A reproducibility ladder with design-aware re-estimation for survey-based travel research: a case study of twenty-two studies of the US National Household Travel Survey

Seyed Saber Naseralavi
Department of Civil Engineering, Shahid Bahonar University of Kerman, Kerman 7616914111, Iran
saber_naseralavi@uk.ac.ir

Abstract

Transport planners use regression coefficients and prediction scores from national travel surveys to judge how much ride-hailing, public transport or electric vehicles change travel. These numbers depend on choices that papers rarely state in full: which records were kept, whether the survey weights were used, how the uncertainty was computed, and how a prediction model was tested. We propose a reproducibility ladder with four rungs (rebuild the sample, rebuild the descriptive tables, refit the published model, refit it under the survey's own sampling design) and apply it to 22 published studies of the United States National Household Travel Survey (NHTS); the public files allowed the models of 13 studies to be refitted. The samples of 9 studies were rebuilt well, and 81 per cent of 1,454 printed descriptive values were matched. Where the weighted and the unweighted fit reproduce different numbers, the estimator the paper declared reproduced its printed values: in one study weighted odds ratios matched 57 per cent of the printed values and unweighted ones 3 per cent. With the replicate weights shipped with the survey, standard errors were a median of 2.2 times those of an unweighted fit. This cost came almost entirely from the unequal weights (2.2 times); clustering of persons within households (1.06 times) and the replicate formula (0.99 times) added little on average, and the standard errors of an unweighted analysis were understated by only 1.10 times. Under each study's declared inference, 24 per cent of the significant coefficients (exact interval 18 to 31 per cent) were not significant under the design, and 8 per cent of the coefficients changed sign between the unweighted and the weighted fit. For classification models, oversampling the rarer class before splitting the data raised the area under the receiver operating characteristic (ROC) curve, a score of chance level at fifty per cent and perfection at one hundred per cent, of a random forest from 63 per cent to 99 per cent; the printed value was reproduced by neither pipeline. A published regression score, the share of variance explained, of 71 per cent was not reproduced (20 per cent on held-out households). The machinery is released as two modules of the open-source Python package pynhts.

Keywords: National Household Travel Survey; reproducibility; survey weights; replicate-weight variance; data leakage; negative binomial regression

Introduction

The US National Household Travel Survey (NHTS) is the main public source on how Americans travel. Its waves of 2001, 2009, 2017 and 2022 hold between about eight thousand (2022) and one hundred and fifty thousand households each, with every member's trips on one assigned day, the vehicles, and a long list of household and person attributes. Researchers use the files to estimate who rides, who pays and who emits, and the resulting regression coefficients, odds ratios and prediction scores are read by planners as evidence on ride-hailing, bike-sharing, transit, electric vehicles and vehicle miles travelled (Alberini et al., 2021, Conway et al., 2018, Wu and MacKenzie, 2021). A coefficient such as "one more ride-hailing trip per month is associated with more trips per day" is a short sentence that rests on a long chain of decisions: which respondents entered the analysis, whether the survey weights were applied, how the standard error was formed, and, for a prediction model, how it was tested on data it had not seen.

Those decisions matter because the NHTS is not a simple random sample. Households are drawn in strata, persons are clustered within households, response differs by group, and the files therefore carry person and household weights and, for the 2001, 2009 and 2017 waves, between ninety-eight and one hundred replicate weights per record, so that analysts can obtain variance estimates that respect the design (Rust and Rao, 1996, Wolter, 2007). Whether to weight a regression is a long-standing question in survey statistics: weights can remove bias when selection depends on the outcome, but they cost efficiency when they do not (Bollen et al., 2016, Gelman, 2007, Solon et al., 2015, Winship and Radbill, 1994). Whether to use the design when computing standard errors is less debatable: ignoring clustering and stratification understates uncertainty (Lumley, 2004). Demonstrations from health surveys show that an association that is significant in an unweighted analysis can cease to be so once weights and design are respected (Birrell et al., 2019, Gard et al., 2023). How often this happens in transport research, where regression on NHTS data is routine, has not been measured.

A second doubt concerns reproducibility more broadly. Computational results are often hard to regenerate from the published description (Hardwicke et al., 2021, Peng, 2011, Stodden et al., 2018), and when independent analysts receive the same data and question, defensible but different choices can lead to different conclusions, sometimes in opposite directions (Schweinsberg et al., 2021, Silberzahn et al., 2018, Simonsohn et al., 2020, Steegen et al., 2016). Prediction models add a further layer. Machine learning has entered travel-behaviour research with claims that flexible models beat classical ones (Hagenauer and Helbich, 2017, Martín-Baos et al., 2023, Wang et al., 2024), while the wider literature documents how leakage between training and test data produces over-optimistic scores in many fields (Kapoor et al., 2024, Kapoor and Narayanan, 2023, Rosenblatt et al., 2024). Two leaks are especially relevant here. Creating synthetic examples of a rare class before the data are split puts copies of test rows into training, which inflates scores sharply (Blagus and Lusa, 2013, Chawla et al., 2002, Demircioğlu, 2024). Splitting rows at random when rows share a household, a place or a person puts near-duplicates on both sides of the split, and validation designs that respect such structure give lower and more honest scores (Ploton et al., 2020, Roberts et al., 2017, Vabalas et al., 2019). NHTS rows are exactly of this kind: persons of one household share almost everything that matters.

Work on the NHTS itself has mostly examined the quality of its data, for example selection bias among respondents who agree to follow-up surveys (Wang et al., 2022), or used it to answer substantive questions (Conway et al., 2018). Open pipelines exist for synthetic travel demand from other national surveys (Hörl and Balac, 2021). We found no study that takes a body of published NHTS analyses, rebuilds their samples and tables, refits their models, and asks whether their conclusions survive the survey's own design and a leakage-free validation. This is a gap in a double sense: the size of the effect of ignoring the design in transport regressions is unknown, and the extent to which published prediction scores depend on the validation pipeline is unknown.

This paper fills the gap with a method and a case study. The method is a reproducibility ladder with four rungs, each of which can be graded against printed numbers, together with a design-aware re-estimation that compares three ways of fitting the same model and a re-scoring of classifiers under four pipelines. The case study applies it to 22 published studies of the NHTS, collected as a convenience set, of which 13 could be refitted from the public files. We ask four questions: how much of each study can be rebuilt; whether the printed estimates are reproduced by the estimator the paper declares; how much the conclusions change under the survey design; and how much a prediction score depends on the pipeline. The same machinery is released as two modules of the open-source package pynhts, so that the check takes a few lines of code instead of a few weeks.

Materials and methods

The case-study set and the data

The set consists of 22 studies published between 2011 and 2025 that analyse the NHTS and print enough numbers to compare with (Table 1). They are listed in the table (Alberini et al., 2021, Barua et al., 2021, Deka and Fei, 2019, Fu et al., 2020, Jia, 2019, Jia et al., 2020, Jiao et al., 2020, Krizek and McGuckin, 2019, Lidbe et al., 2021, Naseralavi et al., 2025, Quinn et al., 2016, Sabouri et al., 2020, Sadeghvaziri and Tawfik, 2020, Shirgaokar and Nobler, 2021, Su et al., 2021, Wu and MacKenzie, 2021, Xue et al., 2021, Yang et al., 2011; Zhang and Zhang, 2018a, 2018b, 2018c; Zhou and Wang, 2014) and were collected by the authors while building the pynhts package, so they are a convenience set and not a systematic sample of the literature; we draw no conclusion about the share of all NHTS papers that behave like them. One study is by the author, who is also the package developer; it is included in the register but was not re-estimated here, and its models are reserved for a separate paper. Public-use NHTS files of 2001, 2009 and 2017 were obtained from the survey's website through pynhts, which verifies each file against its codebook. All 22 studies use the 2001, 2009 or 2017 wave.

Table 1: The 22 case studies and the rung each reached. L1 gives the number of printed descriptive values graded and, in brackets, the share graded EXACT or CLOSE.

