Bridging Models and Modalities: Fusing Models Created from Diverse Multimodal Datasets

Lanlan Yan*, Rico Krueger[1]

[1]Department of Technology, Management, and Economics, Technical University of Denmark, Denmark

Short summary

Predictive generalization is a central challenge in transport choice modeling, especially when models are applied to new settings. One important source of this challenge is out-of-distribution (OOD) prediction, which can arise when data for the same decision problem are collected from different experimental settings. Such settings often produce heterogeneous datasets, including modality-rich and modality-lean data, that provide complementary but incomplete views of behavior. To leverage information from these datasets, we propose a general framework that extends model fusion methods to combine predictive distributions from models trained on heterogeneous datasets. Using two stated choice datasets, we find that PSIS-LOO stacking performs best overall. The results further show that the benefit of model fusion depends on the target domain, and that stacking improves performance mainly by correcting severe prediction errors in OOD cases.
Keywords: model fusion; dataset heterogeneity; shared mobility; discrete choice modelling

1 Introduction

Predicting travel behavior in new settings is a central objective in transport choice modeling. While many models achieve good in-sample fit, their ability to generalize to new contexts remains limited (Fox et al., 2014). In forecasting applications, models are often required to operate under conditions that differ from those observed during estimation, including both out-of-sample and out-of-distribution (OOD) scenarios. For example, OOD settings may arise when input variables such as travel time take values outside the range covered in the training data (Hess & van Cranenburgh, 2025), or when underlying travel behavior patterns change over time (Zhao et al., 2018).

Beyond these distributional shifts, transport choice applications may face a broader challenge. In particular, data for the same decision problem can be collected through multiple experimental settings, which may differ in sampled populations, experimental conditions, and recorded information.

More specifically, different experiments provide distinct observational views of the same choice process. Some datasets contain only observed choices and basic explanatory variables, whereas others additionally collect attitudinal indicators (e.g., Ton et al. (2020)) or process-related measurements such as eye-tracking and physiological responses (Bogacz et al., 2021; Argota Sánchez-Vaquerizo et al., 2024; Derakhshan et al., 2024). These datasets may also be drawn from heterogeneous groups due to different recruitment methods (Lee et al., 2022; Mottelson et al., 2021), implying both varying informational richness and potentially different data-generating processes.

In studies involving multiple experimental settings (e.g., Burris et al. (2023)), models are typically estimated separately for each dataset. A model trained on a modality-lean dataset does not observe richer attitudinal or process signals but can often be deployed at larger scale. Conversely, a model trained on a modality-rich dataset may lack exposure to broader population variability due to higher costs. More fundamentally, models estimated on one experiment do not fully capture the populations, measurement conditions, and information structures present in others. As a result, integrating such datasets introduces a broader form of out-of-distribution (OOD) prediction challenge.

This motivates the following question: can modality-rich and modality-lean datasets from different experiments be effectively combined to improve predictive generalization for the same decision

problem?

To address this question, we propose a framework that extends Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO), along with two other Bayesian model averaging methods from Yao et al. (2018), to settings with multiple heterogeneous datasets. Unlike the standard setting where stacking is applied to models estimated on the same dataset, our context involves models trained on separate datasets. We extend stacking to combine predictive distributions across datasets in a consistent manner. This approach leverages the complementary strengths of these datasets to improve predictive performance. While our focus is on transport choice modeling, the framework is broadly applicable to other problems involving heterogeneous or multimodal datasets.

2 Methodology

This section begins a workflow of the proposed late-level model fusion framework. Next, we introduce the three model fusion methods: PSIS-LOO stacking (our focus), pseudo-BMA, and BB-pseudo-BMA from the study of Yao et al. (2018).

