10.1016/j.jocm.2016.04.002. Elsevier Ltd.

Hurtubia, R., et al., 2014. 'Integrating psychometric indicators in latent class choice models'. Transport. Res. Pol. Pract. 64, 135–146. https://doi.org/10.1016/j.tra.2014.03.010. Elsevier Ltd.

Idris, A.O., Nurul Habib, K.M., Shalaby, A., 2015. An investigation on the performances of mode shift models in transit ridership forecasting. Transport. Res. Pol. Pract. 78, 551–565. https://doi.org/10.1016/j.tra.2015.06.012. Elsevier Ltd.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2 (8), 696–701. https://doi.org/10.1371/journal.pmed.0020124.

Irfan, M., et al., 2018. Policy implications of work-trip mode choice using econometric modeling. J. Transport. Eng., Part A: Systems 144 (8), 04018035. https://doi.org/10.1061/jtepbs.0000158.

Jánošíkova, L., Slavík, J., Koháni, M., 2014. Estimation of a route choice model for urban public transport using smart card data. Transport. Plann. Technol. 37 (7), 638–648. https://doi.org/10.1080/03081060.2014.935570. Taylor & Francis.

Ji, Y., et al., 2017. Public bicycle as a feeder mode to rail transit in China: the role of gender, age, income, trip purpose, and bicycle theft experience. International Journal of Sustainable Transportation 11 (4), 308–317. https://doi.org/10.1080/15568318.2016.1253802. Taylor & Francis.

Justice, A.C., Covinsky, K.E., Berlin, J.A., 1999. Assessing the generalizability of prognostic information. Ann. Intern. Med. 130, 515–524.

Kamargianni, M., Ben-Akiva, M., Polydoropoulou, A., 2014. Incorporating social interaction into hybrid choice models. Transportation 41 (6), 1263–1285. https://doi.org/10.1007/s11116-014-9550-5.

Kamruzzaman, M., et al., 2015. Commuting mode choice in transit oriented development: disentangling the effects of competitive neighbourhoods, travel attitudes, and self-selection. Transport Pol. 42, 187–196. https://doi.org/10.1016/j.tranpol.2015.06.003.

Kaplan, S., Nielsen, T.A.S., Prato, C.G., 2016. Walking, cycling and the urban form: a Heckman selection model of active travel mode and distance by young adolescents. Transport. Res. Transport Environ. 44, 55–65. https://doi.org/10.1016/j.trd.2016.02.011. Elsevier Ltd.

Katoshevski, R., et al., 2015. Integrating activity-based travel-demand models with land-use and other long-term lifestyle decisions. Journal of Transport and Land Use 8 (3), 71–93.

Kazagli, E., Bierlaire, M., Flötteröd, G., 2016. Revisiting the route choice problem: a modeling framework based on mental representations. Journal of Choice Modelling 19, 1–23. https://doi.org/10.1016/j.jocm.2016.06.001. Elsevier.

Keyes, A.K.M., Crawford-Brown, D., 2018. 'The changing influences on commuting mode choice in urban England under Peak Car: a discrete choice modelling approach'. *Transport. Res. Part F: Traffic Psychology and Behaviour*. The Authors 58, 167–176. https://doi.org/10.1016/j.trf.2018.06.010.

Khan, M., Kockelman, K.M., Xiong, X., 2014. Models for anticipating non-motorized travel choices, and the role of the built environment. Transport Pol. 35, 117–126. https://doi.org/10.1016/j.tranpol.2014.05.008. Elsevier.

Khan, S., et al., 2016. Toward sustainable urban mobility: investigating nonwork travel behavior in a sprawled Canadian city. International Journal of Sustainable Transportation 10 (4), 321–331. https://doi.org/10.1080/15568318.2014.928838. Taylor & Francis.

Kim, J., 2009. 'Estimating classification error rate : repeated cross-validation , repeated hold-out and bootstrap'. Comput. Stat. Data Anal. 53, 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009. Elsevier B.V.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence (IJCAI).

Koppelman, F., Wilmot, C.G., 1982. Transferability analysis of disaggregate choice models. Transport. Res. Rec. 895, 18–24.

Kristoffersson, I., Daly, A., Algers, S., 2018. Modelling the attraction of travel to shopping destinations in large-scale modelling. Transport Pol. 68, 52–62. https://doi.org/10.1016/j.tranpol.2018.04.013. June 2017.

Kunhikrishnan, P., Srinivasan, K.K., 2018. Investigating behavioral differences in the choice of distinct Intermediate Public Transport (IPT) modes for work trips in Chennai city. Transport Pol. 61 (October 2017), 111–122. https://doi.org/10.1016/j.tranpol.2017.10.006. Elsevier Ltd.