ID	Study	Wave	Model	Sample (L0)	Printed values matched (L1)	Model (L2)
P01	Yang 2011	2001	logit	rebuilt	154 (95%)	reproduced
P02	Zhou 2014	2009	SEM	partly	23 (48%)	not reproduced
P03	Quinn 2016	2009	logit	partly	21 (48%)	partly
P05	Deka 2019	2017	ZINB	partly	86 (70%)	not run
P06	Jia 2019	2017	ML classifiers	rebuilt	42 (86%)	not reproduced
P07	Krizek 2019	2017	descriptive	partly	18 (22%)	not run
P09	Fu 2020	2017	simulation	not rebuilt	13 (8%)	not run
P10	Lidbe 2021	2001/2017	descriptive	rebuilt	336 (90%)	not run
P11	Alberini 2021	2017	OLS	rebuilt	50 (94%)	partly
P12	Xue 2021	2017	SEM	not rebuilt	20 (35%)	not run
P13	Zhang 2018a	2017	ZINB	partly	80 (69%)	no valid target
P14	Zhang 2018b	2017	ZINB	rebuilt	97 (97%)	partly
P15	Zhang 2018c	2017	ZINB	rebuilt	85 (96%)	no valid target
P16	Sabouri 2020	2017	multilevel Poisson	partly	30 (70%)	not run
P17	Jia 2020	2017	ML classifiers	rebuilt	40 (85%)	not reproduced
P18	Jiao 2020	2017	NB	rebuilt	55 (87%)	reproduced
P19	Sadeghvaziri 2020	2009/2017	descriptive	partly	11 (82%)	not run
P20	Shirgaokar 2021	2017	NB	partly	63 (41%)	partly
P21	Barua 2021	2009/2017	GBM	partly	40 (48%)	not reproduced
P22	Wu 2021	2017	PSM	rebuilt	71 (97%)	not reproduced
P23	Su 2021	2017	motif clustering	partly	65 (65%)	not run
P24	Naseralavi 2025	2017	ML per state	partly	54 (91%)	not run

The reproducibility ladder

The ladder has four rungs (Figure 1). **L0**: rebuild the analysis sample from the public files by applying the paper's rules in the order given, and compare the sample size with the printed one. **L1**: rebuild the descriptive tables and grade every printed value as EXACT (equal at the printed precision), CLOSE (within two per cent), ACCEPTABLE (within five per cent) or GAP. **L2**: fit the published model with the same variables and the estimator the paper declares (weighted or unweighted), and grade each printed coefficient or score in the same way, with a tolerance of at least the last printed digit. **L3**: refit the model under the survey design. No rule was altered to reach a printed number: a difference is recorded as a gap, and the most natural reading of an unwritten rule is documented with the evidence that supported it. Rungs L0 and L1 were completed before this paper's models were run, and their grading is reproduced from the audit files of the package repository. The analysis plan was frozen after three studies (Yang, Quinn, Jiao) had been run during the development of the engine, and these hold about half of the coefficients and all three headline examples; deviations from the plan, including a reformulated fourth hypothesis, are logged in it and summarised in the supplement.

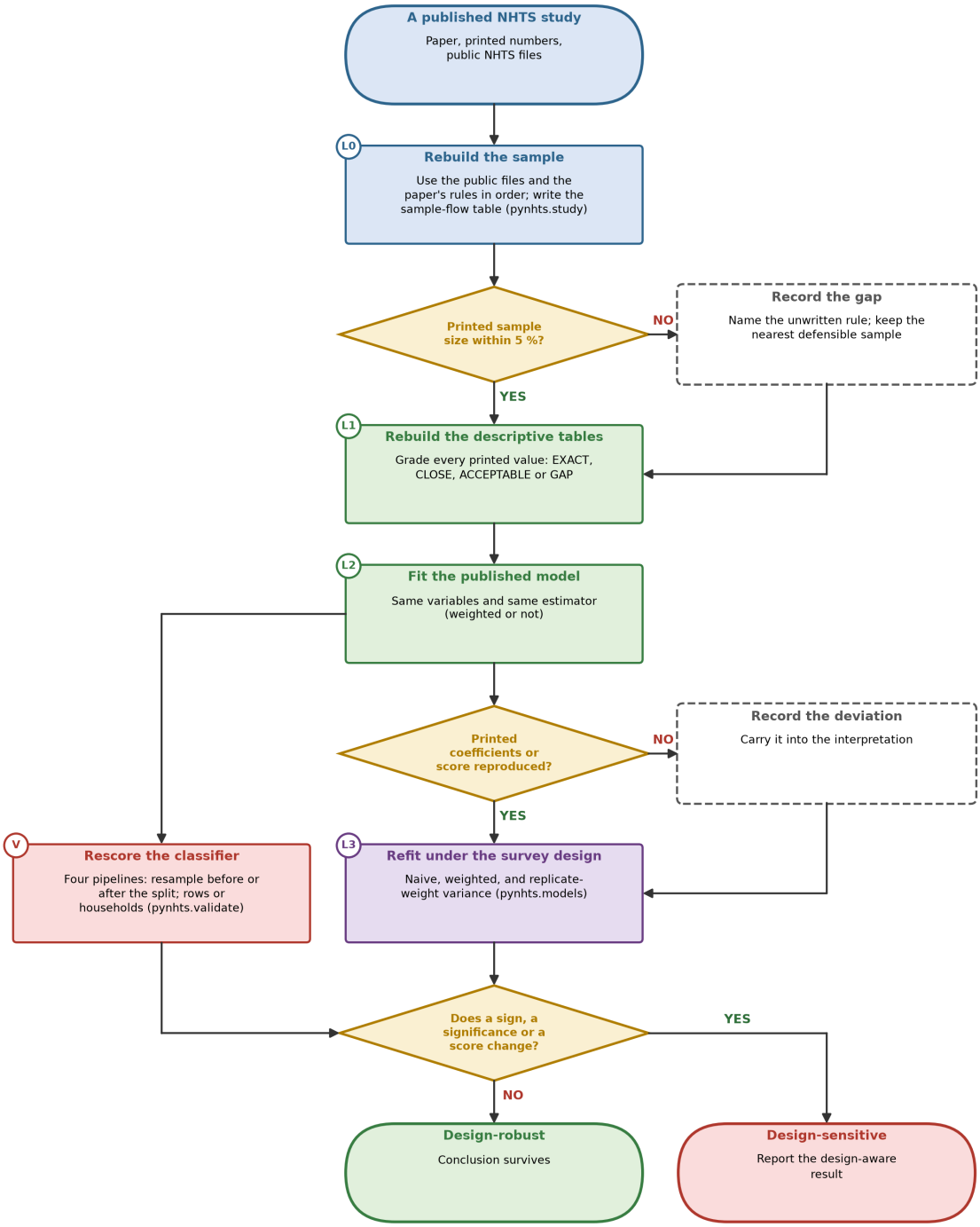


Figure 1: Method of the paper. Rungs L0 to L3 are applied to each published study; the box V rescoring a classifier applies to machine-learning studies. Diamonds are decisions; dashed boxes record what is carried forward.

Four ways to fit the same model

Each regression was fitted four ways. The unweighted fit uses maximum likelihood with model-based standard errors, which is what most of the studies report. The clustered fit keeps the unweighted estimate and forms the variance with a sandwich in which observations of the same household form one cluster. The weighted fit uses the person, household or person-day weight of the observation (scaled to the sample size for estimation) with the same household-clustered sandwich; the same fit with every observation its own cluster measures what the unequal weights alone cost. The design-aware fit uses the same weighted estimate with the variance design of the wave. The fits answer different questions: the unweighted estimate describes the association among respondents and the weighted one the association in the population the weights represent, so a comparison of their standard errors mixes a change of estimand with a change of precision; we therefore report the standard errors as a chain of ratios (weights, clustering, replicate formula) and not only their product. For the waves with replicate weights this is the jackknife: the model is refitted with each of the R replicate weights w_r, and the variance of a coefficient $\hat{\beta}$ is

$$\hat{V}(\hat{\beta}) = f \sum_{r=1}^{R} \left(\hat{\beta}_r - \hat{\beta}\right)^2,$$

with $R = 98$ and $f = 6/7$ for 2017, $R = 100$ and $f = 99/100$ for 2009, and $R = 99$ and $f = 98/99$ for 2001, as in the User's Guides. Logistic, linear and Poisson regressions are refitted exactly from warm starts. Negative binomial and zero-inflated negative binomial regressions, whose refitting takes minutes, use Newton steps from the full-sample optimum, a step that we checked against exact refits. The count-model likelihoods and their gradients are our own, checked against finite differences and against the statsmodels library (Lambert, 1992, Zeileis et al., 2008). As independent checks, the logistic, Poisson and linear fits on adults of the 2017 person file were compared with the survey package in R (a replicate-weight design with the same factor): coefficients agreed to one part in one hundred thousand and standard errors to 0.2 per cent; the Taylor-series variance of the 2022 wave was checked against the same package in the same way. On the weekend negative binomial model of one study, 98 exact refits gave standard errors within 0.2 per cent of the Newton ones. The regime in which a study's printed numbers were reproduced is reported as its declared estimator.

Models of the case studies

Eight published regression studies were refitted: logistic regression of walking and cycling (two studies); ordinary least squares with fixed effects for vehicle make, state and density (one); negative binomial regression of daily trips (two); and zero-inflated negative binomial regression of monthly shared-mobility use (three). A structural equation model with two latent propensities, a propensity-score matching of ride-hailing user groups, a gradient-boosting regression of household online shopping and five classifiers of alternative-fuel and electric vehicle adoption complete the set. Where a printed coefficient table exists, each printed coefficient is a target of L2; where only a headline is printed (the public-transit effect in two ZINB studies, the elasticity in the OLS study), that headline is the target, and where it was derived from a percentage in the abstract this is stated in the supplement.