Model fusion framework

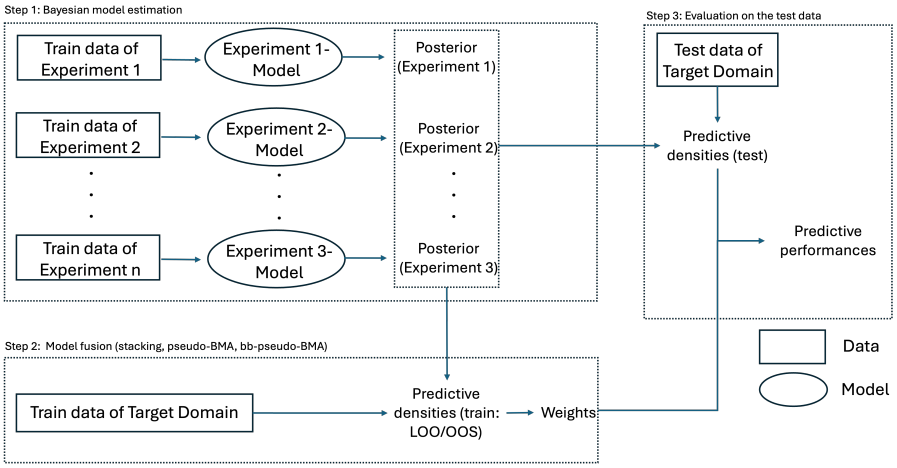


Figure 1: Workflow of model fusion

Figure 1 illustrates the general workflow of the proposed model fusion framework, which consists of three main steps. First, all datasets from different experiments (not limited to two, can be several) are split into training and test sets at the individual level using the same proportion. Then, the base behavioral model is estimated separately on the training data under a Bayesian framework, yielding posterior distributions of model parameters.

Second, model fusion weights are learned based on predictive performance evaluated on the training data. A key design choice in this framework is the definition of the *target domain*, which specifies the data distribution on which predictive performance is optimized and evaluated. The target domain is defined by the researcher prior to applying the framework and reflects the intended application setting. In practice, two common types of target domains can be considered: (i) a *single-domain* setting, where the target domain corresponds to one specific dataset, and (ii) a *pooled-domain* setting, where multiple datasets are combined to form a broader evaluation domain. These choices correspond to different research objectives: improving prediction within a specific dataset versus improving overall prediction by leveraging heterogeneous data sources.

For a given target domain, pointwise predictive densities are constructed for each model by combining (1) leave-one-out (LOO) estimates for within-domain predictions and (2) out-of-sample (OOS) predictions for cross-domain evaluations, which correspond to out-of-distribution (OOD) settings. Based on these predictive densities, model weights are computed using PSIS-LOO stacking, pseudo-BMA, and BB-pseudo-BMA. Third, the learned weights are applied to predictive densities on the

test data in the target domain to obtain predictive performance. The overall workflow remains the same across different settings, with differences arising solely from the definition of the target domain.

Last but not least, when combining datasets collected under different experimental settings, a structural mismatch between models may arise due to differences in available modalities (e.g., process data versus choice-only data). To enable meaningful model fusion, it is necessary to introduce an alignment mechanism that ensures predictive distributions are defined over a consistent input space. The specific form of this alignment may vary depending on the application (see section 3 on how the alignment works in this study's application).

PSIS-LOO stacking, pseudo-BMA and BB-pseudo-BMA

To combine predictive distributions from multiple models, we consider PSIS-LOO stacking, pseudo-BMA, and BB-pseudo-BMA. All three methods assign weights based on predictive performance, but differ in how these weights are constructed.
Suppose that M candidate models are available, each with predictive density $p_m(y_i \mid x_i)$ for observation i, where $m = 1, \ldots, M$. Let

$$\ell_{im} = \log p_m(y_i \mid y_{-i}, x_i) \tag{1}$$

denote the leave-one-out (LOO) log predictive density, approximated using Pareto smoothed importance sampling (PSIS). The corresponding expected log predictive density (ELPD) is

$$\widehat{\mathrm{ELPD}}_m = \sum_{i=1}^{N} \ell_{im}. \tag{2}$$

PSIS-LOO stacking combines predictive distributions through a weighted mixture

$$p(y \mid x) = \sum_{m=1}^{M} w_m \, p_m(y \mid x), \quad w_m \geq 0, \ \sum_{m=1}^{M} w_m = 1, \tag{3}$$

where the weights are obtained by maximizing the LOO predictive performance of the combined distribution:

$$\hat{w} = \arg\max_{w} \sum_{i=1}^{N} \log \left(\sum_{m=1}^{M} w_m \exp(\ell_{im}) \right). \tag{4}$$