Lai, X., Bierlaire, M., 2015. Specification of the Cross-Nested Logit Model with Sampling of Alternatives for Route Choice models Transportation Research Part B: Methodological, vol. 80, pp. 220–234. https://doi.org/10.1016/j.trb.2015.07.005. Elsevier Ltd.

Larsen, K., et al., 2018. A matched pairs approach to assessing parental perceptions and preferences for mode of travel to school. Journal of Transport and Health 11 (September), 56–63. https://doi.org/10.1016/j.jth.2018.09.004. Elsevier Ltd.

Lee, J., 2015. Impact of neighborhood walkability on trip generation and trip chaining: case of Los angeles. J. Urban Plann. Dev. 142 (3), 05015013 https://doi.org/10.1061/(asce)up.1943-5444.0000312.

Lee, J.S., Nam, J., Lee, S.S., 2014. Built environment impacts on individual mode choice: an empirical study of the houston-galveston metropolitan area. International Journal of Sustainable Transportation 8 (6), 447–470. https://doi.org/10.1080/15568318.2012.716142.

Lee, J., He, S.Y., Sohn, D.W., 2017. Potential of converting short car trips to active trips: the role of the built environment in tour-based travel. Journal of Transport and Health 7 (October), 134–148. https://doi.org/10.1016/j.jth.2017.08.008. Elsevier Ltd.

Li, W., Kamargianni, M., 2018. Providing quantified evidence to policy makers for promoting bike-sharing in heavily air-polluted cities: a mode choice model and policy simulation for Taiyuan-China. Transport. Res. Pol. Pract. 111 (January), 277–291. https://doi.org/10.1016/j.tra.2018.01.019. Elsevier.

Li, D., et al., 2016. Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets. Transport. Res. C Emerg. Technol. 67, 31–46. https://doi.org/10.1016/j.trc.2016.02.002. Elsevier Ltd.

Lin, J.J., et al., 2018. 'Built environment and public bike usage for metro access: a comparison of neighborhoods in Beijing, Taipei, and Tokyo'. Transport. Res. Transport Environ. 63 (1), 209–221. https://doi.org/10.1016/j.trd.2018.05.007. Elsevier.

Liu, C., Susilo, Y.O., Karlström, A., 2015. 'The influence of weather characteristics variability on individual's travel mode choice in different seasons and regions in Sweden'. Transport Pol. 41, 147–158. https://doi.org/10.1016/j.tranpol.2015.01.001. Elsevier.

Liu, C., Susilo, Y.O., Karlström, A., 2017. 'Jointly modelling individual's daily activity-travel time use and mode share by a nested multivariate Tobit model system'. Transportmetrica: Transport. Sci. 13 (6), 491–518. https://doi.org/10.1080/23249935.2017.1298682.

Liu, Y., et al., 2018. 'Investigating the effect of the spatial relationship between home, workplace and school on parental chauffeurs' daily travel mode choice'. Transport Pol. 69 (June), 78–87. https://doi.org/10.1016/j.tranpol.2018.06.004. Elsevier Ltd.

Llorca, C., et al., 2018. The usage of location based big data and trip planning services for the estimation of a long-distance travel demand model. Predicting the impacts of a new high speed rail corridor'. Elsevier Res. Transport. Econ. 72, 27–36. https://doi.org/10.1016/j.retrec.2018.06.004. October 2017.

Lorenzo Varela, J.M., Börjesson, M., Daly, A., 2018. Public transport: one mode or several? Transport. Res. Pol. Pract. 113 (March), 137–156. https://doi.org/10.1016/j.tra.2018.03.018.

Ma, T.Y., et al., 2015. Mode choice with latent preference heterogeneity: a case study for employees of the EU institutions in Luxembourg. Transportmetrica: Transport. Sci. 11 (5), 441–463. https://doi.org/10.1080/23249935.2015.1007175.

Ma, X., et al., 2018. Joint analysis of the commuting departure time and travel mode choice: role of the built environment. J. Adv. Transport. 2018, 1–13. https://doi.org/10.1155/2018/4540832.

Mahmoud, M.S., Weiss, A., Habib, K.N., 2016. Myopic choice or rational decision making? An investigation into mode choice preference structures in competitive modal arrangements in a multimodal urban area, the City of Toronto. Can. J. Civ. Eng. 43 (5), 420–428. https://doi.org/10.1139/cjce-2015-0336.

Mahpour, A., et al., 2018. 'Shopping destination choice in Tehran: an integrated choice and latent variable approach'. Transport. Res. F Traffic Psychol. Behav. 58, 566–580. https://doi.org/10.1016/j.trf.2018.06.045. Elsevier Ltd.