Machine-learning models, validation designs and tuning

Five classifiers were re-run on the sample of 30,511 new-vehicle owners in 27,413 households, with the outcome alternative-fuel vehicle (AFV, 1,236 owners) or electric vehicle (EV, 389 owners). The sample rule (newest vehicle of each main driver, at most two model years old, complete key variables) was recovered from the printed sample size and is slightly smaller than the printed one. The models are explained in Table 2. Logistic regression turns a weighted sum of the inputs into a probability; the naive Bayes classifier assumes each input to be bell-shaped within a class and independent of the others; a decision tree asks a sequence of yes/no questions on the inputs; a random forest averages the votes of many trees grown on random parts of the data; and a support vector machine with a radial kernel separates the classes with a smooth boundary built from the training points nearest to it. A sixth family, gradient boosting, which adds small trees one after another so that each corrects the errors of the sum of the earlier ones, was used for the household-level online-shopping regression together with linear regression, a support vector regression and a random forest. These four regression models were not tuned: they were run with fixed, documented settings, those printed in the study where it printed them (cost ten and kernel width one hundredth for the support vector regression, a minimum leaf of twenty-five rows for the forest), because the question was whether the printed scores can be approached with the printed settings.

Table 2: The five classifiers and the search space of their settings (the same ten-trial random search for every model).

Model	Idea in plain words	Searched hyper-parameters (ten random trials each)
LR	Logistic regression: a weighted sum of the inputs turned into a probability	penalty strength C (0.001-100, log scale); penalty type (L1 or L2)
NB	Gaussian naive Bayes: each input is assumed bell-shaped within each class and independent of the others	variance floor (1e-11 to 1e-3, log scale)
DT	Decision tree: a sequence of yes/no questions on the inputs that ends in a leaf with a class	maximum depth (3, 5, 8, 12, 20); minimum leaf size (1, 5, 20, 50); split criterion (Gini or entropy)
RF	Random forest: many trees grown on random parts of the data whose votes are averaged	number of trees (50, 100, 200); maximum depth (5, 10, 20, 40); minimum leaf size (1, 2, 5, 20); share of inputs tried per split (0.3, 0.5, 1)
SVM	Support vector machine with a radial kernel: separates the classes with a smooth boundary built from the training points nearest to it	cost C (0.1-100, log scale); kernel width gamma (0.001-1, log scale); trained on at most 12,000 rows

Every model was tuned alike. The search space of every model is given in Table 2. The search method is random search with the same budget (ten trials) for every model; each trial is scored by three-fold validation inside the training rows, grouped by household in the pipelines that split by household and by row in those that split by row, with oversampling inside the inner training folds only when the pipeline oversamples after the split; the criterion is the area under the receiver operating characteristic curve. The selected configuration is refitted on all training rows and scored once on the held-out rows (nested validation). The experiment was run with two master seeds. The four pipelines differ only in where the oversampling happens and in what is split. In *oversample, then split rows* the rarer class is oversampled on the whole file by interpolating between near neighbours (Chawla et al., 2002), and the rows are then split at random; in *split rows, then oversample* the split comes first and only the training rows are oversampled; in *split households, then oversample* every household lies wholly on one side; and in *split households, no oversampling* the training rows are used as they are and the held-out score is weighted by the survey weight. About 9,189 rows, from about 27,413 households, are held out in every split of the household designs (Table 3), and the frequency histogram of the rare class in all, training and held-out rows is in Figure 8 of the supplementary material. Because the held-out households are a random subset of all households, the designs test generalisation to new households within the same sample; no held-out group lies outside the range of the training groups, so none is an extrapolation in the sense of a convex hull. For the regression study, a second design holds out whole states, which is a harder test of transfer.

Table 3: Rows removed and kept in the household split and the share of each vehicle class in them (the frequency of the rare class before and after removal).

Group	Rows	Share AFV	Share EV
All new-vehicle owners	30,511	4.1%	1.3%
Training households (mean of three splits)	21,322	4.1%	1.3%
Held-out households (mean of three splits)	9,189	3.9%	1.3%

Software and reproducibility

All numbers, tables and figures are generated by code from the result files; the text contains no typed result. The analysis code, the engine of the regression regimes and the validation pipelines are in the repository, run by one command, and the regression and validation machinery is part of pynhts (modules `pynhts.models` and `pynhts.validate`), with unit tests, a real-data test against R and documentation. The analysis plan, with the hypotheses and thresholds, was written before the models of the second half of the studies were run and is dated.

Results

What could be rebuilt

The ladder is shown in Figure 2. Of the 22 studies, the audit judged the public files sufficient to rebuild the data of 9, partly sufficient for 11, and insufficient for 2, for which the population or sampling rule is not stated. Over all studies, 879 of 1,454 printed descriptive values were

reproduced exactly at the printed precision, and 81 per cent exactly or within two per cent. The gaps came mostly from the papers: unstated rules for dropping records, arithmetic that does not add up inside a table, and variables that are not in the public file. 13 studies could be taken to the model rung; for the others the reasons are listed in Table 1.

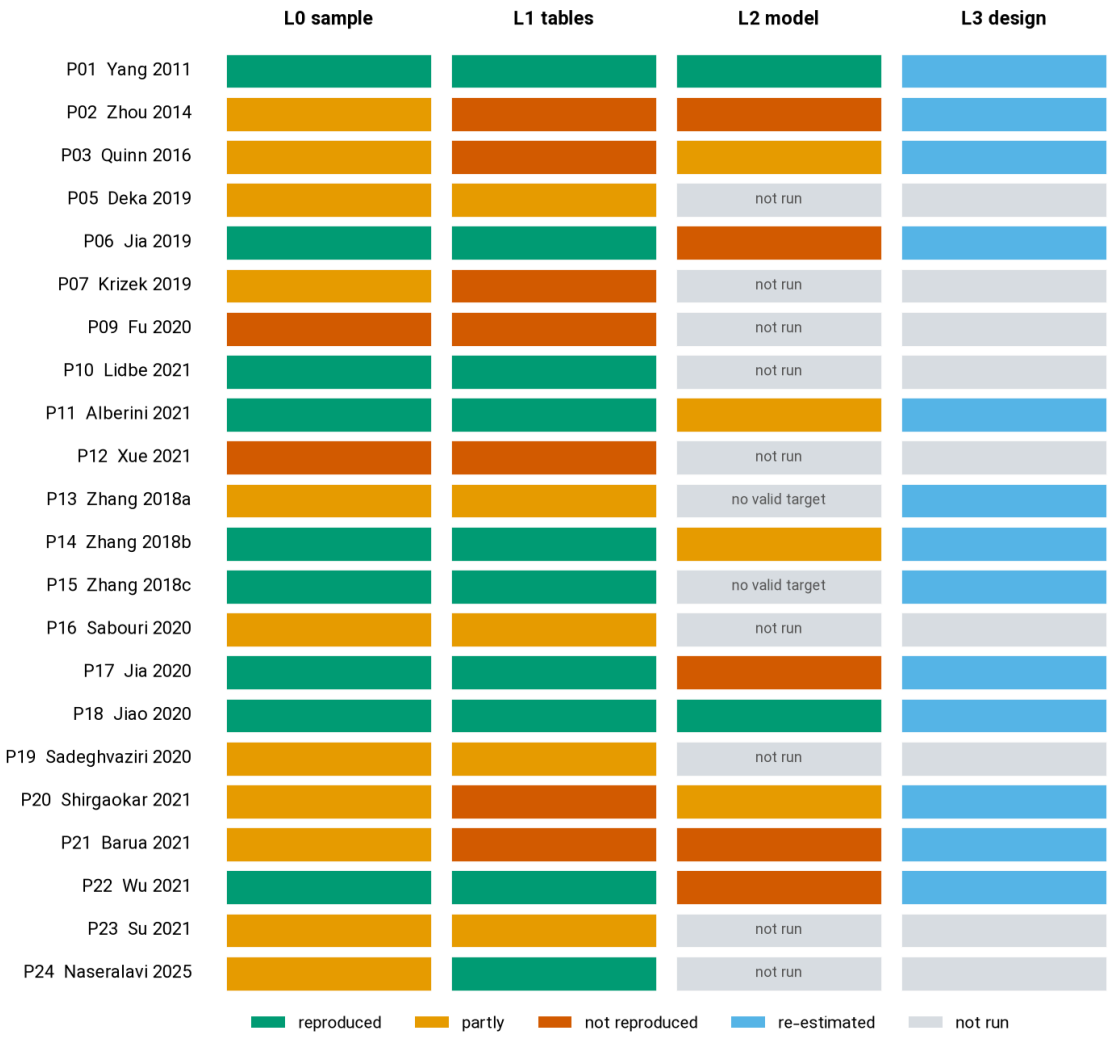


Figure 2: The ladder for the 22 studies. Each row is a study, each column a rung; colours give the grade. The L3 column shows which models were re-estimated, not a grade.

