Willingness to pay quality estimates in commute mode choice: Model performance comparison under sample size and balance impacts.

Nikita Gusarov *
GAEL
Univ. Grenoble Alpes, CNRS, INRAE, Grenoble INP
38000 Grenoble, France
nikita.gusarov@univ-grenoble-alpes.fr

Iragaël Joly
GAEL
Univ. Grenoble Alpes, CNRS, INRAE, Grenoble INP
38000 Grenoble, France
iragael.joly@grenoble-inp.fr

Pierre Lemaire
G-SCOP
Univ. Grenoble Alpes, CNRS, Grenoble INP
38000 Grenoble, France
pierre.lemaire@grenoble-inp.fr

Abstract

In economics studies one of the wide-spread target metrics is the *Willingness to Pay* (WTP) of individuals for particular attributes of transportation mode choices. There already exists a wast literature addressing some major issues of the WTP elicitation task. We propose a performance comparison framework, allowing to systematize the previous research. With its help, in this work we explore models perform in WTP elicitation task under potential misspecifications, sample size and dataset balance changes. The `swissmetro` dataset is used for application purposes. We use simulation to vary sample size and configuration, which are used for model estimation and WTP elicitation. The results illustrate the variability in WTP estimates under different configurations.

***K*eywords** Transportation · Mode Choice · Willingness to Pay · Model Performance Comparison · Discrete Choice Modelling · Econometrics

1 Research question

In economics studies one of the wide-spread target metrics is the *Willingness to Pay* (WTP) of individuals for particular attributes of goods or services. In transportation studies popular manifestation of WTP are *Value of Time* (VOT) or *Value of Comfort* (VOC). The WTP elicitation lies at the heart of various tasks in the transportation mode choice analysis: adoption of sustainable transportation modes (Ilahi et al., 2021), perception of the resilient shared transportation modes (Ardeshiri et al., 2021), consumer preferences for delivery services (Merkert et al., 2022), attitudes towards trip attributes (Boto-García et al., 2022).

There exist multiple ways to deduce WTP from the data, most of which rely on the *Random Utility Maximisation* (RUM) framework (McFadden, 1974). The obtained results are affected not only by the selected methodology, but by the modelling strategy as well. With time the number of available models and estimation techniques increases, many of them remaining *RUM-compliant*. Following McFadden (1981) the RUM-compliance translates in the independence of the ranking of the choice probabilities of the alternatives by any monotonically increasing transformation of the utility functions of all elemental alternatives. One can also observe a growing number of papers focusing on interpretable *Machine Learning* (ML) techniques in

*Corresponding author.

application to choice modelling analysis (Aboutaleb et al., 2021; Han et al., 2022; Wang et al., 2020), some of which address the WTP elicitation (Bergtold and Ramsey, 2015; Wang et al., 2020). The multitude of available models and techniques makes it sometimes difficult for the researcher to select the best modelling approach for the particular situation.

There exists several studies in the literature exploring model performances in general (Jong et al., 2019; Zeng et al., 2018). The majority of researchers focus on the predictive accuracy as the main performance metrics for their sample size requirements calculation. However, according to the interdisciplinary works (Japkowicz and Shah, 2011) the performance of competing models may be assessed over several criteria: (1) quality of data adjustments; (2) predictive capacity; (3) quality of the field specific (ex: economic and behavioural) indicators derived from estimates; and (4) algorithmic efficiency and computational costs. We attempt to complete previous findings with a more extended view on the derived metrics, WTP in particular. **How the various models perform in WTP elicitation task under potential misspecifications? Does the sample size and class balance impact the WTP estimates in various *RUM-compliant* models?**

2 Methodology and context

Many of the listed above studies focusing on the WTP metrics rely on *Stated Preference* (SP) data. However, while setting up a DCE little is known about the exact behaviour within the target population. The researchers typically rely on the previous studies in selecting the most plausible theoretical assumptions while conducting a DCE, but there are always some limitations. One of the important elements in the WTP elicitation tasks is tied to the model requirements in terms of sample size and overall dataset configuration. This pushes us to explore empirically the potential consequences of inadequate model usage under changes in data.

Some may say that the research questions were already addressed in the literature and they will not be wrong. There is a number of studies, which in one way or another proposed some insight into the data requirements for particular model families, or explored the data quality impacts on the estimates.

Among the reference works we may encounter, a revision of WTP elicitation approaches performed by Daly et al. (2022). Or a criticised estimation of WTP under utility specification restrictions of Carson and Czajkowski (2019). Paper of Bazzani et al. (2018) addressing the usage of flexible mixing distributions in WTP space. As well as a rather complete comparison of confidence intervals measures for WTP under sample size changes published by Hole (2007). Among the data focused studies we encounter the mitigation of class balance effects for NL models by Bierlaire et al. (2008). The study if impacts of sample size, attribute variance and choice distribution on the accuracy in the paper of Zeng et al. (2018). An extensive analysis of ample size requirements for stated choice experiments of Rose and Bliemer (2013).

All of the above works are relatively close to the research questions we have outlined in the introduction. However, as most of the research is focused on the theoretical fundamentals with scarce empirical illustrations, we attempt to complement the existing literature with a more accessible evidence. For this purpose we propose a theoretical performance comparison framework, which should simplify the empirical theory testing procedure.

In this section we are offering a short focus on the WTP elicitation approach, which will be used further on. Then we outline the proposed performance comparison framework that will guide our data-driven study.

2.1 Willingness to Pay

For the purposes of this study we use the simplest WTP definition. We assume that individual deterministic utility of an alternative j (from a set of available alternatives Ω) is given as function with parameters $\beta_{\mathbf{j}}$: $V_j = f(\beta_{\mathbf{j}})$. The simplest option is then to provide the point analytical estimates of the WTP values, which is justified if V_j is linear in attributes. The total variation of V_j with respect to joint variations in the k-th attribute $x_{k,j}$ and the cost attribute $x_{cost,j}$ is $\Delta V_j = \Delta x_{k,j} + \Delta x_{cost,j}$. Resolving this equation for the case of $\Delta V_j = 0$ we obtain the change in cost, which keeps the deterministic utility unchanged given a change in k-th attribute:

$$WTP_{k,j} = \frac{\Delta V_j / \Delta x_{k,j}}{\Delta V_j / \Delta x_{cost,j}}$$