In contrast, pseudo-BMA assigns weights based on model-level ELPD:

$$w_m^{\mathrm{pBMA}} = \frac{\exp(\widehat{\mathrm{ELPD}}_m)}{\sum_{j=1}^{M} \exp(\widehat{\mathrm{ELPD}}_j)}. \tag{5}$$

BB-pseudo-BMA extends this approach by incorporating uncertainty in ELPD via Bayesian bootstrap.

The key distinction is that stacking operates at the observation level by directly optimizing the predictive performance of the mixture distribution, whereas pseudo-BMA and BB-pseudo-BMA rely on aggregated model-level performance. Following Yao et al. (2018), stacking is expected to provide more robust predictive performance when models are complementary.

3 Application

This section starts with an introduction to the modality-rich laboratory data and modality-lean street-intercept data. To address the structural mismatch caused by missing attention information in the street data, we introduce an attention transfer mechanism within the MCPT model to achieve structural alignment. Then we discuss the two target domains in the application.

Dataset

The dataset is obtained from Ding et al. (2025), consisting of stated choice data from two experimental settings. A total of 152 ride-hailing drivers participated in street-intercept experiments, while 40 drivers participated in a laboratory experiment with eye-tracking measurements. Participants completed 2 (street) and 5 (lab) choice tasks respectively, resulting in 506 observations in total. This leads to **two datasets with complementary characteristics: a relatively large but modality-lean street dataset, and a smaller but modality-rich laboratory dataset**. Each task involves a choice between a conventional vehicle (CV) and an electric vehicle (EV), described by attributes including rental cost (RC), operating cost (OC, uncertain for EV), and driving range (DR). The two samples exhibit similar demographic characteristics, suggesting a relatively consistent population across datasets.

Base behavioral model and structural mismatch

We adopt the multi-attribute cumulative prospect theory (MCPT) model as the base behavioral model, following Ding et al. (2025). In MCPT, utility depends on attribute values, reference levels, and probability weighting under uncertainty. In the laboratory dataset, fixation counts are used to construct task-level attention weights, which modulate attribute sensitivities through a softplus transformation.

A key challenge is that attention information is only observed in the laboratory dataset. As a result, the predictive distributions differ in their information sets:

$$p_{\text{lab}}(y_i \mid x_i, A_i, \theta), \qquad p_{\text{street}}(y_i \mid x_i, \theta). \tag{6}$$

This discrepancy creates a structural mismatch: the two models are defined on different inputs, making it difficult to apply the lab model to street data without further adjustment.

Attention transfer and model alignment

To address this issue, we introduce an attention transfer mechanism that approximates the missing attention information in the street-intercept dataset. We assume that attention allocation is primarily driven by the structure of the choice scenario. Accordingly, for each scenario s, we compute the average attention vector from the laboratory data across observations sharing the same scenario, where attention weights are derived from fixation counts:

$$\bar{A}_s = E(A_i \mid s_i = s, D_{\text{lab}}). \tag{7}$$

For each street observation, the unobserved attention is approximated by assigning the corresponding scenario-level average, i.e., $A_i^{\text{street}} \approx \bar{A}_{s_i}$, yielding an augmented predictive model

$$p_{\text{street}}(y_i \mid x_i, \bar{A}_{s_i}, \theta), \tag{8}$$

which is structurally aligned with the laboratory-based model.