The declared estimator reproduces the printed numbers

Among the studies in which the unweighted and the weighted fit reproduce clearly different numbers of printed values, the declared estimator reproduced more. The study of walking and cycling by Quinn and colleagues states that population weights and the jackknife were used: the weighted odds ratios matched 34 of 60 printed values and the unweighted ones 2; for the two commuting models, 29 of 30 matched to the printed precision or within two per cent, but for the two transport models few did, which we attribute to the sample (we could not rebuild the printed

one). In the study of physical activity by season the weighted odds ratios matched 21 of 21 printed values and the unweighted ones 6. In the other direction, the negative binomial study of shared mobility and trips, which reports an unweighted fit in R, was reproduced by the unweighted fit (30 of 32 coefficients within two per cent, 20 at the printed precision), and the zero-inflated study of vehicle ownership and ridesharing for 15 of 32 coefficients within two per cent and 25 within five per cent. In the other studies neither fit reproduced much: the elasticity of annual miles with respect to fuel cost per mile matched for one of 4 printed values (−0.515 against −0.510). The first hypothesis is therefore supported for the four studies in which the two estimators can be told apart, and not tested for the others; we do not generalise it to all studies.

The estimator matters for the substance, not only for the match. The most striking case is the education effect on cycling to work. The printed odds ratio for a graduate degree compared with the lowest education group is 0.41, the weighted fit gives 0.41, and the unweighted fit gives 1.71: the sign of the association with education reverses (Figure 3, reference letter H). The same happens for the Midwest effect on active travel in 2001, with printed 0.67, weighted 0.67 and unweighted 0.83. An unweighted fit estimates the association among respondents, and these results show that respondents differ from the population in ways that matter for this association; we did not test formally whether the weights are informative.

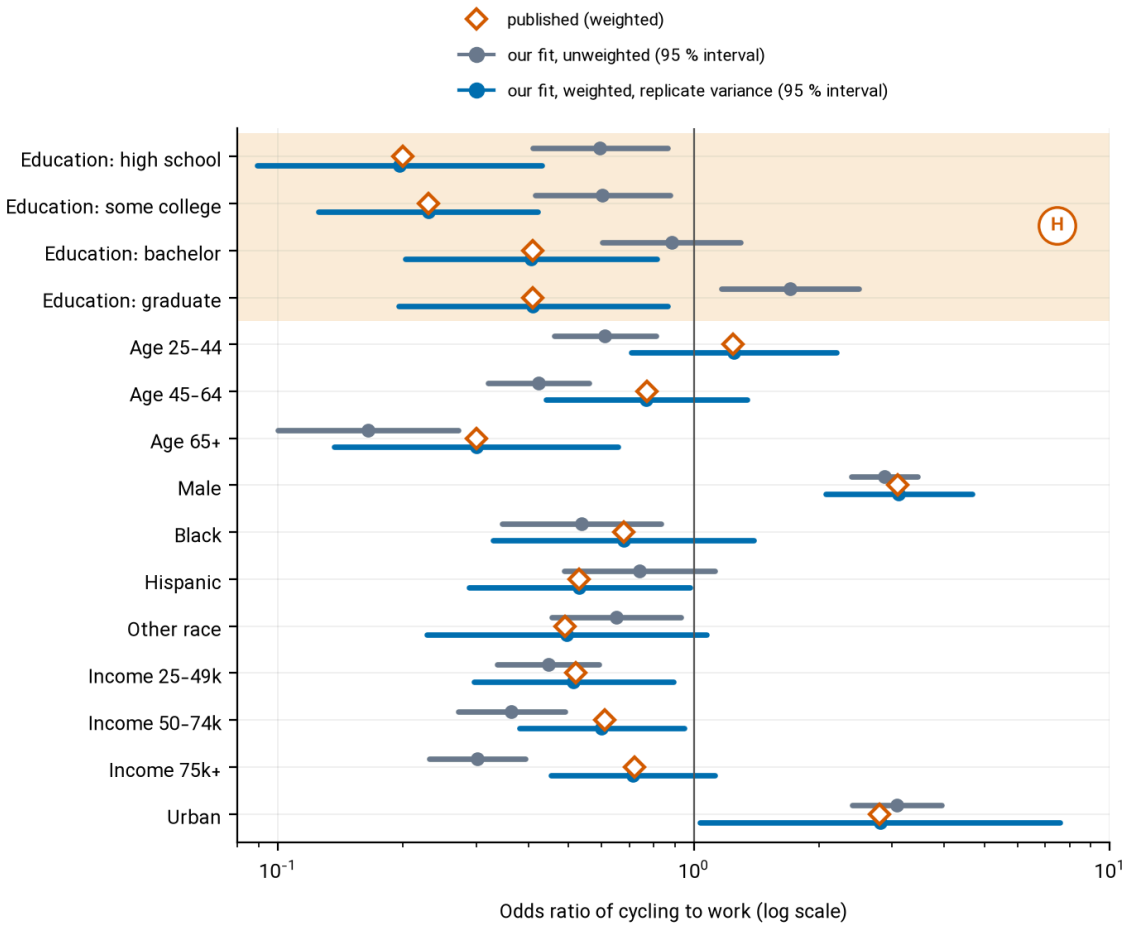


Figure 3: Odds ratios of cycling to work in the study of Quinn and colleagues: printed, our unweighted fit and our weighted fit with replicate-weight variance. The shaded rows (reference letter H) are the education effects discussed in the text.

Standard errors and significance

Across the 310 re-estimated coefficients of eight regression studies, the design-aware standard error was a median of 2.19 times the unweighted one (interquartile range 1.86 to 2.66). Figure 4 (panel A) shows where this comes from. Weighting alone, with every observation its own cluster, raised the standard error by a median factor of 2.17; clustering persons within households added 1.06 on top of that, and the replicate-weight formula a further 0.99 (fifth to ninety-fifth percentile 0.70 to 1.44). An unweighted fit whose standard errors were clustered on households differed from its model-based standard errors by a median factor of only 1.10. The unweighted analyses were therefore not wrong about the precision of what they estimate, an association among respondents; the large standard errors are the price of estimating the population association with unequal weights.

Under each study's declared inference (unweighted for six studies, weighted with household-clustered errors for one; the study that itself used the jackknife is excluded), 42 of 174 coefficients

that were significant at the five per cent level were not significant under the design-aware fit, 24 per cent with an exact ninety-five per cent interval of 18 to 31 per cent; pooled against the unweighted fit for all studies, the figures are 78 of 244. The share differs by study (Figure 4, panel B; Table 4). The point estimates of 26 coefficients (8.4 per cent) differ in sign between the unweighted and the weighted fit, but only 8 of these were significant in the unweighted fit and 1 in both, so most sign changes concern coefficients that are close to zero. The second hypothesis (median ratio above one and a half) and the third (more than one tenth of significant coefficients lose significance, here with a lower interval limit above one tenth) are supported.

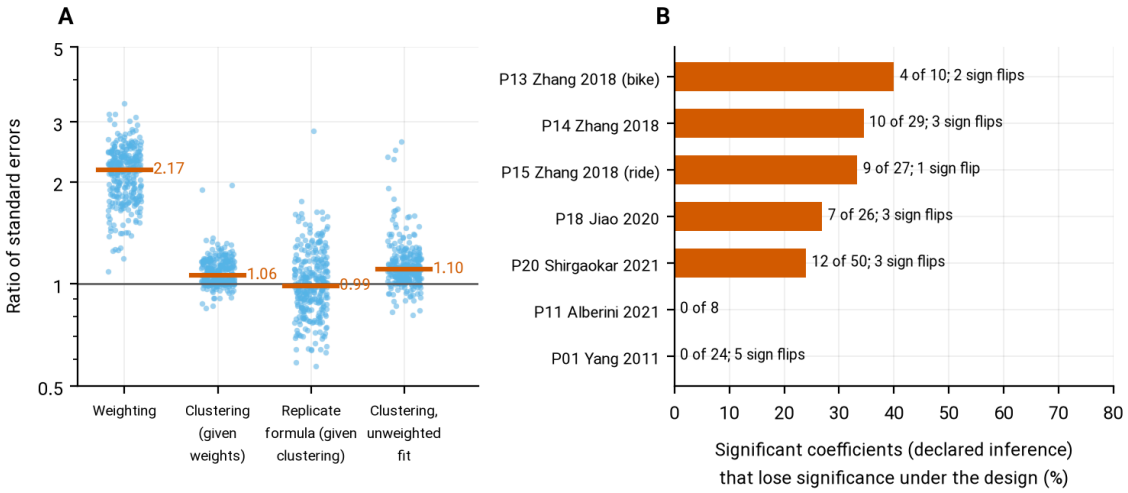


Figure 4: Where the standard-error difference comes from and what it changes. Panel A: ratios of standard errors over all re-estimated coefficients, by source (orange: median). Panel B: share of coefficients, significant under each study's declared inference, that lose significance under the design-aware fit, by study.