Mai, T., 2016. 'A method of integrating correlation structures for a generalized recursive route choice model'. Transp. Res. Part B Methodol. 93, 146–161. https://doi.org/10.1016/j.trb.2016.07.016. Elsevier Ltd.

Mai, T., Fosgerau, M., Frejinger, E., 2015. A nested recursive logit model for route choice analysis. Transp. Res. Part B Methodol. 75, 100–112. https://doi.org/10.1016/j.trb.2015.03.015. Elsevier Ltd.

Mai, T., Bastin, F., Frejinger, E., 2017. On the similarities between random regret minimization and mother logit: the case of recursive route choice models. Journal of Choice Modelling 23 (December 2016), 21–33. https://doi.org/10.1016/j.jocm.2017.03.002. Elsevier Ltd.

Manoj, M., Verma, A., 2015a. Activity-travel behaviour of non-workers belonging to different income group households in Bangalore, India. J. Transport Geogr. 49, 99–109. https://doi.org/10.1016/j.jtrangeo.2015.10.017. Elsevier Ltd.

Manoj, M., Verma, A., 2015b. 'Activity-travel behaviour of non-workers from Bangalore City in India. Transport. Res. Pol. Pract. 78, 400–424. https://doi.org/10.1016/j.tra.2015.06.006. Elsevier Ltd.

Marquet, O., Miralles-Guasch, C., 2016. City of Motorcycles. On how objective and subjective factors are behind the rise of two-wheeled mobility in Barcelona. Transport Pol. 52, 37–45. https://doi.org/10.1016/j.tranpol.2016.07.002. Elsevier.