The stacked predictive distribution can be written as

$$p(y_i \mid x_i) = w_{\text{street}}\, p_{\text{street}}(y_i \mid x_i, \bar{A}_{s_i}, \theta) + w_{\text{lab}}\, p_{\text{lab}}(y_i \mid x_i, A_i, \theta), \tag{9}$$

The attention transfer mechanism is applied uniformly across the target domains of this study and may serve as a general solution to reconcile modality differences in other cases.

Target domains and evaluation settings

Figure 2 illustrates the application of the framework. A key design choice is the definition of the **target domain** that determines the data distribution for weight estimation and model evaluation.

We consider two target domains:
Target Domain 1 (street-only): This domain includes only street-intercept observations and represents a setting where performance is evaluated on modality-lean data. It allows us to assess whether information from the modality-rich (but smaller) laboratory dataset improves prediction

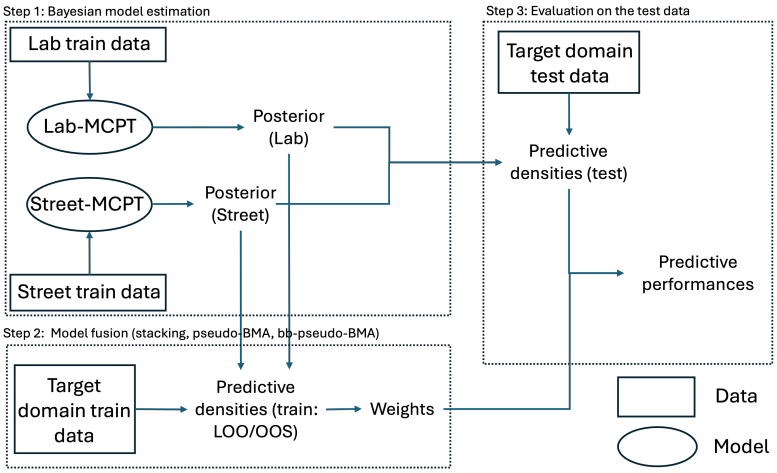


Figure 2: Workflow of model fusion for the dataset in Ding et al. (2025)

for street data.

Target Domain 2 (pooled): This domain includes both laboratory and street-intercept observations, representing a joint-domain setting where performance is evaluated over a combined data distribution. It allows us to assess whether modality-rich and modality-lean models provide complementary predictive value across datasets.

In Target Domain 1, performance is evaluated on street data by combining leave-one-out (LOO) estimates for the street-based model with out-of-sample (OOS) predictions from applying the laboratory-based model. In Target Domain 2, performance is constructed by combining within-dataset (LOO) and cross-dataset (OOS) evaluations across both laboratory and street data.

4 Results and discussion

This section is organized into four parts. First, we compare the three model fusion methods with the baseline models. Second, we examine performance differences between Target Domain 1 and Target Domain 2. Third, we analyze the contributions of the two base models to the stacked predictions. Finally, we extend stacking to a logit-based formulation with individual-level weights and compare it with standard stacking.

Table 1: Predictive performance and stacking weights (Target Domain 1: Street data only)

Model	Avg. Log Score	Variance	Weights_Street	Weights_Lab
Street-MCPT (baseline)	-0.4936	0.5863	–	–
Pooled-MCPT (street + lab, baseline)	-0.5452	0.1159	–	–
Pseudo-BMA	-0.4666	0.2101	1.0000	0.0000
BB-pseudo-BMA	-0.4666	0.2101	1.00000	0.0000
Stacking	**-0.4666**	0.2101	1.0000	0.0000

Baseline models vs. model fusion

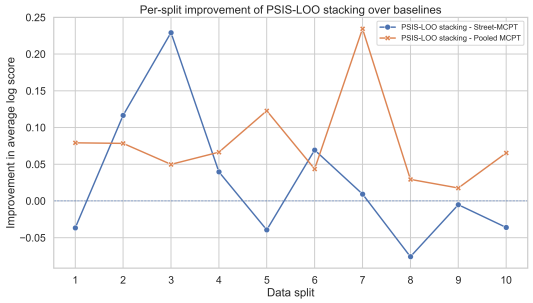
To evaluate whether model fusion improves predictive performance, we compare three fusion methods with several baseline MCPT models in both target domains (Tables 1 and 2). Results are averaged over 10 data splits. The baseline models differ in their training data. In Target Domain 1, both models are evaluated on street-intercept test data: *Street-MCPT* is trained on street data only, while *Pooled-MCPT* is trained on the combined dataset. In Target Domain 2, all models are evaluated on the pooled test set, including *Street-MCPT*, *Lab-MCPT*, and *Pooled-MCPT*. All baseline models are estimated using MLE.