Table 4: Re-estimation of eight regression studies. The three ratio columns are median ratios of standard errors: weighted model-based over unweighted; weighted household-clustered over weighted model-based; replicate-weight over weighted household-clustered. Significance lost: coefficients significant under the study's declared inference that are not under the design-aware fit (the study that itself used the jackknife has none). Sign flip: unweighted and design-aware estimates differ in sign.

Study	Coefficients	Printed values reproduced: unweighted / weighted / printed	Weights	Clustering	Replicate formula	Significance lost (declared inference)	Sign flips
P01 Yang 2011	50	6 / 21 / 21	1.78	1.12	0.94	0 / 24	5
P03 Quinn 2016	72	2 / 34 / 60	2.33	1.01	0.99	-	9
P11 Alberini 2021	8	1 / 0 / 4	2.28	1.01	0.75	0 / 8	0
P13 Zhang 2018 (bike)	30	-	1.67	1.02	1.09	4 / 10	2
P14 Zhang 2018	32	15 / 4 / 32	2.19	1.12	1.05	10 / 29	3
P15 Zhang 2018 (ride)	30	-	2.18	1.09	1.06	9 / 27	1
P18 Jiao 2020	32	30 / 12 / 32	2.12	1.04	1.15	7 / 26	3
P20 Shirgaokar 2021	56	8 / 6 / 14	2.34	1.16	0.94	12 / 50	3

A concrete example is the negative binomial study of shared mobility and trips, whose conclusion is that ride-hailing use is a significant predictor of daily trips. The unweighted fit reproduces the printed coefficient (0.44 per cent more trips per additional use in the past month on weekdays) and gives a p-value < 0.001. The weighted estimate is smaller (0.33 per cent, a change of −24 per cent); its household-clustered p-value is = 0.045 and the replicate-weight standard error of 0.0024 (against 0.0008 unweighted) gives a p-value = 0.17; on weekends the replicate p-value is = 0.11 (Figure 5, reference letter R). Significance, on which the conclusion rests, therefore depends on both the weighting and the variance estimator. A second example is the effect of being a driver on the count of rides in the zero-inflated model: the estimate changes from −0.43 to −0.19, and the p-value from < 0.001 (unweighted) to = 0.015 (weighted, clustered) and = 0.14 (replicate weights).

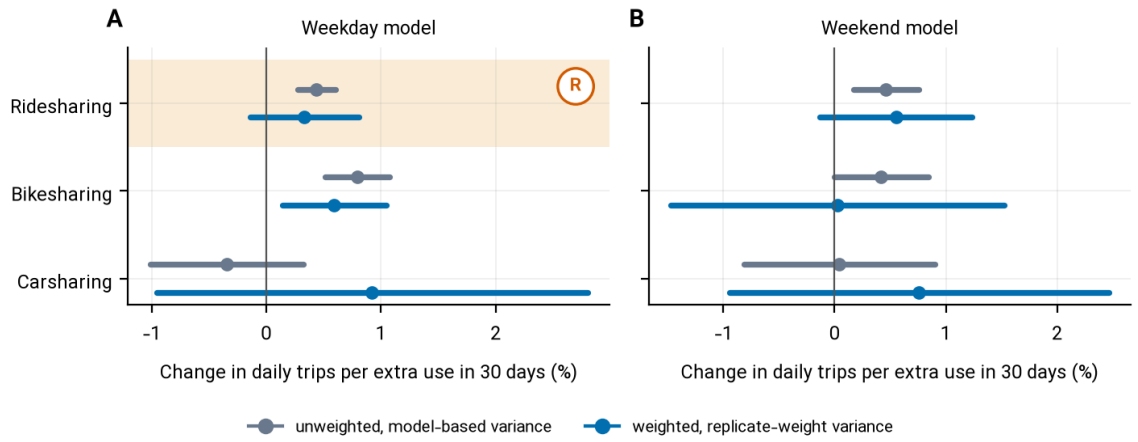


Figure 5: Shared-mobility terms of the weekday (panel A) and weekend (panel B) trip models of Jiao and colleagues, with ninety-five per cent intervals from the unweighted and the design-aware fit. Reference letter R marks the ridesharing row discussed in the text.

Which layer of the design matters

The decomposition in Figure 4 (panel A) says that the weights drive the difference and that, on average, the household clustering and the replicate formula add little. This is not a licence to skip the replicate formula. Per study, the median ratio of the replicate-weight to the household-clustered standard error ranged from 0.75 to 1.15, and for individual coefficients from 0.70 to 1.44 (fifth to ninety-fifth percentile); of the 182 coefficients significant under the weighted household-clustered fit, 16 (9 per cent) lost significance under the replicate variance, including the ride-hailing coefficient above. Where replicate weights exist they should be used; the household-clustered weighted fit is a reasonable fallback for the 2022 wave, where the design is expressed through strata and clusters, but it is an approximation that fails in some of exactly the cases that matter.

Models without a printed coefficient table

The structural equation model of online shopping and shopping trips, which prints standardised coefficients, was fitted with the same variables except two that come from other sources, on 106,136 persons against the printed 85,663. Its two cross-effects were −0.08 and 0.17 in our weighted fit against 0.12 and 0.21 printed, with design standard errors of about 0.02. One cross-effect has the opposite sign and the other a different size, and the sign convention of the latent variables cannot repair both; we grade the model as not reproduced. For the propensity-score matching study of ride-hailing and vehicle miles, a greedy nearest-neighbour matching in the same strata gave 2,664 matched triples of persons with household vehicle access against about five thousand printed. Matched mean daily miles differed from the printed means (for frequent users 29.8 against 24.8 miles) and only one of the three vehicle contrasts was close to its printed value (non-users to occasional users −0.067 against −0.065); we grade the study as not reproduced. A design-aware comparison of matched samples is not straightforward, because weighting the two arms separately breaks the pairing and the replicate variance treats the matched sample as

fixed; we report only that none of the three vehicle contrasts has a t statistic above two when the arms are weighted (0.95, 0.00 and 1.62), without comparing it with the paired statistic of the paper (2.45 for the first contrast).

Machine-learning models

Classifiers of alternative-fuel and electric vehicle adoption showed the largest dependence on the pipeline (Figure 6). When the rarer class was oversampled before the rows were split, the area under the curve of the random forest was 0.99 (alternative-fuel vehicles) and 1.00 (electric vehicles) and of the decision tree 0.98 and 0.99. When the split came first, the same models scored 0.64 and 0.73 for the random forest and 0.52 and 0.54 for the tree; splitting by household changed this little (0.63 and 0.71), as expected because only 20 per cent of the rows share a household with another row, so the leak that matters here is the oversampling. Weighted by the survey weight, the random forest scored 0.65 and 0.69. The inflation of the random forest was 0.36 (alternative-fuel) and 0.29 (electric); the decision tree, which scored 0.98 in the leaky pipeline, is barely better than chance (0.54) when evaluated honestly.

The printed values for alternative-fuel vehicles, 0.92 for the random forest and 0.91 for the decision tree, lie between the leaky and the honest scores and are reproduced by neither (Figure 6, reference letter L); our sample is also smaller than the printed one. The printed accuracies, above ninety per cent for all five models, are consistent with an honest test set of a rare class, where predicting the majority class gives 0.96, and not with the leaky pipeline, where accuracy was 0.64 for logistic regression, 0.66 for naive Bayes and 0.83 for the support vector machine. We therefore do not claim that the published pipeline leaked; we show that oversampling before splitting inflates the score of three of five models by more than fifteen percentage points (the random forest by 0.36, the decision tree by 0.44 and the support vector machine by 0.30) and of the other two, logistic regression and naive Bayes, by 0.02 and 0.11. The fourth hypothesis, as originally worded (the printed score is reproduced only in a leaky pipeline), is not supported; the inflation it rests on is. Tuning all models alike did not remove the gap. The tuned random forest scored 0.99 when oversampling came first and 0.66 by household (an inflation of 0.33; for electric vehicles 0.23), and the tuned decision tree rose from 0.54 to 0.63 but stayed far below its leaky score of 0.98. The largest standard deviation of a mean score over the seeds was 0.033 for the default settings and 0.039 (0.075 for the weighted score) for the tuned models.