The easiest option focuses on confidence interval calculation for WTP values using the less resource heavy **Delta method** (Daly et al., 2022), which avoids simulation step (Scaccia et al., 2023). Such method is

usually used to calculate the standard error for a function of the parameter estimates. For simplicity in this study we do not use any alternative WTP confidence intervals identification strategies. This methods add some more prerequisites, as we should assume that WTP_k is given as $\omega_k = h(\beta_k, \beta_{cost}) = \frac{\beta_k}{\beta_{cost}}$ is a differentiable function. The formula for the standard error of ω_k is hence (Daly et al., 2012) is:

$$\sigma_{\hat{\omega_k}} = \frac{1}{\beta_{cost}} \sqrt{\sigma^2_{\hat{\beta_k}} + 2\omega_k \sigma_{\hat{\beta_k}\hat{\beta_{cost}}} + \omega_k^2 \sigma^2_{\hat{\beta_{cost}}}}$$

2.2 Performance comparison framework

To address the model misspecification and data imbalance issues under a new angle we propose a performance comparison framework. It incorporates all essential steps from the research question definition to the performance comparison in relation to the given context. This framework is based on the concepts described by Williams and Ortuzar (1982), revised and extended.

We believe that the most rational way to construct such framework is to mimic in its structure the traditional scientific **research procedure**. In the literature, regardless of the actual case, all the research takes its root in some problematic: a question to be answered, a barrier to be overcome. Once the task delimited, there are different strategies on how to proceed. Some of them are conventional and described in every practical guide (Baltagi, 2008; Wooldridge, 2012), while other are more obscure and are sometimes criticized for uncommon practices (Daly et al., 2022). As one can see, those topics we'll rise here are mainly discussed in the epistemology works, rather than in more abundant applied studies. Nevertheless, it's extremely important to have the general understanding of the typical procedures and paths implemented in applied research to make the next leap towards framework construction.

The procedure may be in general divided into several major steps (Figure 1). First of all, every research starts with a *problematic* identification and *operational* or *economic question* definition[2]. Every study begins with a particular need - *operational problematic* to be addressed. The first step reflect the transition of the real world problem to be treated into the more restricted context of a research specific question. The next stage in the research requires the researcher to make some assumptions about the nature of the data and the underlying processes. Typically it's during this stage that hypothetical interaction model is defined based on the theoretical assumptions or the preliminary analysis of the available (if available) data. Thus the second step is a further extension of the *problematic* narrowing and translation into numerical terms: target *metrics* identification. Those *metrics* should allow the researcher to answer to the research question. For example, one may be interested in causality exploration, which may be translated into the analysis of particular coefficient significance in an econometric model. Another example is the prediction task: researchers may be interested to offer the best prediction of consumer behaviour (ex: to identify the market shares), which may be translated into comparison of various performance metrics for different predictive models. Once the target defined, the research may proceed differently, depending on the available information. Without loss of generality this step may be summarized as *data collection and analysis* process. Either the actors already have access to some data and build the model using available information. Or the model is prebuilt and drives the data collection step. Finally, the data analysis provides the actor with information on the target *metrics* (*estimates*). Those allow to answer the initial question and offer a *solution* to the initial problematic.

The performance of a model can scarcely be assessed without any particular context. In fact we redefine the **performance** as **the model's capacity to bring a correct answer in the context of explored problematic**. This performance term redefinition brings us to the necessity to step aside from the typical *model performance comparison* and switches our focus to another conceptual unit - *the procedure*. By *the procedure* we understand the entire process starting with the research question definition to the answer to this question. This means that the procedure in this case includes such steps as data collection, processing and analysis. This also includes all the eventual (be it arbitrary or not) choice in terms of model configuration, selection and fine-tuning.

Consequently, the framework should be inevitably dependent on the research question: some models are simply not capable to answer some questions or there are no known or established practices of their usage. The definition of the research question should therefore be considered as the first step in the proposed

[2]Here we avoid speaking about *research question*, as sometimes it may not be directly linked with the *economic question* treated in the study. Moreover, the *question* may be purely *operational*, without production of any particular new knowledge and be purely context specific for particular application.

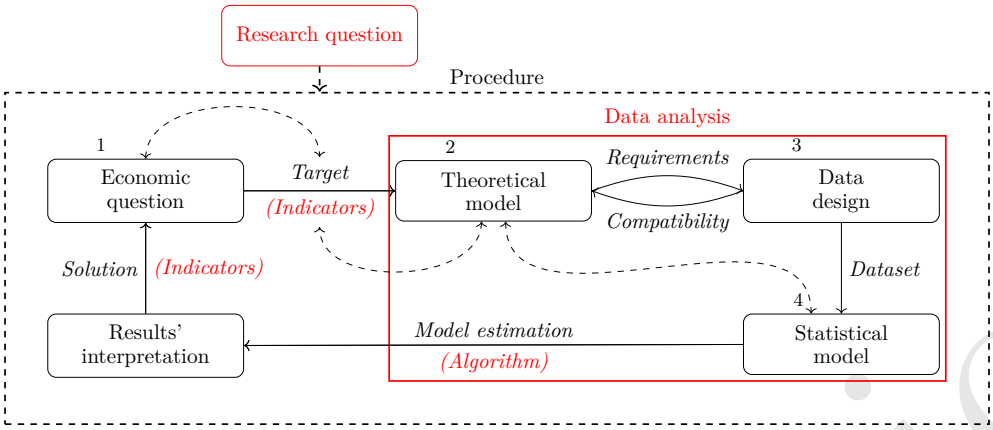

Figure 1: Proposed performance comparison framework

framework. It will provide the researcher with particular metrics to consider while performing the model comparison.

The second stage in the framework should be left for the dataset choice or dataset generation procedure. This includes all the potential assumptions and *a priori* choices on the assumed individual behaviour within population, external effects and potential biases. At this point the opinions may vary as discussed by Japkowicz and Shah (2011). Even though in statistical modelling when speaking about model performance assessment and comparison the focus is typically made on the classification (prediction) accuracy (Andersson et al., 1999; Askin and Gokalp, 2013; Hand, 2012) this is not always the best option. On the one hand, in model comparison, whatever is the research question, one will always have some target metrics or criteria in mind. It means, that for a complete comparison procedure one should be able to compare not only the models between themselves, but compare the results with some externally defined target as well.

The next stage is represented by the *modelling procedure* itself. This includes the choice of the model and its implementation, the configuration of the estimated utility functions, etc. Later it will be equally subject to the numerical specificity: the choice of the estimation algorithm and its implementation, the particular code base and approach to the problem solving.