Table 2: Predictive performance and stacking weights (Target Domain 2: Pooled Street and Lab data)

Model	Avg. Log Score	Variance	Weights_Street	Weights_Lab
Street-MCPT (baseline)	-0.7882	1.2200	–	–
Lab-MCPT (baseline)	-0.7981	0.1032	–	–
Pooled-MCPT (street + lab, baseline)	-0.6661	0.2624	–	–
Pseudo-BMA	-0.6249	0.4069	1.0000	0.0000
BB-pseudo-BMA	-0.6166	0.3356	0.9372	0.0628
Stacking	**-0.6133**	0.1502	0.6851	0.3149



(a) Predictive performance across data splits



(b) Per-split improvement over baseline models

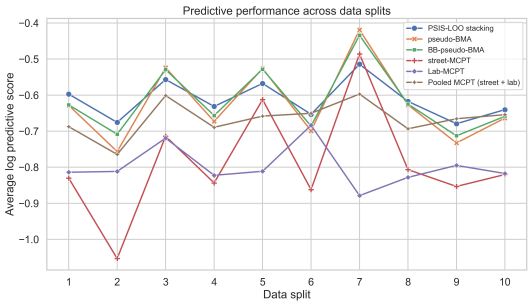
Figure 3: Predictive performance comparison in Domain 1

Across both target domains, fusion methods consistently outperform all baseline models in terms of average log predictive score. In Target Domain 1, *Street-MCPT* outperforms *Pooled-MCPT*, indicating that adding lab data may introduce noise when predicting street-based observations. In Target Domain 2, *Pooled-MCPT* performs better than single-source models, as the test data includes both lab and street observations.

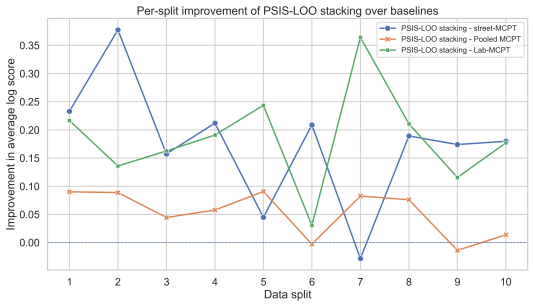
Figures 3a and 4a show that stacking achieves stable performance across data splits, while baseline models exhibit higher variability. The improvement is consistent across splits, rather than driven by a few cases. Among the three fusion methods, PSIS-LOO stacking performs best overall, in line with Yao et al. (2018). We therefore focus on stacking in the following analysis.

Target Domain 1 vs. Target Domain 2

In Target Domain 1, all fusion methods yield identical results and assign full weight to the street-based model (Table 1). This indicates that the lab-based model does not provide additional predictive value when the target domain consists only of street-intercept data. As a result, model



(a) Predictive performance across data splits



(b) Per-split improvement over baseline models

Figure 4: Predictive performance comparison in Domain 2

fusion collapses to the street model.

In Target Domain 2, the fusion methods produce different weights and performances (Table 2). In particular, stacking assigns non-zero weight to the lab-based model, suggesting that lab data provides complementary information when both data sources are included. In contrast, pseudo-BMA and BB-pseudo-BMA tend to favor the street-based model more strongly.

From a performance perspective, fusion yields larger improvements in Target Domain 2 than in Target Domain 1, as shown in Figures 4a and 4b. This indicates that model fusion is more beneficial when the target domain is heterogeneous.