Table 5: Area under the ROC curve on held-out rows (mean of three seeds for default settings, two for tuned models). Leaky: oversample, then split rows. Rows: split rows, then oversample. Households: split households, then oversample. Weighted: split households, no oversampling, survey-weighted score.

Outcome	Model	Leaky	Rows	House-holds	Weighted	Tuned, leaky	Tuned, house-holds
Alterna-tive-fuel	LR	0.69	0.67	0.67	0.66	0.69	0.67
Alterna-tive-fuel	NB	0.72	0.61	0.62	0.63	0.73	0.62
Alterna-tive-fuel	DT	0.98	0.52	0.54	0.52	0.98	0.63
Alterna-tive-fuel	RF	0.99	0.64	0.63	0.65	0.99	0.66
Alterna-tive-fuel	SVM	0.91	0.60	0.62	0.56	0.97	0.65
Electric	LR	0.78	0.75	0.77	0.77	0.78	0.77
Electric	NB	0.81	0.73	0.73	0.73	0.81	0.74
Electric	DT	0.99	0.54	0.53	0.53	0.99	0.69
Electric	RF	1.00	0.73	0.71	0.69	1.00	0.77
Electric	SVM	0.97	0.70	0.70	0.63	1.00	0.76

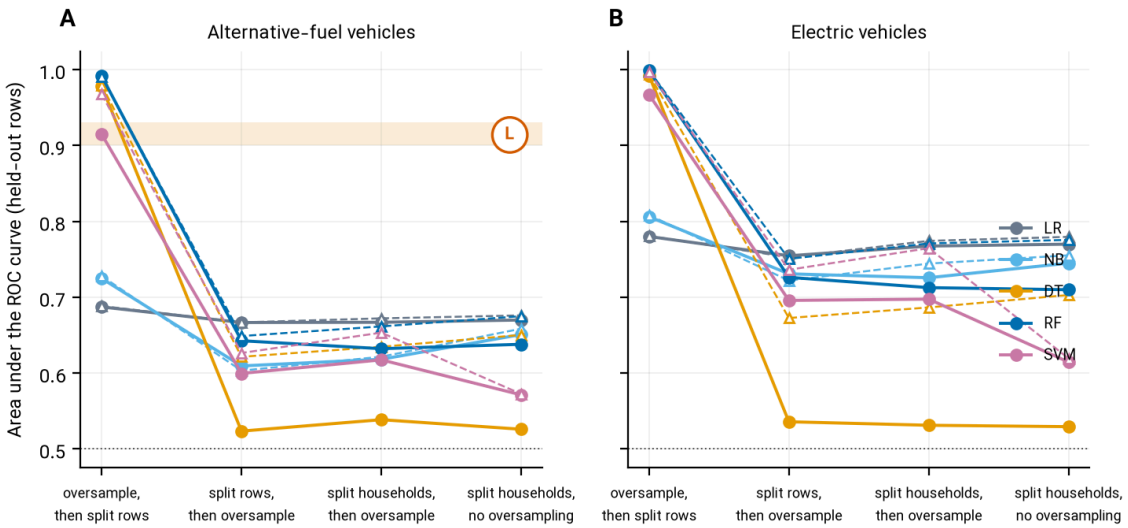


Figure 6: Area under the ROC curve of five classifiers under four pipelines, for alternative-fuel vehicles (panel A) and electric vehicles (panel B). Filled circles: default settings of the first experiment; open triangles: all models tuned alike. Reference letter L marks the printed random-forest score of the study.

The regression study of household online shopping printed held-out R-squared values of 0.71 for gradient boosting in 2017 and 0.72 in 2009. Our boosting model on the public variables that correspond to the printed list reached 0.20 and 0.08 on randomly held-out households and 0.18 when whole states were held out (Figure 7). The validation design does not explain the gap: even on its training households the model explained only 0.29 in 2017, and our linear regression scored 0.20 against 0.11 printed, so the target or the variables differ. The printed variable table contains percentages above one hundred and a variable listed twice, so we cannot say which variables entered the printed model, and we make no claim about its correctness or about leakage.

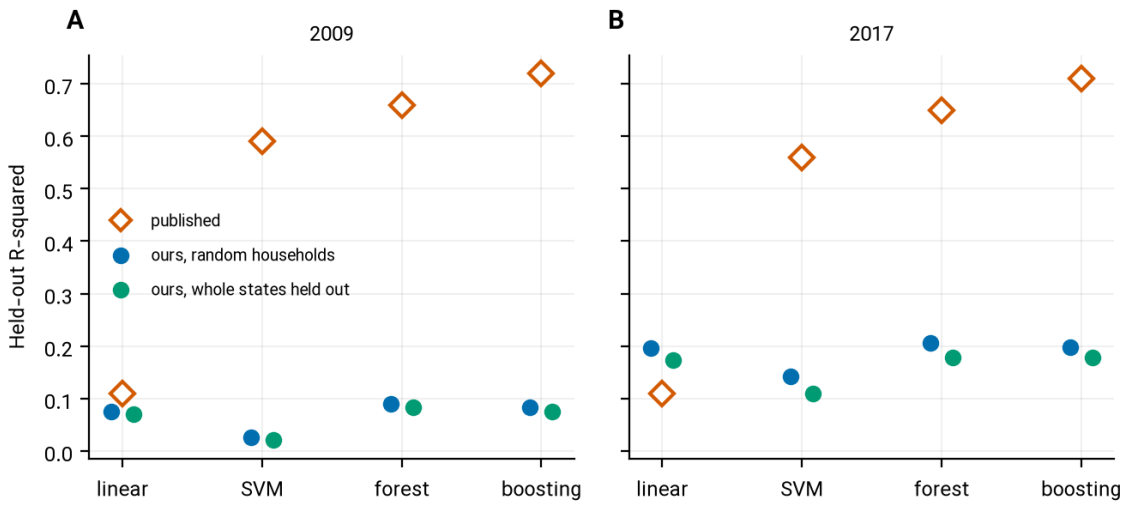


Figure 7: Held-out R-squared of four regression models for household online purchases in 2009 (panel A) and 2017 (panel B): printed values and our values for random and for whole-state hold-out.

Discussion

What the ladder says. For a typical NHTS regression, the printed numbers are recoverable from the public files, provided the reader knows whether the authors weighted, which the paper often has to be read twice to give; where the two estimators can be told apart, the declared one reproduces the printed numbers. The next rung is where the findings move. The design-aware standard error is a median of 2.2 times the unweighted one, and under the studies' own declared inference 24 per cent of the significant coefficients are not significant under the design. These are not errors in the original papers, most of which state what they did; they show that the choice between an unweighted and a weighted, design-aware analysis has consequences that the articles rarely discuss.

What the standard errors mean. The unweighted analyses estimate an association among respondents, and their standard errors are nearly right for that estimand (understated by a median factor of only 1.10). The design-aware standard errors are large because estimating the population association with unequal weights is less precise. A planner who wants to project demand needs the population association and therefore the weighted fit with its honest uncertainty; a researcher interested in respondents may legitimately stay unweighted, but should say so. In the ride-hailing example, the coefficient is small under both fits, its significance, on which the headline rests, depends on the estimator, and the weighted estimate is a quarter smaller than the unweighted one. For the vehicle-miles matching study, our matched samples do not reproduce the printed contrasts, so we make no statement about vehicle shedding.

Why clustering matters little here. Persons of a household share location, income and vehicles, but the outcomes studied (trip counts, mode use) are only weakly correlated within households once the covariates are in the model, and many rows have no other member in the sample. The

effect of weights on the standard error is much larger because the NHTS weights vary strongly across persons. Replicate weights add the stratification and the unequal selection and change individual verdicts, so they should be used when available.

The prediction scores. Oversampling before splitting inflates the score of a decision tree, a random forest and a support vector machine by more than fifteen percentage points, a leak documented in other fields (Demircioğlu, 2024, Kapoor and Narayanan, 2023). The printed scores are not reproduced by either pipeline and the printed accuracies argue against the leaky one, so the result is a warning, not a verdict on the original studies. A reader can check any pipeline with `pynhts.validate.audit_classifier`, which takes a few seconds. When the data are split before oversampling, a random forest of alternative-fuel adoption has an area under the curve of 0.63, about that of logistic regression (0.67), which is modest and worth knowing before using it to target adoption policy. The regression-score gap of the online shopping study is of another kind: the validation design does not explain it.

Implications for the package. The study required tools that did not exist in the package: a regression that returns weighted estimates with the wave's variance design, a check of prediction pipelines for leakage, and a way to compare a printed number with a rebuilt one. The first two are released in pynhts as `pynhts.models` (now with the unweighted, clustered, weighted and design-aware fits and a warning when the verdicts differ) and `pynhts.validate`, validated against R for the replicate-weight and the Taylor designs, and the rules of the package's agent guide say to use them. Three further developments follow from the evidence and are planned: a formula interface and ordinal and multinomial families for the discrete-choice papers that the present set does not include; household-level multilevel models with partial pooling for regional comparisons; and a library of replication recipes, one per published study, so that the registered gaps become regression tests that any user can run.