Finally, comes the *post-treatment* of the obtained estimates. In our particular case this step involves the WTP calculation, assuming the model was not estimated directly in the WTP space. All the essential indicators obtained on this step should be evaluated in the context of the research question and, if possible, compared to the target values used as inputs for the simulation task. Now, once the framework is fully described we can proceed with the application.

3 Application

For illustration purposes we use the rather popular and publicly available dataset `swissmetro`. The dataset firstly appeared in the paper of Bierlaire et al. (2001), where it was used to asses the acceptance of the state proposed modal innovation (Nash et al., 2007). A more in-depth description of the dataset, as well as the dataset itself are available on the `biogeme` project website. This data was used in many illustration of newly created model capabilities, as well as in several model performance comparison tasks. The most closely related works to our usecase are: Bierlaire et al. (2001), Bierlaire et al. (2008) and Newman et al. (2013).

We rely on preceding works to construct artificial datasets of different sample sizes and configurations. The conventional *Nested Logit* (NL) structure is imposed, which reflects quite common in reality decision rule structure. Several models are then estimated over the resulting datasets. The tests for WTP estimates validity are then performed, through a comparison with expected target results, as well as their overall significance.

3.1 Dataset description

The original dataset (as presented by Bierlaire et al. (2001)) is based on a combination of the *revealed preferences* (RP) and *stated preferences* (SP) data collected in Switzerland, during March 1998. At the first stage, the study relied on collection of the initial information (observation) of the trip performed by subject.

This step was followed by a SP data collection step where they were proposed a novel *hypothetical* alternative: the `swissmetro`. To ensure that the new *hypothetical* transportation mode was pertinent for the subjects the sampling was performed through approaching the subjects while they travelled on the target routes. 470 observations (435 suitable ones) were collected in the train between St. Gallen and Geneva. Another 770 usable SP surveys were collected among the car user, this part being performed by mail with the support of central Swiss car licence agency. In the SP part of the study authors used *fractional factorial design* offering the following set of alternatives: (1) rail (`TRAIN`), (2) swissmetro (`SM`) and (3) car (`CAR`, only for car owners). All the alternatives were designed by *travel time*, *fare/cost* and *headway* (for rail based alternatives only).

For this study we adopt the approach described by Bierlaire et al. (2008) and later used by Newman et al. (2013). The original dataset will be used for simulation purposes, which allows us to observe the model performances in a more controlled environment. Prior to simulation the dataset is filtered, excluding the observations for which there is no choice made and limiting our attention to the commute and business purpose trips. The descriptive statistics for the resulting dataset are presented in the table 1 (only reused explicative variables are shown).

Table 1: Descriptive statistics

Variable	N	Mean	St. Dev.	Min	Max
Cost					
TRAIN_CO	6,768	490.885	1,062.594	9	5,040
CAR_CO	6,768	78.656	55.922	0	520
SM_CO	6,768	641.066	1,411.658	11	6,720
Travel time					
TRAIN_TT	6,768	166.077	69.796	35	1,022
CAR_TT	6,768	123.155	91.718	0	1,560
SM_TT	6,768	84.507	47.113	12	796

3.2 Simulation

We proceed with a simulated dataset, which is based on the original one. The simulation approach adopted is identical to the one performed by Bierlaire et al. (2008). Each observation is replicated 100 times to provide us with synthetic observations. The alternative attributes values were overwritten by draws from normal distribution $N(\lambda, \sigma^2)$, where λ is the value of the corresponding attribute in the original dataset, and $\sigma = 0.05\lambda$ (Bierlaire et al., 2008).

Speaking about the decision rules, we decide to adopt the identical nested logit structure as in the other studies (Bierlaire et al., 2008; Bierlaire et al., 2001). The choice model specification is given in the Table 2.

Table 2: Utility specification

Utility	Value	TRAIN	SM	CAR
Parameter				
ASC_{CAR}	-0.1880	0	0	1
ASC_{SM}	0.1470	0	1	0
β_{TRAIN_TIME}	-0.0107	TT	0	0
β_{SM_TIME}	-0.0081	0	TT	0
β_{CAR_TIME}	-0.0071	0	0	TT
β_{COST}	-0.0083	COST	COST	COST
Nests				
$\lambda_{EXISTING}$	0.4405	1	0	1
λ_{FUTURE}	1.0000	0	1	0

Nesting structure was introduced through error components following the specification provided by Bierlaire et al. (2008)[3]. This structure assumed that alternatives can be separated according to their real availability.

[3]For this purpose we used the `evd::rmvevd()` function in `R`

Meaning that while error components behave identically for existing transportation modes (car and train), the effects may differ for non-existing (*future*) alternative.

The WTP (VOT in this particular case) true values can be calculated as $\omega_k = \frac{\beta_k}{\beta_{cost}}$ (ex. for `TRAIN` alternative we would calculate $WTP_{TRAIN\_TIME} = \frac{\beta_{TRAIN\_TIME}}{\beta_{COST}}$). This computation is justified because we assume, in our simulation, that effects are fixed within population. This gives us the values as presented in Table 3.

Table 3: True WTP (VOT) values

WTP_{CAR_TIME}	WTP_{SM_TIME}	WTP_{TRAIN_TIME}
0.8554217	0.9759036	1.289157

The final step includes drawing random observations from the resulting database to compose individual datasets of desired size and class-distribution. We vary the sample size from 500 observations, a number quite often encountered in econometric studies, to 10000 observations[4], which approaches the frontier of the datasets available for some very simple ML tasks. The different configurations are tested for all possible combinations of classes with a step of 0.2[5], as well as the perfectly balanced class distribution with equally distributed observations. For each pair of sample size and configuration parameters we randomly draw 50 datasets and estimate selected model over them.

This approach to simulation allows us not only to obtain a consistent baseline for performance assessment, but also the possibility to compare our results with similar papers, where identical simulation strategy was implemented.

3.3 Estimation

For the purposes of this study we implement three closely related econometric models, which might be potentially used by novices in choice modelling. Among them: (1) the optimal NL model, (2) the misspecified Multinomial Logit (MNL) model and (3) the Mixed MNL (MMNL) model, which still allows to capture non-uniform error structure. For easier results interpretation we use scaling during the estimation step for all the models.