Overall, these results suggest that the effectiveness of model fusion depends on the target domain. When the domain is homogeneous, fusion may reduce to a single model, whereas in heterogeneous settings, stacking can effectively combine complementary information from different data sources.

Understanding the gains from model stacking

Table 3: Predictive performance (Avg. Log Score) in Target Domain 2 (averaged over 10 data splits)

Model	Street-test	Lab-test	Pooled
Street-MCPT	-0.4666	-0.8625	-0.6645
Lab-MCPT	-0.7363	-0.7420	-0.7392
Stacking (PSIS-LOO)	-0.5163	-0.7589	**-0.6133**

Among the three fusion methods, PSIS-LOO stacking consistently achieves the best performance. We therefore **focus on stacking** and examine how it combines the two base models under in-distribution (ID) and out-of-distribution (OOD) settings.

As found in table 1, stacking collapses to the street-based model in Target Domain 1, providing no improvement over *Street-MCPT*. In this case, we will **focus on Target Domain 2**, which includes observations from both datasets and therefore involves cross-domain prediction. This settinhuog provides a more informative evaluation, as it captures how each model performs both within and outside its training distribution, and is where stacking can effectively leverage complementary strengths.

Table 3 reports average log predictive scores in Target Domain 2 (over 10 splits), and Figure 5 shows the distribution of pointwise log predictive densities for a representative split (seed = 42). They jointly illustrate the predictive patterns in Target Domain 2. The *Street-MCPT* model performs well on street-test data but deteriorates substantially on lab-test data, while the *Lab-MCPT* model shows more stable but less accurate performance across both subsets. The distributional results provide further insight into these patterns. On the street-test subset, stacking closely overlaps with *Street-MCPT*, effectively recovering the in-distribution model. On the lab-test subset, *Street-MCPT* exhibits a pronounced left tail, indicating severe prediction failures on a subset of observations, whereas *Lab-MCPT* remains more stable. The stacking model lies between the two, reducing extreme errors while maintaining reasonable predictive accuracy.

To further understand where the performance gains originate, Figure 6 presents the distribution of the pointwise improvement of stacking over Street-MCPT on the lab-test subset:

$$\Delta_i = \log p_{\text{stack}}(y_i) - \log p_{\text{Street}}(y_i).$$

We focus on the lab-test subset as an illustrative example of out-of-distribution (OOD) prediction for the *Street-MCPT* model, where it is evaluated on data generated from a different experimental setting.

The histogram shows that stacking does not uniformly improve performance; most observations exhibit small negative differences (left tail). However, the distribution has a heavy right tail, indicating substantial gains for a subset of observations where *Street-MCPT* performs particularly