Limitations. The case-study set is a convenience set of 22 studies, and 13 of them were refitted; the share of fragile conclusions among all NHTS papers is therefore not estimated. The pooled census mixes studies of different size; the declared-inference version restricts the comparison to what each study did. Two zero-inflated studies could not be graded because their printed headline cannot be mapped onto a coefficient; they contribute design effects only. Our fits reproduce some printed values only approximately because the published samples cannot be rebuilt exactly. The matching study used a greedy matching, not the optimal matching of the original. The machine-learning experiments use five classifiers with ten tuning trials each, a modest search that makes the comparison fair but not exhaustive, and score person rows with the household weight. We did not test whether the weights are informative about the outcomes, which would be the formal basis for choosing between the weighted and the unweighted estimand. The 2022 wave, which has no replicate weights, would use linearisation; it is implemented and checked against R but not exercised on the case studies. Finally, we describe conclusions as design-sensitive; we do not say they are wrong.

Recommendations. For authors: state whether weights were used and why, and for which population the coefficients are meant; report standard errors from the survey design or, when only strata and clusters are available, with household-clustered errors; and give the sample-flow

table with the number of records removed by each rule. For prediction models: split by household before anything else, oversample only the training rows, report a weighted score, and tune every model alike. For reviewers: a significance star from an unweighted NHTS regression is a statement about respondents, and an area under the curve from a random split of oversampled rows should trigger a question.

Conclusions

A four-rung reproducibility ladder with design-aware re-estimation turns a general worry about survey-based transport research into numbers. On 22 published NHTS studies, the printed numbers were recoverable with the declared estimator where the estimators can be told apart; design-aware standard errors were about 2.2 times the unweighted ones, almost entirely because of the unequal weights; 24 per cent of the coefficients that were significant under the studies' own inference were not significant under the design; and oversampling before splitting inflated the scores of classifiers of vehicle adoption by up to 0.44 area under the curve, while the printed scores were reproduced by neither pipeline. The machinery is released as `pynhts.models` and `pynhts.validate`, so that other researchers can run the same check on their own analyses.

Declarations

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process. During the preparation of this work the author used Claude (Anthropic) as an assistant for study planning, writing and running analysis code, literature screening, and drafting and editing the text. After using this tool, the author reviewed, verified and edited the material (code, results, references and text) and takes full responsibility for the content of the publication. Numbers in the text are injected by code from the result files, and the references were checked against bibliographic databases.

Competing interests. The author developed the pynhts package whose modules are described here and is the author of one of the case-study papers; that paper was included in the register and not re-estimated. [[AUTHOR-TODO: confirm the standard competing-interests sentence and any co-authors.]]

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. [[AUTHOR-TODO: confirm.]]

Data and code availability. The NHTS public-use files are available from the survey's website. The code, result tables and the package are in the open repository of pynhts (https://github.com/sabernaseralavi-60/pynhts), archived with a persistent identifier on Zenodo.

Supplementary material

Deviations from the analysis plan (the full log is in `analysis_plan.md`): three studies were run before the plan was frozen; the grading tolerance had a floor of two units of the last printed digit; hypothesis three was evaluated against each study's declared inference (the first draft pooled against the unweighted fit); hypothesis four was reformulated after the printed score proved reproducible by neither pipeline; the tuned run used two seeds; a decomposition of the standard-

error ratio, an unweighted household-clustered fit and a check of Newton against exact replicate refits were added after an external review; the printed targets of two zero-inflated studies were dropped because they cannot be mapped onto coefficients.

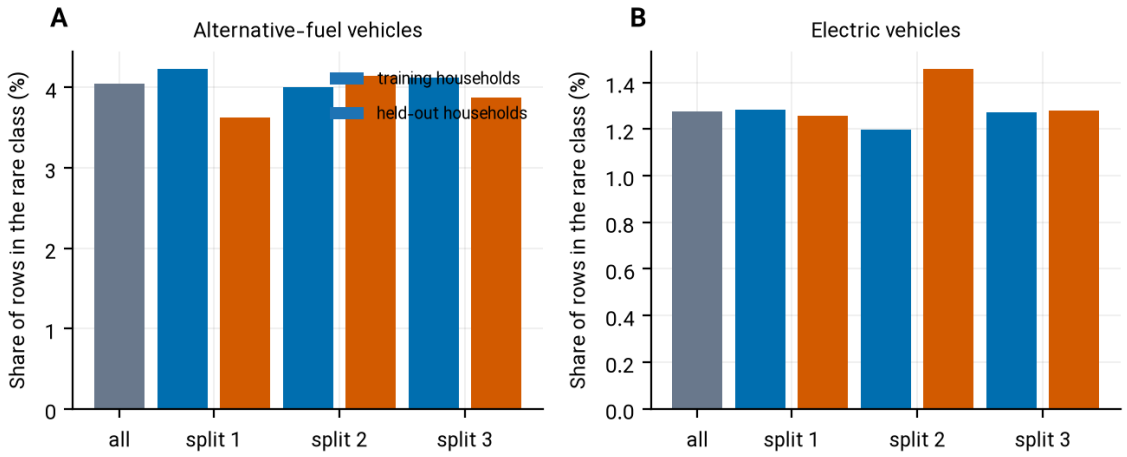


Figure 8: Frequency histogram of the rare class (share of alternative-fuel vehicles, panel A, and of electric vehicles, panel B) in all rows, in the training households and in the held-out households of each of the three splits.

Table 6: Selected hyper-parameters of the tuned models in the household-split pipeline (first master seed). C: cost or penalty strength; vs: variance floor; depth, leaf, crit: tree depth, minimum leaf size, split criterion; n, mf: number of trees and share of inputs per split; gamma: kernel width.

Outcome	Model	Selected values	Inner validation AUC
Alternative-fuel	LR	C = 0.0256, penalty_l1 = False	0.67
Alternative-fuel	NB	vs = 1.79e-09	0.63
Alternative-fuel	DT	depth = 3, leaf = 50, crit = entropy	0.63
Alternative-fuel	RF	n = 200, depth = 10, leaf = 1, mf = 0.3	0.65
Alternative-fuel	SVM	C = 2.59, gamma = 0.00407	0.62
Electric	LR	C = 0.0175, penalty_l1 = True	0.75
Electric	NB	vs = 1.79e-09	0.72
Electric	DT	depth = 3, leaf = 50, crit = entropy	0.68
Electric	RF	n = 200, depth = 10, leaf = 1, mf = 0.3	0.74
Electric	SVM	C = 2.59, gamma = 0.00407	0.72

References

Bibliography

Alberini, A., Teja Burra, L., Cirillo, C., Shen, C., 2021. Counting vehicle miles traveled: What can we learn from the NHTS?. Transportation Research Part D: Transport and Environment 98, 102984.. https://doi.org/10.1016/j.trd.2021.102984

Barua, L., Zou, B., Zhou, Y., Liu, Y., 2021. Modeling household online shopping demand in the U.S.: a machine learning approach and comparative investigation between 2009 and 2017. Transportation 50, 437–476.. https://doi.org/10.1007/s11116-021-10250-z

Birrell, C.L., Steel, D.G., Batterham, M.J., Arya, A., 2019. How to use replicate weights in health survey analysis using the National Nutrition and Physical Activity Survey as an example. Public Health Nutrition 22, 3315–3326.. https://doi.org/10.1017/s1368980019001927

Blagus, R., Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106.. https://doi.org/10.1186/1471-2105-14-106

Bollen, K.A., Biemer, P.P., Karr, A.F., Tueller, S., Berzofsky, M.E., 2016. Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. Annual Review of Statistics and Its Application 3, 375–392.. https://doi.org/10.1146/annurev-statistics-011516-012958

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research 16, 321–357.. https://doi.org/10.1613/jair.953

Conway, M.W., Salon, D., King, D.A., 2018. Trends in Taxi Use and the Advent of Ridehailing, 1995–2017: Evidence from the US National Household Travel Survey. Urban Science 2, 79.. https://doi.org/10.3390/urbansci2030079

Deka, D., Fei, D., 2019. A comparison of the personal and neighborhood characteristics associated with ridesourcing, transit use, and driving with NHTS data. Journal of Transport Geography 76, 24–33.. https://doi.org/10.1016/j.jtrangeo.2019.03.001

Demircioğlu, A., 2024. Applying oversampling before cross-validation will lead to high bias in radiomics. Scientific Reports 14, 11563.. https://doi.org/10.1038/s41598-024-62585-z

Fu, Z., Liu, W., Zhu, J., Yuan, Y., 2020. Trip Chain Simulation of Electric Vehicles based on NHTS Data. 2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC) 1–6.. https://doi.org/10.1109/appeec48164.2020.9220726