The NL model follows the specification used during the simulation step and is expected to perform the best on the available data. The MNL model differs from it only by the absence of the nests (Table 4), meaning the nesting parameter α is omitted.

Table 4: Utility specification for MNL model

Utility	TRAIN	SM	CAR
Parameter			
ASC_{CAR}	0	0	1
ASC_{SM}	0	1	0
β_{TRAIN_TIME}	TT	0	0
β_{SM_TIME}	0	TT	0
β_{CAR_TIME}	0	0	TT
β_{COST}	COST	COST	COST

With a deterministic alternative specific utility given as:

$$V_j = ASC_j + \beta_{TIME,j} x_{TIME,j} + \beta_{COST} x_{COST,j}$$

Table 5: Utility specification for MMNL model

Utility	TRAIN	SM	CAR
Parameter			
ASC_{CAR}	0	0	1
ASC_{SM}	0	1	0
β_{TRAIN_TIME}	TT	0	0
β_{SM_TIME}	0	TT	0
β_{CAR_TIME}	0	0	TT
β_{COST}	COST	COST	COST
σ_v	1	0	1

With a deterministic alternative specific utility given as:

$$V_j = ASC_j + \beta_{TIME,j} x_{TIME,j} + \beta_{COST} x_{COST,j} + v_j(0, \sigma_v)$$

The MMNL model (Table 5) mimics the NL model structure, although it is a theoretically incorrect way to introduce nesting in the model as it was illustrated by Munizaga and Alvarez-Daziano (n.d.). We introduce

[4]Those values may vary by ±1 for the datasets with balanced shares.

[5]This results in a plain defined as `SHARE_TRAIN + SHARE_SM + SHARE_CAR = 1`.

the random term with zero mean and variance σ_v for the alternatives within a single nest. This allows to address the differences in errors variances, but also introduces a biased covariance structure to the estimated model.

3.4 WTP and model performance

As we have previously shown, in the literature there is no known consensus on the performance metrics and the "*model performance*" definition. As our study focuses on the WTP estimates, we assume that the objective of a model can be viewed as correct estimation of the target metrics. The WTP in its turn relies on the correct estimation of the effects within the model, assuming that the functional form is known and true.

Hence we are interested to observe the shares of estimation routines which manage to correctly identify the effects. Here, the term "*correctly estimate*" means a production of human readable results, which are not contradictory with real world (simulated in our case) scenario. To properly analyse this information, we are going to explore two different shares: (1) a share of models reporting estimates significantly different from 0, meaning that in the real world application the researcher would take the estimates into account; and (2) a share of models reporting estimates not significantly different from target values (the true values used for simulation of individual behaviour). One of the main advantages for this approach is that we can use basic t-test for hypothesis verification in each of the estimations and report the results in a convenient human readable form.

The same reasoning may be applied to the WTP estimates directly (Daly et al., 2022; Hole, 2007). For WTP estimates we set $\alpha/2$ to 0.125 for confidence interval specification, as in the work of Bierlaire et al. (2008). The WTP variance estimates are obtained using the *Delta* method, as suggested in the manuscript of Daly et al. (2022).

Performing similar test in over our simulated dataset estimates results in the following shares (Table 6). Here we observe the shares of models in dependence of the sample size. Each entry relies on $10 \times 50 = 500$ estimated models, mixing all available class balance configurations within sample. The WTP estimates are considered as appropriate if the desired condition (test) is fulfilled across all three alternatives, as facing three alternative mode choices makes us compute three distinct WTP values. The results presented in this part are for traditional estimation method, without any transformations (Carson and Czajkowski, 2019) nor transitions into the WTP space (Train and Weeks, 2005). We can observe that the number of estimates different from zero increases with sample size.

However, the same cannot be said about the shares of results not different from analytical targets (Table 7). Obviously, the simple *difference from zero* test is not the only one interesting for us. We might be interested with an additional test - the exploration of whether or not the obtained WTP estimates are significantly different from zero. Obviously, it's important that the estimator is unbiased, but from operational point of view it's equally important to obtain a meaningful result, which correctly reflects the reality. Which is extremely important in the context of potential strategic decision making based on the estimated values. We can see that those shares decrease with sample size, which has two potential explanations. Assuming the simulation procedure and random sampling has no apparent flaws, we may imply that such behaviour might be explained by the changes in class balance within the dataset.

Table 6: Shares of WTP estimates not different from target, by sample size.

Observations	MMNL	MNL	NL
500	46.80	42.60	64.52
1000	42.80	36.00	60.20
5000	17.80	13.20	40.00
10000	9.60	8.00	31.60

Table 7: Shares of WTP estimates different from zero, by sample size.

Observations	MMNL	MNL	NL
500	97.80	97.80	78.11
1000	99.80	99.80	89.20
5000	100.00	100.00	99.00
10000	100.00	100.00	99.60

Finally we explore the shares of WTP completing both of the above conditions, as presented in Table **??**. While WTP estimates non-distinguishable from zero may be discarded by researcher leading to non-concluding results, the biased estimates are not so easy to detect in the field.

Table 8: Shares of all correctly estimated WTP by sample size.

Observations	MMNL	MNL	NL
500	46.80	42.60	56.63
1000	42.80	36.00	57.20
5000	17.80	13.20	40.00
10000	9.60	8.00	31.60

A similar analysis can be applied to the results aggregater by class balance (Table 9). In this case each *shares combination* regroups $4 \times 50 = 200$ entries with all available class balances within sample.

Table 9: Shares of all correctly estimated WTP by dataset balance.

Share TRAIN	Share SM	Share CAR	MMNL	MNL	NL
0.10	0.10	0.80	34.50	31.00	53.50
0.10	0.80	0.10	30.50	26.50	51.50
0.20	0.20	0.60	33.50	29.50	61.50
0.20	0.40	0.40	29.50	21.50	63.00
0.20	0.60	0.20	24.50	18.50	47.00
0.33	0.33	0.33	26.00	24.00	54.50
0.40	0.20	0.40	30.00	28.50	53.50
0.40	0.40	0.20	30.50	20.50	34.00
0.60	0.20	0.20	27.00	28.00	25.76
0.80	0.10	0.10	26.50	21.50	19.19

4 Conclusion

In this paper we have empirically explored the effects of the model misspecification and changes in sample size and class balance within dataset on the WTP estimates. This study offers primarily a case dependent evidence, which is intended to raise the awareness of the perverse effects of the modelling strategy choice and data selection in empirical work.