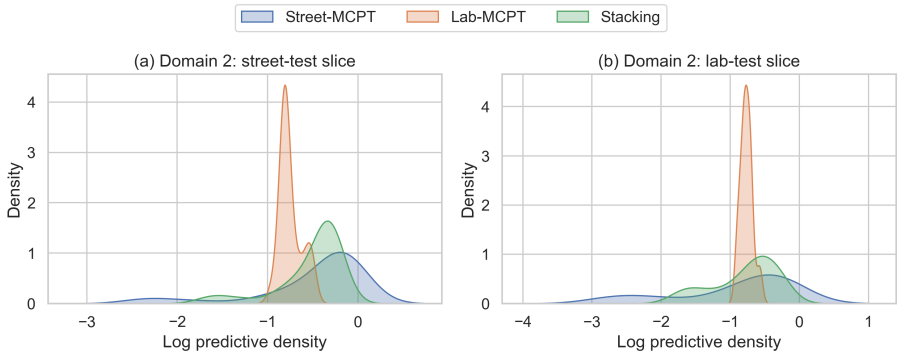


Figure 5: Distribution of pointwise log predictive densities in Target Domain 2

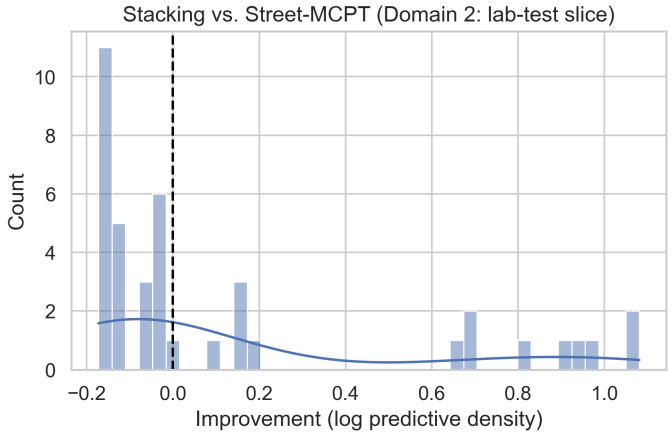


Figure 6: Distribution of pointwise improvement of stacking over Street-MCPT on the lab-test slice in Target Domain 2

poorly. These large positive improvements outweigh the small degradations elsewhere, leading to an overall increase in average log predictive performance.

These results suggest that the advantage of stacking arises from selectively correcting severe prediction errors in the out-of-distribution region, rather than uniformly improving all observations.

Logit-based individual weighting for model stacking

To account for heterogeneity in the relative importance of modality-rich and modality-lean data, we extend the standard stacking framework by introducing a logit-based formulation of model weights. While standard stacking estimates a single global weight shared across all observations, this assumption may be restrictive in our setting, where individuals are exposed to different experimental conditions.

We therefore specify individual-level weights using a binary logit model, where the street-based model is taken as the reference alternative. The resulting weights are:

$$w_{r,\text{lab}} = \frac{\exp(\eta_r)}{1+\exp(\eta_r)}, \quad w_{r,\text{street}} = 1 - w_{r,\text{lab}}, \tag{10}$$

where r indexes individuals and η_r is a linear function of individual-level characteristics. These weights are mapped to each observation in the panel data via $r(i)$, yielding the combined predictive distribution:

$$p(y_i \mid x_i) = w_{r(i),\text{street}}\, p_{\text{street}}(y_i \mid x_i) + w_{r(i),\text{lab}}\, p_{\text{lab}}(y_i \mid x_i). \tag{11}$$

In our implementation, the weights depend on four individual-level characteristics, including driving frequency, daily driving distance, attitudinal preferences, and driving experience. Selected variables are discretized into categorical indicators, while continuous variables are standardized. To ensure stable estimation, mild regularization and numerical stabilization (i.e., ridge penalty and

probability clipping) are applied.

From Table 4, logit-based stacking consistently improves predictive performance over standard stacking, with more pronounced gains in the pooled domain (Domain 2). In Domain 1, standard stacking collapses to a degenerate solution assigning full weight to the street-based model, whereas logit-based stacking introduces a non-zero contribution from the lab-based model, though the performance gain remains modest.

In Domain 2, logit-based stacking improves both predictive accuracy and stability, with more balanced weights between the two models. This suggests that modality-rich and modality-lean data provide complementary information in the pooled setting, and that individual-level weighting allows adaptive combination across individuals.

Overall, logit-based stacking outperforms standard stacking and mitigates degenerate weight solutions, although its effectiveness depends on the choice of target domain.

Table 4: Comparison between stacking and logit-based stacking across the two target domains (average in 10 splits).