Mattisson, K., et al., 2018. Modelling the association between health indicators and commute mode choice: a cross-sectional study in southern Sweden. Journal of Transport and Health 11 (June), 110–121. https://doi.org/10.1016/j.jth.2018.10.012.
Mehdizadeh, M., Nordfjaern, T., Mamdoohi, A.R., 2018. 'The role of socio-economic, built environment and psychological factors in parental mode choice for their children in an Iranian setting'. Transportation 45 (2), 523–543. https://doi.org/10.1007/s11116-016-9737-z. Springer US.
Meng, M., Koh, P., Wong, Y., 2016. Influence of socio-demography and operating streetscape on last-mile mode choice. Journal of Public Transportation 19 (2), 38–54. https://doi.org/10.5038/2375-0901.19.2.3.
Minal, S., Ravi, S.C., 2016. 'Commuter's sensitivity in mode choice: an empirical study of New Delhi'. J. Transport Geogr. 57, 207–217. https://doi.org/10.1016/j.jtrangeo.2016.11.001. Elsevier B.V.
Mitra, R., Buliung, R.N., 2014. The influence of neighborhood environment and household travel interactions on school travel behavior: an exploration using geographically-weighted models'. J. Transport Geogr. 36, 69–78. https://doi.org/10.1016/j.jtrangeo.2014.03.002. Elsevier Ltd.
Mitra, R., Buliung, R.N., 2015. Exploring differences in school travel mode choice behaviour between children and youth. Transport Pol. 42, 4–11. https://doi.org/10.1016/j.tranpol.2015.04.005. Elsevier.
Mitra, R., Papaioannou, E.M., Habib, K.M.N., 2015. Past and present of active school transportation: an exploration of the built environment effects in toronto, Canada from 1986 to 2006. Journal of Transport and Land Use 25–41. https://doi.org/10.5198/jtlu.2015.537.
Mitra, R., Ziemba, R.A., Hess, P.M., 2017. 'Mode substitution effect of urban cycle tracks: case study of a downtown street in Toronto, Canada. International Journal of Sustainable Transportation 11 (4), 248–256. https://doi.org/10.1080/15568318.2016.1249443. Taylor & Francis.
Moniruzzaman, M., Farber, S., 2018. 'What drives sustainable student travel? Mode choice determinants in the Greater Toronto Area'. International Journal of Sustainable Transportation 12 (5), 367–379. https://doi.org/10.1080/15568318.2017.1377326. Taylor & Francis.
Moran, M.R., Rodríguez, D.A., Corburn, J., 2018. 'Examining the role of trip destination and neighborhood attributes in shaping environmental influences on children's route choice'. Transport. Res. Transport Environ. 65 (August), 63–81. https://doi.org/10.1016/j.trd.2018.08.001. Elsevier.
Myung-Jin, J., Ji-Eun, J., Hyun-Ju, A., 2018. The welfare effects of the free subway fare scheme for seniors: a discrete choice approach with the case of Seoul. Case Studies on Transport Policy 6 (4), 642–650. https://doi.org/10.1016/j.cstp.2018.08.003. Elsevier.
Nguyen-Phuoc, D., et al., 2018. 'Mode choice among university students to school in Danang, Vietnam'. Elsevier Travel Behaviour and Society 13, 1–10. https://doi.org/10.1016/j.tbs.2018.05.003. June 2016.
Nguyen-Phuoc, D.Q., et al., 2018. 'Transit user reactions to major service withdrawal – a behavioural study'. Transport Pol. 64 (November 2017), 29–37. https://doi.org/10.1016/j.tranpol.2018.01.004. Elsevier Ltd.
Noland, R.B., et al., 2014. A mode choice analysis of school trips in New Jersey. Journal of Transport and Land Use 7 (2), 111. https://doi.org/10.5198/jtlu.v7i2.444.
Ortúzar, J. de D., Willumsen, L., 2011. Modelling Transport. John wiley & sons.
Owen, A., Levinson, D.M., 2015. Modeling the commute mode share of transit using continuous accessibility to jobs. Transport. Res. Pol. Pract. 74, 110–122. https://doi.org/10.1016/j.tra.2015.02.002. Elsevier Ltd.
Oyama, Y., Hato, E., 2018. 'Link-based measurement model to estimate route choice parameters in urban pedestrian networks'. Transport. Res. C Emerg. Technol. 93 (June), 62–78. https://doi.org/10.1016/j.trc.2018.05.013. Elsevier.
Paleti, R., et al., 2017. An integrated model of intensity of activity opportunities on supply side and tour destination & departure time choices on demand side. Journal of Choice Modelling 24 (October 2015), 63–74. https://doi.org/10.1016/j.jocm.2017.03.003. Elsevier Ltd.
Pang, H., Khani, A., 2018. 'Modeling park-and-ride location choice of heterogeneous commuters'. Transportation 45 (1), 71–87. https://doi.org/10.1007/s11116-016-9723-5. Springer US.
Papola, A., 2016. A new random utility model with flexible correlation pattern and closed-form covariance expression: the CoRUM. Transp. Res. Part B Methodol. 94, 80–96. https://doi.org/10.1016/j.trb.2016.09.008. Elsevier Ltd.