In the particular application we have demonstrated that the increase of the sample size may is not always the best solution. In particular the attention should be paid to the modelling technique implemented and the reliability of the underlying assumptions. Those observations underline the problematic of model performance assessment and toolset selection in the empirical work.

Finally, but not less importantly, we have outlined the baseline of a model performance comparison framework, which can be extended to the other domains. The proposed toolset allows to efficiently contrast the performance impacts of the changes in the research procedure, which is invaluable for the empirical studies. Such toolset may allow to reduce studies' costs and time through prior experimentation.

5 References

Aboutaleb, Y.M., Danaf, M., Xie, Y., Ben-Akiva, M.E., 2021. Discrete Choice Analysis with Machine Learning Capabilities. Machine Learning 19.

Andersson, A., Davidsson, P., Lindén, J., 1999. Measure-based classifier performance evaluation. Pattern Recognition Letters 20, 1165–1173.

Ardeshiri, A., Safarighouzhdi, F., Rashidi, T.H., 2021. Measuring willingness to pay for shared parking. Transportation Research Part A: Policy and Practice 152, 186–202.

Askin, O.E., Gokalp, F., 2013. Comparing the Predictive and Classification Performances of Logistic Regression and Neural Networks: A Case Study on Timss 2011. Procedia - Social and Behavioral Sciences 106, 667–676.

Baltagi, B., 2008. Econometrics, 4th edition. Berlin: Springer.

Bazzani, C., Palma, M.A., Nayga, R.M., 2018. On the use of flexible mixing distributions in WTP space: An induced value choice experiment. Aust J Agric Resour Econ 62, 185–198.

Bergtold, J.S., Ramsey, S.M., 2015. Neural Network Estimators of Binary Choice Processes: Estimation, Marginal Effects and WTP (2015 AAEA \& WAEA Joint Annual Meeting, July 26-28, San Francisco, California No. 205649). Agricultural and Applied Economics Association.
Bierlaire, M., Axhausen, K., Abay, G., 2001. The acceptance of modal innovation: The case of Swissmetro 17.
Bierlaire, M., Bolduc, D., McFadden, D., 2008. The estimation of generalized extreme value models from choice-based samples. Transportation Research Part B: Methodological 42, 381–394.
Boto-García, D., Mariel, P., Pino, J.B., Alvarez, A., 2022. Tourists' willingness to pay for holiday trip characteristics: A Discrete Choice Experiment. Tourism Economics 28, 349–370.
Carson, R.T., Czajkowski, M., 2019. A new baseline model for estimating willingness to pay from discrete choice models. Journal of Environmental Economics and Management 95, 57–61.
Daly, A., Hess, S., Ortúzar, J.D.D., 2022. Estimating Willingness-to-Pay from Discrete Choice Models: Setting the Record Straight. SSRN Journal.
Daly, A., Hess, S., Train, K., 2012. Assuring finite moments for willingness to pay in random coefficient models. Transportation 39, 19–31.
Han, Y., Pereira, F.C., Ben-Akiva, M., Zegras, C., 2022. A neural-embedded discrete choice model: Learning taste representation with strengthened interpretability. Transportation Research Part B: Methodological 163, 166–186.
Hand, D.J., 2012. Assessing the Performance of Classification Methods. International Statistical Review 80, 400–414.
Hole, A.R., 2007. A comparison of approaches to estimating confidence intervals for willingness to pay measures. Health Econ. 16, 827–840.
Ilahi, A., Belgiawan, P.F., Balac, M., Axhausen, K.W., 2021. Understanding travel and mode choice with emerging modes; a pooled SP and RP model in Greater Jakarta, Indonesia. Transportation Research Part A: Policy and Practice 150, 398–422.
Japkowicz, N., Shah, M., 2011. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press.
Jong, V.M.T., Eijkemans, M.J.C., Calster, B., Timmerman, D., Moons, K.G.M., Steyerberg, E.W., Smeden, M., 2019. Sample size considerations and predictive performance of multinomial logistic prediction models. Statistics in Medicine 38, 1601–1619.
McFadden, D., 1981. Econometric models of probabilistic choice. Structural analysis of discrete data with econometric applications 198272.
McFadden, D., 1974. The measurement of urban travel demand. Journal of Public Economics 3, 303–328.
Merkert, R., Bliemer, M.C.J., Fayyaz, M., 2022. Consumer preferences for innovative and traditional last-mile parcel delivery. International Journal of Physical Distribution & Logistics Management 52, 261–284.
Munizaga, M.A., Alvarez-Daziano, R., n.d. MIXED LOGIT VS. NESTED LOGIT AND PROBIT MODELS.
Nash, A., Weidmann, U., Buchmueller, S., Rieder, M., 2007. Assessing Feasibility of Transport Megaprojects: Swissmetro European Market Study. Transportation Research Record 1995, 17–26.
Newman, J.P., Ferguson, M.E., Garrow, L.A., 2013. Estimating GEV models with censored data. Transportation Research Part B: Methodological 58, 170–184.
Rose, J.M., Bliemer, M.C.J., 2013. Sample size requirements for stated choice experiments. Transportation 40, 1021–1041.
Scaccia, L., Marcucci, E., Gatta, V., 2023. Prediction and confidence intervals of willingness-to-pay for mixed logit models. Transportation Research Part B: Methodological 167, 54–78.
Train, K., Weeks, M., 2005. Discrete Choice Models in Preference Space and Willingness-to-Pay Space, in: Scarpa, R., Alberini, A. (Eds.), Applications of Simulation Methods in Environmental and Resource Economics, The Economics of Non-Market Goods and Resources. Springer Netherlands, Dordrecht, pp. 1–16.
Wang, S., Wang, Q., Zhao, J., 2020. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. Transportation Research Part C: Emerging Technologies 118, 102701.
Williams, H.C.W.L., Ortuzar, J.D., 1982. Behavioural theories of dispersion and the mis-specification of travel demand models. Transportation Research Part B: Methodological 16, 167–219.
Wooldridge, J.M., 2012. Introductory Econometrics: A Modern Approach 910.
Zeng, M., Zhong, M., Hunt, J.D., 2018. Analysis of the Impact of Sample Size, Attribute Variance and Within-Sample Choice Distribution on the Estimation Accuracy of Multinomial Logit Models Using Simulated Data. J. Syst. Sci. Syst. Eng. 27, 771–789.