Domain	Method	Avg. log score	Variance	Avg. w_{street}	Avg. w_{lab}
Target Domain 1	Stacking	-0.4666	0.2101	1.0000	0.0000
	Stacking-logit	-0.4617	0.1531	0.7938	0.2062
Target Domain 2	Stacking	-0.6133	0.1502	0.6851	0.3149
	Stacking-logit	-0.5821	0.1293	0.5920	0.4080

5 Conclusions

In this paper, we investigate how to leverage modality-lean and modality-rich datasets to improve predictive performance in transport choice modeling. Using two stated choice datasets collected from different experimental settings but addressing the same decision problem, we apply three model fusion methods and find that PSIS-LOO stacking consistently provides the best performance. In particular, stacking improves prediction both within a single dataset (Target Domain 1) and for the pooled dataset (Target Domain 2). Our results also show that the definition of the target domain plays an important role. Modality-rich data do not necessarily improve predictions in out-of-domain settings when used alone, but can provide substantial gains when combined with in-domain data. Observation-level analysis further reveals that stacking improves performance by correcting severe prediction errors in out-of-distribution regions, rather than uniformly improving all observations. These findings highlight the value of combining heterogeneous datasets to improve predictive generalization by stacking.

These results suggest that extending PSIS-LOO stacking to heterogeneous datasets with different modalities is a promising approach for improving predictive performance under distributional shifts, and that the proposed framework can be applied to other domains with similar data characteristics.

We further explore a logit-based extension of stacking with individual-level weights. The results show that this approach can outperform standard stacking while producing more balanced weights, suggesting that incorporating individual-level heterogeneity is a promising direction for improving model fusion in heterogeneous data environments, which we can continue refining in the future.

References

Argota Sánchez-Vaquerizo, J., Hausladen, C. I., Mahajan, S., Matter, M., Siebenmann, M., van Eggermond, M. A., & Helbing, D. (2024). A virtual reality experiment to study pedestrian perception of future street scenarios. *Scientific reports*, *14*(1), 4571.

Bogacz, M., Hess, S., Calastri, C., Choudhury, C. F., Mushtaq, F., Awais, M., ... Erath, A. (2021). Modelling risk perception using a dynamic hybrid choice model and brain-imaging data:

an application to virtual reality cycling. *Transportation research part C: emerging technologies*, *133*, 103435.

Burris, M., Brown, A., Gupta, H., Wang, J., Jha, K., Walk, M., ... others (2023). Examining the impact of stress and fear of covid-19 on transit travel.

Derakhshan, S., Nosrat Nezami, F., Wächter, M. A., Stephan, A., Pipa, G., & König, P. (2024). A situated inspection of autonomous vehicle acceptance–a population study in virtual reality. *International Journal of Human–Computer Interaction*, 1–20.

Ding, J., Yu, J. G., & Bansal, P. (2025). Preferences for electric vehicles under uncertain charging prices: An eye-tracking study. *Transportation Research Part D: Transport and Environment*, *140*, 104608. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1361920925000185` doi: https://doi.org/10.1016/j.trd.2025.104608

Fox, J., Daly, A., Hess, S., & Miller, E. (2014). Temporal transferability of models of mode-destination choice for the greater toronto and hamilton area. *Journal of Transport and Land Use*, *7*(2), 41–62.

Hess, S., & van Cranenburgh, S. (2025). Combine and conquer: model averaging for out-of-distribution forecasting. *arXiv preprint arXiv:2506.03693*.

Lee, J., Natarrajan, R., Rodriguez, S. S., Panda, P., & Ofek, E. (2022). Remotelab: A vr remote study toolkit. In *Proceedings of the 35th annual acm symposium on user interface software and technology* (pp. 1–9).

Mottelson, A., Vandeweerdt, C., Atchapero, M., Luong, T., Holz, C., Böhm, R., & Makransky, G. (2021). A self-administered virtual reality intervention increases covid-19 vaccination intention. *Vaccine*, *39*(46), 6746–6753.

Ton, D., Zomer, L.-B., Schneider, F., Hoogendoorn-Lanser, S., Duives, D., Cats, O., & Hoogendoorn, S. (2020). Latent classes of daily mobility patterns: the relationship with attitudes towards modes. *Transportation*, *47*(4), 1843–1866.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average bayesian predictive distributions (with discussion).

Zhao, Z., Koutsopoulos, H. N., & Zhao, J. (2018). Detecting pattern changes in individual travel behavior: A bayesian approach. *Transportation research part B: methodological*, *112*, 73–88.