Park, S., Kang, J., Choi, K., 2014. 'Finding determinants of transit users' walking and biking access trips to the station: a pilot case study'. KSCE Journal of Civil Engineering 18 (2), 651–658. https://doi.org/10.1007/s12205-014-0073-6.
Park, S., Choi, K., Lee, J.S., 2015. 'To walk or not to walk: testing the effect of path walkability on transit users' access mode choices to the station'. International Journal of Sustainable Transportation 9 (8), 529–541. https://doi.org/10.1080/15568318.2013.825036, 2015.
Paulssen, M., et al., 2014. Values, attitudes and travel behavior: a hierarchical latent variable mixed logit model of travel mode choice. Transportation 41 (4), 873–888. https://doi.org/10.1007/s11116-013-9504-3.
Picard, N., Dantan, S., de Palma, A., 2018. Mobility decisions within couples. Theor. Decis. 84 (2), 149–180. https://doi.org/10.1007/s11238-017-9648-2. Springer US.
Pike, S., Lubell, M., 2018. The conditional effects of social influence in transportation mode choice. Res. Transport. Econ. 68, 2–10. https://doi.org/10.1016/j.retrec.2018.05.010. June.
Pnevmatikou, A.M., Karlaftis, M.G., Kepaptsoglou, K., 2015. 'Metro service disruptions: how do people choose to travel?'. Transportation 42 (6), 933–949. https://doi.org/10.1007/s11116-015-9656-4. Springer US.
Popovich, N.D., Handy, S., 2015. Downtown, strip centers, and big-box stores: mode choice by shopping destination type in Davis, California. Journal of Transport and Land Use 8 (2), 149–170. https://doi.org/10.5198/jtlu.2015.739.
Prato, C.G., 2014. Expanding the applicability of random regret minimization for route choice analysis. Transportation 41 (2), 351–375. https://doi.org/10.1007/s11116-013-9489-y.
Prato, C.G., Halldórsdóttir, K., Nielsen, O.A., 2017. Latent lifestyle and mode choice decisions when travelling short distances. Transportation 44 (6), 1343–1363. https://doi.org/10.1007/s11116-016-9703-9.
Prato, C.G., Halldórsdóttir, K., Nielsen, O.A., 2018. 'Evaluation of land-use and transport network effects on cyclists' route choices in the Copenhagen Region in value-of-distance space'. International Journal of Sustainable Transportation 12 (10), 770–781. https://doi.org/10.1080/15568318.2018.1437236. Taylor & Francis.
Rahman, M.L., Baker, D., 2018. 'Modelling induced mode switch behaviour in Bangladesh: a multinomial logistic regression approach'. Transport Pol. 71 (June 2017), 81–91. https://doi.org/10.1016/j.tranpol.2018.09.006. Elsevier Ltd.
Rahul, T.M., Verma, A., 2018. 'Sustainability analysis of pedestrian and cycling infrastructure – a case study for Bangalore'. Case Studies on Transport Policy 6 (4), 483–493. https://doi.org/10.1016/j.cstp.2018.06.001. Elsevier.
Ramezani, S., Pizzo, B., Deakin, E., 2018a. 'An integrated assessment of factors affecting modal choice: towards a better understanding of the causal effects of built environment. Transportation 45 (5), 1351–1387. https://doi.org/10.1007/s11116-017-9767-1. Springer US.
Ramezani, S., Pizzo, B., Deakin, E., 2018b. Determinants of sustainable mode choice in different socio-cultural contexts: a comparison of Rome and San Francisco. International Journal of Sustainable Transportation 12 (9), 648–664. https://doi.org/10.1080/15568318.2017.1423137. Taylor & Francis.
Rashedi, Z., et al., 2017. 'On the factors affecting the choice of regional transit for commuting in Greater Toronto and Hamilton Area: application of an advanced RP-SP choice model'. Transport. Res. Pol. Pract. 105 (June), 1–13. https://doi.org/10.1016/j.tra.2017.08.008. Elsevier.
Raveau, S., et al., 2014. A behavioural comparison of route choice on metro networks: time, transfers, crowding, topology and socio-demographics. Transport. Res. Pol. Pract. 66 (1), 185–195. https://doi.org/10.1016/j.tra.2014.05.010.
Rubin, O., Mulder, C.H., Bertolini, L., 2014. The determinants of mode choice for family visits - evidence from Dutch panel data. J. Transport Geogr. 38, 137–147. https://doi.org/10.1016/j.jtrangeo.2014.06.004. Elsevier Ltd.
Rybarczyk, G., Gallagher, L., 2014. Measuring the potential for bicycling and walking at a metropolitan commuter university. J. Transport Geogr. 39, 1–10. https://doi.org/10.1016/j.jtrangeo.2014.06.009. Elsevier Ltd.
Sadhu, S.L.N.S., Tiwari, G., 2016. An activity pattern - destination land use choice model of low income households of informal settlements - case study of Delhi. Transport. Res. Pol. Pract. 85, 265–275. https://doi.org/10.1016/j.tra.2016.01.017. Elsevier Ltd.
Sanko, N., 2014. Travel demand forecasts improved by using cross-sectional data from multiple time points. Transportation 41 (4), 673–695. https://doi.org/10.1007/s11116-013-9464-7.