A Appendix - detailed results

A.1 Focus on WTP estimates by sample size

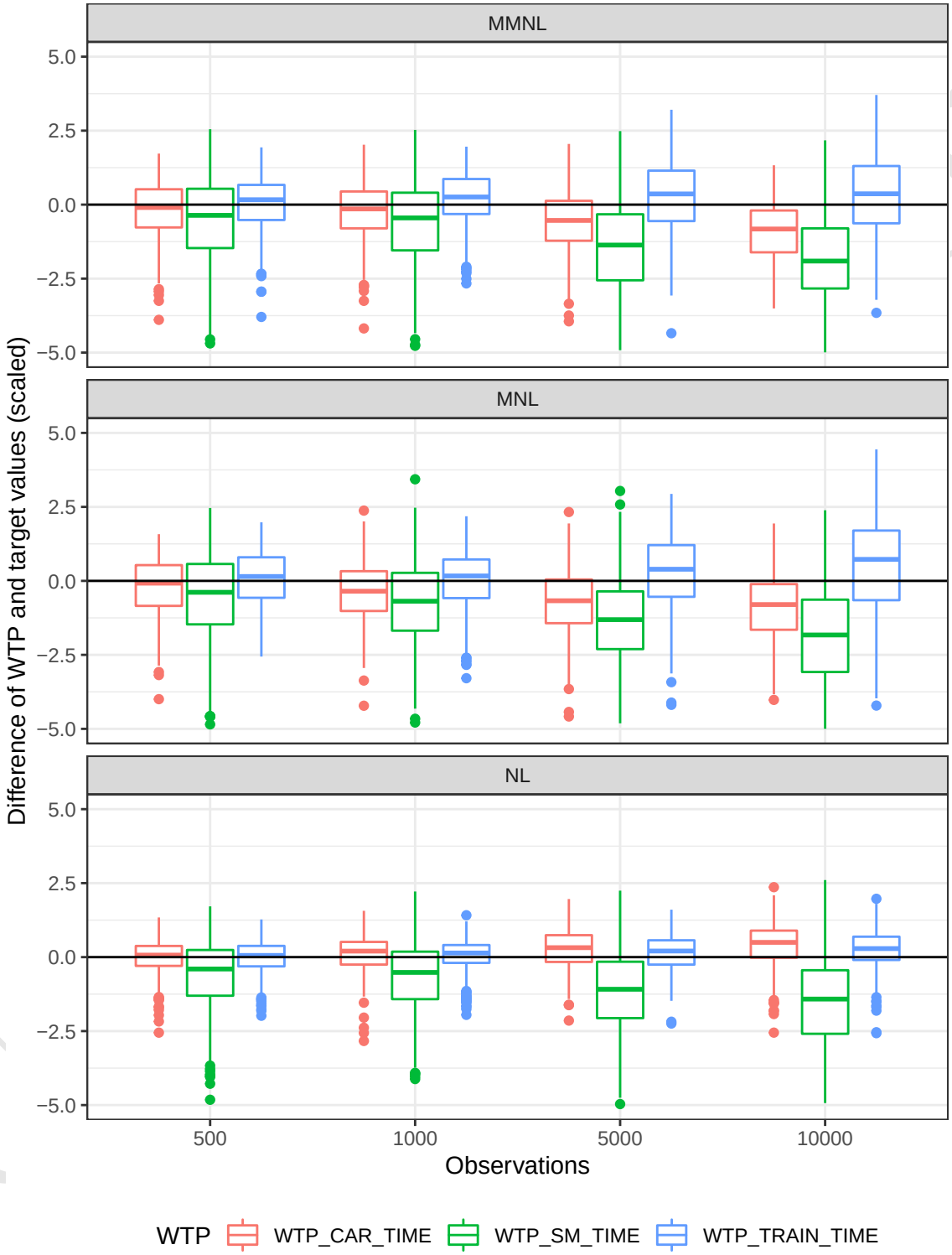

Figure A1: Difference of WTP and target values (scaled), by sample size.

Table A1: Shares of correctly estimated WTP by alternative, by sample size.

Model	Observations	WTP_{CAR_TIME}	WTP_{SM_TIME}	WTP_{TRAIN_TIME}	All
MMNL	500	82.20	59.00	79.20	46.80
MMNL	1000	81.20	60.00	75.60	42.80
MMNL	5000	69.40	39.60	62.80	17.80
MMNL	10000	61.40	28.20	58.80	9.60
MNL	500	78.40	59.20	75.00	42.60
MNL	1000	72.40	51.80	72.00	36.00
MNL	5000	63.60	39.40	58.00	13.20
MNL	10000	59.60	28.60	43.60	8.00
NL	500	87.58	60.48	91.38	56.85
NL	1000	93.20	60.40	95.20	57.20
NL	5000	91.20	47.00	94.60	40.00
NL	10000	83.20	39.60	89.00	31.60

A.2 Focus on WTP estimates by class balance

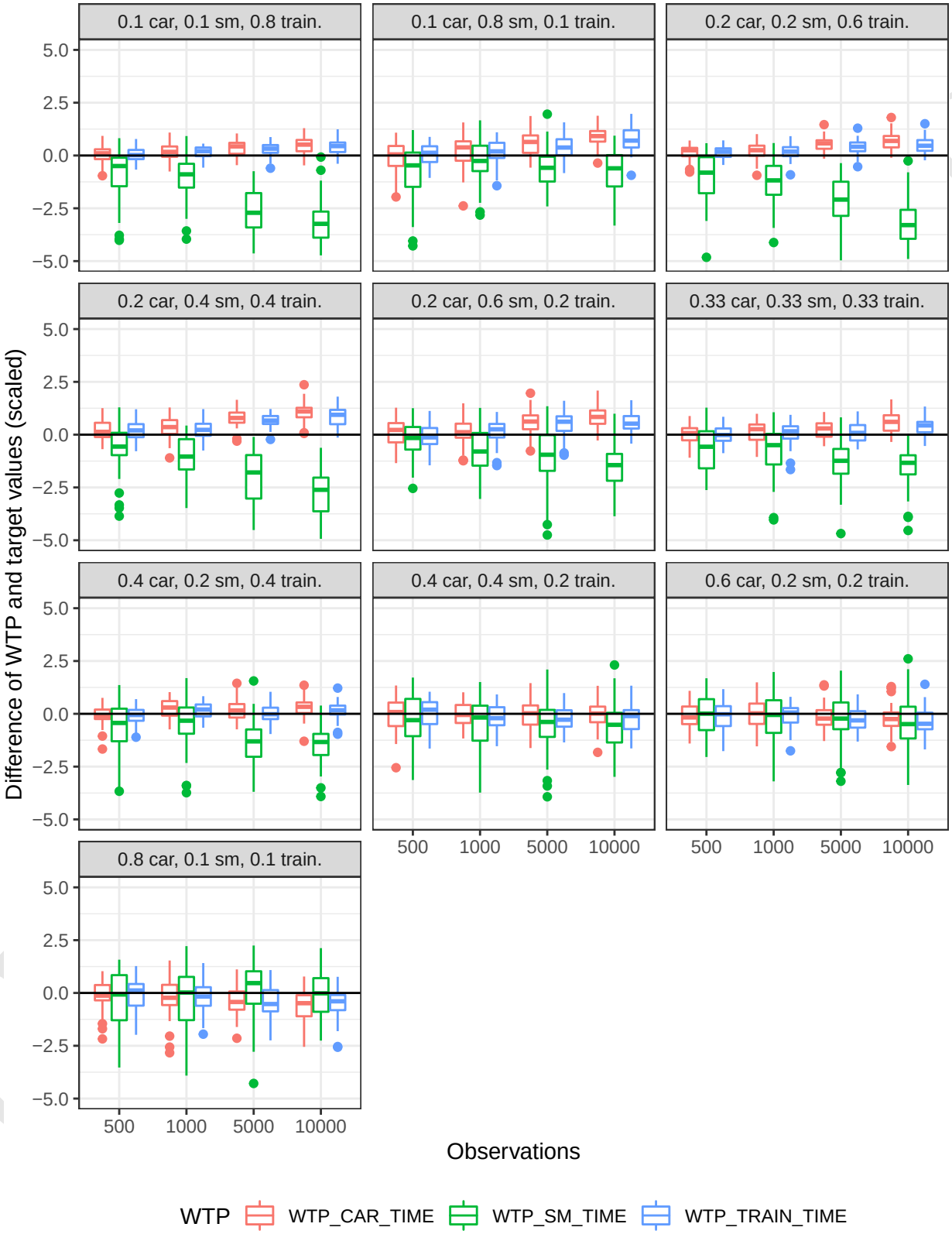

Figure A2: Difference of WTP and target values (scaled), by sample size and balance, NL model only.

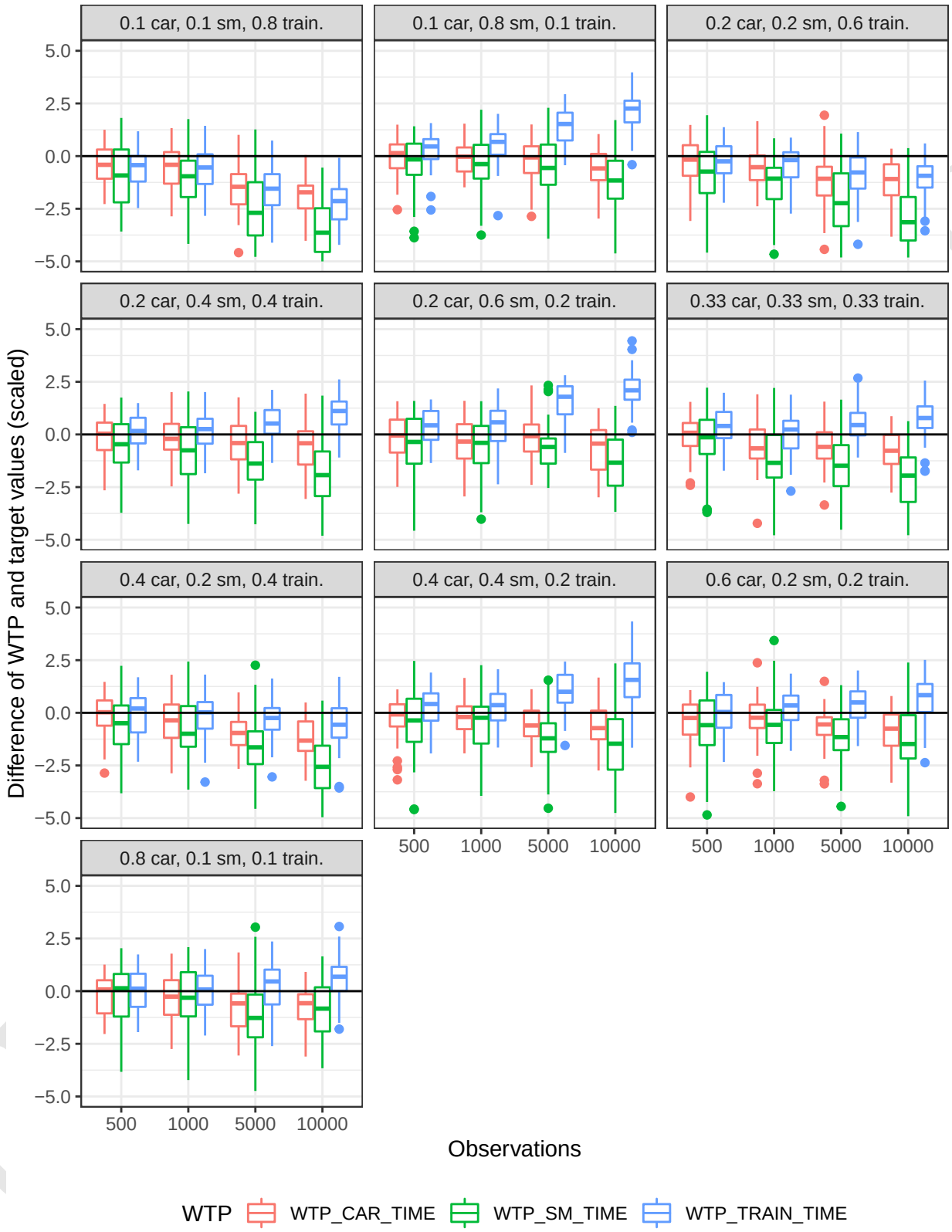

Figure A3: Difference of WTP and target values (scaled), by sample size and balance, MNL model only.