Sanko, N., 2016. 'Factors affecting temporal changes in mode choice model parameters'. Transport. Plann. Technol. 39 (7), 641–652. https://doi.org/10.1080/03081060.2016.1204088. Taylor & Francis.
Sanko, N., 2017. 'Temporal transferability: trade-off between data newness and the number of observations for forecasting travel demand'. Springer US Transportation 44 (6), 1403–1420. https://doi.org/10.1007/s11116-016-9707-5.
Sanko, N., 2018. 'Travel demand forecasts improved by using cross-sectional data from multiple time points: enhancing their quality by linkage to gross domestic product'. Springer US Transportation 45 (3), 905–918. https://doi.org/10.1007/s11116-016-9755-x.
Sarkar, P.P., Chunchu, M., 2016. Quantification and analysis of land-use effects on travel behavior in smaller Indian cities: case study of agartala. J. Urban Plann. Dev. 142 (4), 04016009 https://doi.org/10.1061/(asce)up.1943-5444.0000322.
Sarkar, P.P., Mallikarjuna, C., 2018. 'Effect of perception and attitudinal variables on mode choice behavior: a case study of Indian city, Agartala. Travel Behaviour and Society 12, 108–114. https://doi.org/10.1016/j.tbs.2017.04.003. Hong Kong Society for Transportation Studies.
Scheepers, C.E., et al., 2016. Perceived accessibility is an important factor in transport choice - results from the AVENUE project. Journal of Transport and Health 3 (1), 96–106. https://doi.org/10.1016/j.jth.2016.01.003. Elsevier.
Schoner, J.E., Cao, J., Levinson, D.M., 2015. Catalysts and magnets: built environment and bicycle commuting. J. Transport Geogr. 47, 100–108. https://doi.org/10.1016/j.jtrangeo.2015.07.007. Elsevier Ltd.
Sener, I.N., Reeder, P.R., 2014. 'An integrated analysis of workers' physically active activity and active travel choice behavior'. Transport. Res. Pol. Pract. 67, 381–393. https://doi.org/10.1016/j.tra.2014.03.007. Elsevier Ltd.
Shaheen, S.A., Chan, N.D., Gaynor, T., 2016. Casual carpooling in the san francisco bay area: understanding user characteristics, behaviors, and motivations. Transport Pol. 51, 165–173. https://doi.org/10.1016/j.tranpol.2016.01.003. Elsevier.
Shao, J., 1993. Linear model selection by cross-validation. Journal of the American Statistical Association ISSN 88 (422), 486–494. https://doi.org/10.2307/2290328.
Shao, C., et al., 2015. 'Logistic regression models for the nearest train station choice: a comparison of captive and non-captive stations', *Case Studies on Transport Policy*. World Conference on Transport Research Society 3 (4), 382–391. https://doi.org/10.1016/j.cstp.2015.06.002.
Sharmeen, F., Timmermans, H., 2014. Walking down the habitual lane: analyzing path dependence effects of mode choice for social trips. J. Transport Geogr. 39, 222–227. https://doi.org/10.1016/j.jtrangeo.2014.07.012. Elsevier Ltd.
Shen, Q., Chen, P., Pan, H., 2016. 'Factors Affecting Car Ownership and Mode Choice in Rail Transit-Supported Suburbs of a Large Chinese City', *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 31–44. https://doi.org/10.1016/j.tra.2016.08.027. Elsevier Ltd.
Shirgaokar, M., Nurul Habib, K., 2018. How does the inclination to bicycle sway the decision to ride in warm and winter seasons? International Journal of Sustainable Transportation 12 (6), 397–406. https://doi.org/10.1080/15568318.2017.1378779. Taylor & Francis.
Singh, N., Vasudevan, V., 2018. 'Understanding school trip mode choice – the case of Kanpur (India)'. J. Transport Geogr. 66 (January), 283–290. https://doi.org/10.1016/j.jtrangeo.2017.12.007. Elsevier.
Soltani, A., Shams, A., 2017. Analyzing the influence of neighborhood development pattern on modal choice. J. Adv. Transport. 2017, 1–11. https://doi.org/10.1155/2017/4060348.
Standen, C., et al., 2017. Determinants of mode and route change following the opening of a new cycleway in Sydney, Australia. Journal of Transport and Health 4, 255–266. https://doi.org/10.1016/j.jth.2016.10.004. Elsevier Ltd.
Steyerberg, E.W., et al., 2001. 'Internal validation of predictive models : efficiency of some procedures for logistic regression analysis'. J. Clin. Epidemiol. 54, 774–781.
Stone, M., 1974. Cross-Validatory Choice and Assessment of Statistical Predictions 36 (2), 111–147.
Stone, M., 1977. 'An asymptotic equivalence of choice of model by cross-validation and akaike's. Criterion 39 (1), 44–47.
Stone, M., et al., 2014. Predictors of driving among families living within 2km from school: exploring the role of the built environment. Transport Pol. 33, 8–16. https://doi.org/10.1016/j.tranpol.2014.02.001. Elsevier.
Suel, E., Polak, J.W., 2017. 'Development of joint models for channel, store, and travel mode choice: grocery shopping in London'. *Transport. Res. Part A: Policy and Practice*. The Authors 99, 147–162. https://doi.org/10.1016/j.tra.2017.03.009.
Sun, G., et al., 2018. Living in school catchment neighborhoods: perceived built environments and active commuting behaviors of children in China. Journal of Transport and Health 8 (January), 251–261. https://doi.org/10.1016/j.jth.2017.12.009. Elsevier Ltd.
Thigpen, C.G., Driller, B.K., Handy, S.L., 2015. Using a stages of change approach to explore opportunities for increasing bicycle commuting. Transport. Res. Transport Environ. 39, 44–55. https://doi.org/10.1016/j.trd.2015.05.005. Elsevier Ltd.
Thomas, T., Tutert, B., 2015. Route choice behavior in a radial structured urban network: do people choose the orbital or the route through the city center? J. Transport Geogr. 48, 85–95. https://doi.org/10.1016/j.jtrangeo.2015.08.013. Elsevier B.V.
Tilahun, N., et al., 2016. Transit use and the work commute: analyzing the role of last mile issues. J. Transport Geogr. 54, 359–368. https://doi.org/10.1016/j.jtrangeo.2016.06.021. Elsevier B.V.
Toşa, C., et al., 2018. Commuting behavior in emerging urban areas: findings of a revealed-preferences and stated-intentions survey in Cluj-Napoca, Romania. J. Transport Geogr. 68 (March), 78–93. https://doi.org/10.1016/j.jtrangeo.2018.02.011.
Train, K., 1978. A validation test of a disaggregate mode choice model. Transport. Res. 12 (3), 167–174. https://doi.org/10.1016/0041-1647(78)90120-X.
Train, K.E., 1979. A comparison of the predictive ability of mode choice models with various levels of complexity. Transport. Res. Gen. 13 (1), 11–16. https://doi.org/10.1016/0191-2607(79)90082-7.
Tran, M.T., et al., 2016. A joint analysis of residential location, work location and commuting mode choices in Hanoi, Vietnam. J. Transport Geogr. 54, 181–193. https://doi.org/10.1016/j.jtrangeo.2016.06.003. Elsevier B.V.
Troncoso Parady, G., Takami, K., Harata, N., 2017. Built environment and travel behavior: validation and application of a continuous-treatment propensity score stratification method. Journal of Transport and Land Use 10 (1). https://doi.org/10.5198/jtlu.2017.890.
Venigalla, M., Faghri, A., 2015. A quick-response discrete transit-share model for transit-oriented developments. Journal of Public Transportation 18 (3), 107–123. https://doi.org/10.5038/2375-0901.18.3.7.
Verma, A., Rahul, T.M., Dixit, M., 2015. 'Sustainability impact assessment of transportation policies - a case study for Bangalore city', *Case Studies on Transport Policy*. World Conference on Transport Research Society 3 (3), 321–330. https://doi.org/10.1016/j.cstp.2014.06.001.
Vij, A., Krueger, R., 2017. Random taste heterogeneity in discrete choice models: flexible nonparametric finite mixture distributions. Transp. Res. Part B Methodol. 106, 76–101. https://doi.org/10.1016/j.trb.2017.10.013. Elsevier Ltd.
Vij, A., Walker, J.L., 2014. Preference endogeneity in discrete choice models. Transp. Res. Part B Methodol. 64, 90–105. https://doi.org/10.1016/j.trb.2014.02.008. Elsevier Ltd.
Vij, A., Gorripaty, S., Walker, J.L., 2017. 'From trend spotting to trend 'splaining: understanding modal preference shifts in the San Francisco Bay Area. Transport. Res. Pol. Pract. 95, 238–258. https://doi.org/10.1016/j.tra.2016.11.014. Elsevier Ltd.
von Elm, E., et al., 2007. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 4 (10), 1623–1627. https://doi.org/10.18601/01229893.n38.03.
Wang, Y., et al., 2014. Modeling traveler mode choice behavior of a new high-speed rail corridor in China. Transport. Plann. Technol. 37 (5), 466–483. https://doi.org/10.1080/03081060.2014.912420. Taylor & Francis.
Wang, C.H., Akar, G., Guldmann, J.M., 2015. Do your neighbors affect your bicycling choice? A spatial probit model for bicycling to The Ohio State University. J. Transport Geogr. 42, 122–130. https://doi.org/10.1016/j.jtrangeo.2014.12.003. Elsevier Ltd.
Wang, Y., et al., 2017. Using metro smart card data to model location choice of after-work activities: an application to Shanghai. J. Transport Geogr. 63 (January), 40–47. https://doi.org/10.1016/j.jtrangeo.2017.06.010. Elsevier.
Weiss, A., Habib, K.N., 2018. A generalized parallel constrained choice model for intra-household escort decision of high school students. Transp. Res. Part B Methodol. 114, 26–38. https://doi.org/10.1016/j.trb.2018.05.008. Elsevier Ltd.