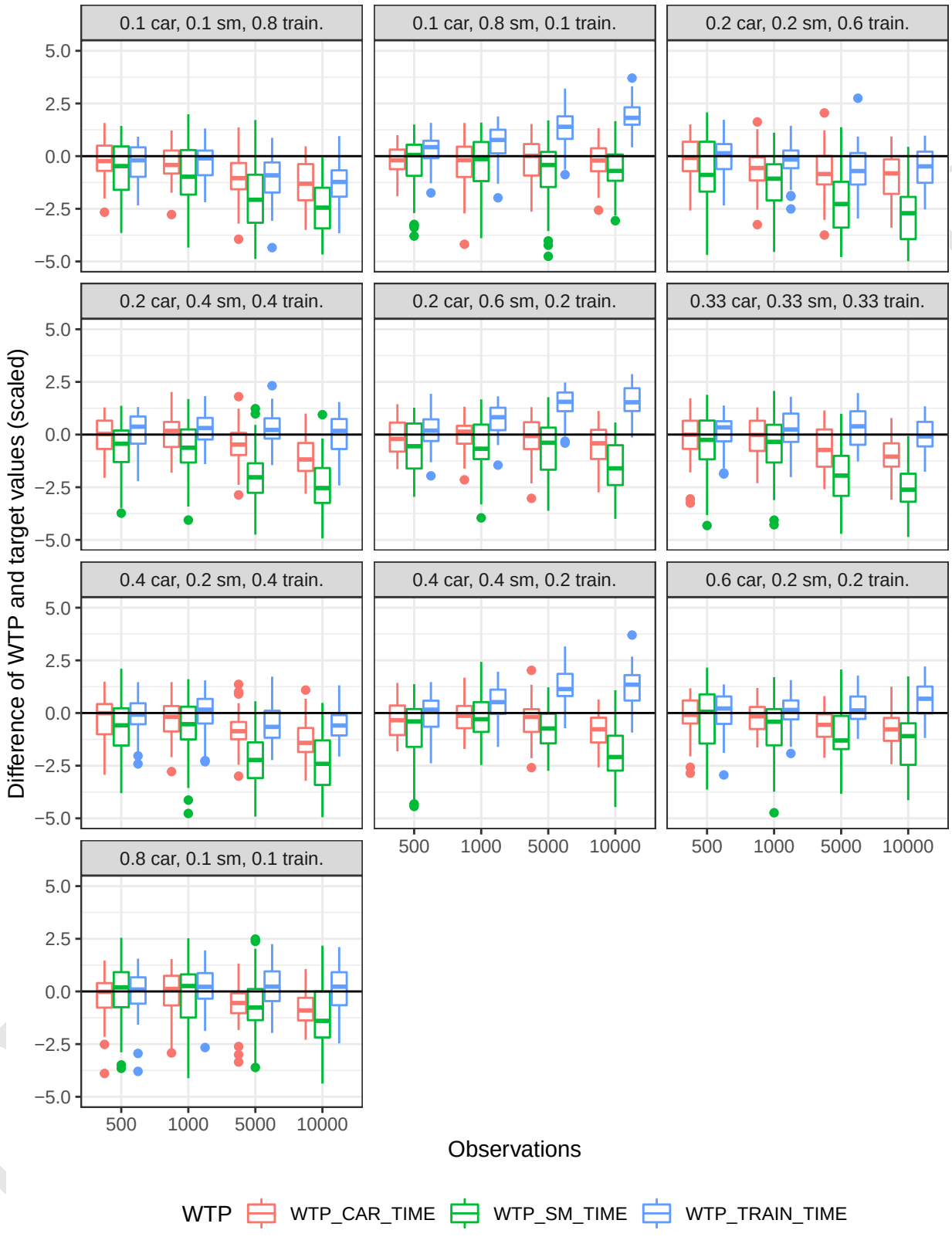

Figure A4: Difference of WTP and target values (scaled), by sample size and balance, MMNL model only.

Table A2: Shares of correctly estimated WTP by alternative, by sample size.

Model	S_TRAIN	S_SM	S_CAR	WTP_{CAR_TIME}	WTP_{SM_TIME}	WTP_{TRAIN_TIME}	All
MMNL	0.10	0.10	0.80	73.00	54.50	73.50	34.50
MMNL	0.10	0.80	0.10	84.00	68.50	50.00	30.50
MMNL	0.20	0.20	0.60	77.00	50.50	79.50	33.50
MMNL	0.20	0.40	0.40	75.50	56.00	61.00	29.50
MMNL	0.20	0.60	0.20	85.00	52.00	51.50	24.50
MMNL	0.33	0.33	0.33	68.50	39.50	80.00	26.00
MMNL	0.40	0.20	0.40	64.50	37.50	73.50	30.00
MMNL	0.40	0.40	0.20	73.50	40.00	84.00	30.50
MMNL	0.60	0.20	0.20	68.00	32.00	72.50	27.00
MMNL	0.80	0.10	0.10	66.50	36.50	65.50	26.50
MNL	0.10	0.10	0.80	65.00	47.50	70.00	31.00
MNL	0.10	0.80	0.10	77.00	58.00	53.00	26.50
MNL	0.20	0.20	0.60	79.00	52.50	68.00	29.50
MNL	0.20	0.40	0.40	76.50	48.50	59.00	21.50
MNL	0.20	0.60	0.20	69.50	55.50	45.50	18.50
MNL	0.33	0.33	0.33	70.00	42.00	73.00	24.00
MNL	0.40	0.20	0.40	63.50	36.50	70.50	28.50
MNL	0.40	0.40	0.20	71.50	42.00	67.50	20.50
MNL	0.60	0.20	0.20	62.50	36.00	71.00	28.00
MNL	0.80	0.10	0.10	50.50	29.00	44.00	21.50
NL	0.10	0.10	0.80	85.50	61.50	86.00	53.50
NL	0.10	0.80	0.10	83.50	67.50	89.00	51.50
NL	0.20	0.20	0.60	91.00	67.50	93.50	61.50
NL	0.20	0.40	0.40	92.00	66.50	94.00	63.00
NL	0.20	0.60	0.20	86.00	57.50	92.50	47.00
NL	0.33	0.33	0.33	97.00	56.00	98.00	54.50
NL	0.40	0.20	0.40	95.00	56.00	98.50	53.50
NL	0.40	0.40	0.20	82.00	39.00	88.50	34.00
NL	0.60	0.20	0.20	91.50	26.77	95.50	25.76
NL	0.80	0.10	0.10	84.42	19.70	89.95	19.19