Weng, J., et al., 2018. Modeling mode choice behaviors for public transport commuters in beijing. J. Urban Plann. Dev. 144 (3), 05018013 https://doi.org/10.1061/(asce)up.1943-5444.0000459.

Willmott, C., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82. https://doi.org/10.3354/cr00799.

Xie, S., Wong, S.C., Lam, W., 2016. Microscopic decision model for pedestrian route choice at signalized crosswalks. J. Adv. Transport. 50, 1181–1192. https://doi.org/10.1002/atr.

Xiqun, (Michael) Chen, et al., 2015. Investigations of interactions between bus rapid transit and general traffic flows. J. Adv. Transport. 49 (1), 326–340. https://doi.org/10.1002/atr.

Yamamoto, T., Takamura, S., Morikawa, T., 2018. 'Structured random walk parameter for heterogeneity in trip distance on modeling pedestrian route choice behavior at downtown area. Elsevier Travel Behaviour and Society 11, 93–100. https://doi.org/10.1016/j.tbs.2018.02.006. June 2017.

Yang, L., 2018. Modeling the mobility choices of older people in a transit-oriented city: policy insights'. Habitat Int. 76 (February), 10–18. https://doi.org/10.1016/j.habitatint.2018.05.007. Elsevier.

Yang, L., et al., 2014. Choice of commuting mode among employees: do home neighborhood environment, worksite neighborhood environment, and worksite policy and supports matter? Journal of Transport and Health 2 (2), 212–218. https://doi.org/10.1016/j.jth.2015.02.003. Elsevier.

Yang, L., Shen, Q., Li, Z., 2016. 'Comparing travel mode and trip chain choices between holidays and weekdays. Transport. Res. Pol. Pract. 91, 273–285. https://doi.org/10.1016/j.tra.2016.07.001. Elsevier Ltd.

Yang, Y., et al., 2016. Modeling the charging and route choice behavior of BEV drivers. Transport. Res. C Emerg. Technol. 65, 190–204. https://doi.org/10.1016/j.trc.2015.09.008. Elsevier Ltd.

Yang, L., et al., 2017. 'Urban form and travel patterns by commuters: comparative case study of wuhan and xi'an, China'. J. Urban Plann. Dev. 144 (1), 05017014 https://doi.org/10.1061/(asce)up.1943-5444.0000417.

Yang, Y., et al., 2018. Understanding the determinants of travel mode choice of residents and its carbon mitigation potential. Energy Pol. 115 (January), 486–493. https://doi.org/10.1016/j.enpol.2018.01.033. Elsevier Ltd.

Yazdanpanah, M., Hadji Hosseinlou, M., 2017. The impact of perception toward the built environment in airport access mode choice using hybrid choice modeling. J. Adv. Transport. 2017, 1–13. https://doi.org/10.1155/2017/8268701.

Yen, B.T.H., et al., 2018a. Assessing interchange effects in public transport: a case study of South East Queensland, Australia. Case Studies on Transport Policy 6 (3), 364–375. https://doi.org/10.1016/j.cstp.2018.01.005. Elsevier.

Yen, B.T.H., Mulley, C., Tseng, W.C., 2018b. Inter-modal competition in an urbanised area: heavy rail and busways. Res. Transport. Econ. 69 (September 2017), 77–85. https://doi.org/10.1016/j.retrec.2018.04.007. Elsevier.

Zhang, G., et al., 2014. Enhanced traffic information dissemination to facilitate toll road utilization: a nested logit model of a stated preference survey in Texas. Transportation 41 (2), 231–249. https://doi.org/10.1007/s11116-013-9449-6.

Zhang, N., Zhang, Y., Zhang, X., 2015. Pedestrian choices of vertical walking facilities inside urban rail transit stations. KSCE Journal of Civil Engineering 19 (3), 742–748. https://doi.org/10.1007/s12205-012-0331-4.

Zhang, R., Yao, E., Liu, Z., 2017. School travel mode choice in Beijing, China. J. Transport Geogr. 62 (June), 98–110. https://doi.org/10.1016/j.jtrangeo.2017.06.001. Elsevier.

Zhang, L., et al., 2018. Modeling lane-changing behavior in freeway off-ramp areas from the shanghai naturalistic driving study. J. Adv. Transport. 2018, 1–10. https://doi.org/10.1155/2018/8645709.

Zhang, Y., et al., 2018. 'Estimating metro passengers' path choices by combining self-reported revealed preference and smart card data'. Transport. Res. C Emerg. Technol. 92 (October 2017), 76–89. https://doi.org/10.1016/j.trc.2018.04.019. Elsevier.

Zhao, P., Li, S., 2017. Bicycle-metro integration in a growing city: the determinants of cycling as a transfer mode in metro station areas in Beijing. Transport. Res. Pol. Pract. 99, 46–60. https://doi.org/10.1016/j.tra.2017.03.003. Elsevier Ltd.

Zhuang, Y., et al., 2017. 'Passengers' behavioral intentions towards congestion: observational study of the entry restrictions at traffic bottleneck'. KSCE Journal of Civil Engineering 21 (6), 2393–2402. https://doi.org/10.1007/s12205-016-1688-6.

Ziliak, S., McCloskey, D., 2007. The Cult of Statistical Significance. How the Standard Error Cost Us Jobs, Justice and Lives. The University of Michigan Press.

Zimmermann, M., Mai, T., Frejinger, E., 2017. Bike route choice modeling using GPS data without choice sets of paths. Transport. Res. C Emerg. Technol. 75, 183–196. https://doi.org/10.1016/j.trc.2016.12.009.

Zimmermann, M., et al., 2018. Capturing correlation with a mixed recursive logit model for activity-travel scheduling'. Transport. Res. C Emerg. Technol. 93 (June), 273–291. https://doi.org/10.1016/j.trc.2018.05.032. Elsevier.

Zolnik, E.J., 2015. The effect of gasoline prices on ridesharing. J. Transport Geogr. 47, 47–58. https://doi.org/10.1016/j.jtrangeo.2015.07.009. Elsevier Ltd.