Gard, A.M., Hyde, L.W., Heeringa, S.G., West, B.T., Mitchell, C., 2023. Why weight? Analytic approaches for large-scale population neuroscience data. Developmental Cognitive Neuroscience 59, 101196.. https://doi.org/10.1016/j.dcn.2023.101196

Gelman, A., 2007. Struggles with Survey Weighting and Regression Modeling. Statistical Science 22, 153–164.. https://doi.org/10.1214/088342306000000691

Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications 78, 273–282.. https://doi.org/10.1016/j.eswa.2017.01.057

Hardwicke, T.E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M.B., Peloquin, B.N., deMayo, B.E., Long, B., Yoon, E.J., Frank, M.C., 2021. Analytic reproducibility in articles receiving open data badges at the journal Psychological Science : an observational study. Royal Society Open Science 8, 201494.. https://doi.org/10.1098/rsos.201494

Hörl, S., Balac, M., 2021. Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. Transportation Research Part C: Emerging Technologies 130, 103291.. https://doi.org/10.1016/j.trc.2021.103291

Jia, J., 2019. Analysis of Alternative Fuel Vehicle (AFV) Adoption Utilizing Different Machine Learning Methods: A Case Study of 2017 NHTS. IEEE Access 7, 112726–112735.. https://doi.org/10.1109/ACCESS.2019.2934780

Jia, J., Shi, B., Che, F., Zhang, H., 2020. Predicting the Regional Adoption of Electric Vehicle (EV) With Comprehensive Models. IEEE Access 8, 147275–147285.. https://doi.org/10.1109/ACCESS.2020.3014851

Jiao, J., Bischak, C., Hyden, S., 2020. The impact of shared mobility on trip generation behavior in the US: Findings from the 2017 National Household Travel Survey. Travel Behaviour and Society 19, 1–7.. https://doi.org/10.1016/j.tbs.2019.11.001

Kapoor, S., Cantrell, E.M., Peng, K., Pham, T.H., Bail, C.A., Gundersen, O.E., Hofman, J.M., Hullman, J., Lones, M.A., Malik, M.M., Nanayakkara, P., Poldrack, R.A., Raji, I.D., Roberts, M., Salganik, M.J., Serra-Garcia, M., Stewart, B.M., Vandewiele, G., Narayanan, A., 2024. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. Science Advances 10, eadk3452.. https://doi.org/10.1126/sciadv.adk3452

Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4, 100804.. https://doi.org/10.1016/j.patter.2023.100804

Krizek, K.J., McGuckin, N., 2019. Shedding NHTS Light on the Use of 'Little Vehicles' in Urban Areas. Transport Findings.. https://doi.org/10.32866/10777

Lambert, D., 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics 34, 1.. https://doi.org/10.2307/1269547

Lidbe, A., Adanu, E.K., Penmetsa, P., Jones, S., 2021. Changes in the travel patterns of older Americans with medical conditions: A comparison of 2001 and 2017 NHTS data. Transportation Research Interdisciplinary Perspectives 11, 100463.. https://doi.org/10.1016/j.trip.2021.100463

Lumley, T., 2004. Analysis of Complex Survey Samples. Journal of Statistical Software 9.. https://doi.org/10.18637/jss.v009.i08

Martín-Baos, J.Á., López-Gómez, J.A., Rodriguez-Benitez, L., Hillel, T., García-Ródenas, R., 2023. A prediction and behavioural analysis of machine learning methods for modelling travel mode choice. Transportation Research Part C: Emerging Technologies 156, 104318.. https://doi.org/10.1016/j.trc.2023.104318

Naseralavi, S., Soltanirad, M., Ranjbar, E., Lucero, M., Gorzin, F., Hakiminejad, Y., Azimi, S., Baghersad, M., Mazaheri, A., 2025. Machine Learning Modeling of Household Trip Generation by State Using NHTS Data. Urban Science 9, 353.. https://doi.org/10.3390/urbansci9090353

Peng, R.D., 2011. Reproducible Research in Computational Science. Science 334, 1226–1227.. https://doi.org/10.1126/science.1213847

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications 11, 4540.. https://doi.org/10.1038/s41467-020-18321-y

Quinn, T.D., Jakicic, J.M., Fertman, C.I., Barone Gibbs, B., 2016. Demographic factors, workplace factors and active transportation use in the USA: a secondary analysis of 2009 NHTS data.

Journal of Epidemiology and Community Health 71, 480–486.. https://doi.org/10.1136/jech-2016-207820

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.. https://doi.org/10.1111/ecog.02881

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., Scheinost, D., 2024. Data leakage inflates prediction performance in connectome-based machine learning models. Nature Communications 15, 1829.. https://doi.org/10.1038/s41467-024-46150-w

Rust, K., Rao, J., 1996. Variance estimation for complex surveys using replication techniques. Statistical Methods in Medical Research 5, 283–310.. https://doi.org/10.1177/096228029600500305

Sabouri, S., Brewer, S., Ewing, R., 2020. Exploring the relationship between ride-sourcing services and vehicle ownership, using both inferential and machine learning approaches. Landscape and Urban Planning 198, 103797.. https://doi.org/10.1016/j.landurbplan.2020.103797

Sadeghvaziri, E., Tawfik, A., 2020. Using the 2017 National Household Travel Survey Data to Explore the Elderly's Travel Patterns. International Conference on Transportation and Development 2020 86–94.. https://doi.org/10.1061/9780784483169.008

Schweinsberg, M., Feldman, M., Staub, N., Akker, O.R. van den, Aert, R.C. van, Assen, M.A. van, Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S.M., Robinson, D., Madan, N., Silberzahn, R., Goldstein, P., Tierney, W., Murase, T., Mandl, B., Viganola, D., Strobl, C., Schaumans, C.B., Kelchtermans, S., Naseeb, C., Mason Garrison, S., Yarkoni, T., Richard Chan, C., Adie, P., Alaburda, P., Albers, C., Alspaugh, S., Alstott, J., Nelson, A.A., Rubia, E. Ariño de la, 2021. Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. Organizational Behavior and Human Decision Processes 165, 228–249.. https://doi.org/10.1016/j.obhdp.2021.02.003

Shirgaokar, M., Nobler, E., 2021. Differences in daily trips between immigrants and US-born individuals: Implications for social integration. Transport Policy 105, 103–114.. https://doi.org/10.1016/j.tranpol.2021.03.008

Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M.A., Dalla Rosa, A., Dam, L., Evans, M.H., Flores Cervantes, I., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T.J., Hederos, K., Heene, M., Hofelich Mohr, A.J., Högden, F., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D.M., Lei, R., Lindsay, T.A., Liverani, S., Madan, C.R., Molden, D., Molleman, E., Morey, R.D., 2018. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science 1, 337–356.. https://doi.org/10.1177/2515245917747646

Simonsohn, U., Simmons, J.P., Nelson, L.D., 2020. Specification curve analysis. Nature Human Behaviour 4, 1208–1214.. https://doi.org/10.1038/s41562-020-0912-z

Solon, G., Haider, S.J., Wooldridge, J.M., 2015. What Are We Weighting For?. Journal of Human Resources 50, 301–316.. https://doi.org/10.3368/jhr.50.2.301

Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W., 2016. Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science 11, 702–712.. https://doi.org/10.1177/1745691616658637

Stodden, V., Seiler, J., Ma, Z., 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences 115, 2584–2589.. https://doi.org/10.1073/pnas.1708290115

Su, R., Xiao, J., McBride, E.C., Goulias, K.G., 2021. Understanding senior's daily mobility patterns in California using human mobility motifs. Journal of Transport Geography 94, 103117.. https://doi.org/10.1016/j.jtrangeo.2021.103117

Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. PLOS ONE 14, e224365.. https://doi.org/10.1371/journal.pone.0224365

Wang, S., Mo, B., Zheng, Y., Hess, S., Zhao, J., 2024. Comparing hundreds of machine learning and discrete choice models for travel demand modeling: An empirical benchmark. Transportation Research Part B: Methodological 190, 103061.. https://doi.org/10.1016/j.trb.2024.103061

Wang, X., Shaw, F.A., Mokhtarian, P.L., Watkins, K.E., 2022. Response willingness in consecutive travel surveys: an investigation based on the National Household Travel Survey using a sample selection model. Transportation 50, 2339–2373.. https://doi.org/10.1007/s11116-022-10312-w

Winship, C., Radbill, L., 1994. Sampling Weights and Regression Analysis. Sociological Methods & Research 23, 230–257.. https://doi.org/10.1177/0049124194023002004

Wolter, K.M., 2007. Introduction to Variance Estimation. Statistics for Social, Behavioral Sciences.. https://doi.org/10.1007/978-0-387-35099-8

Wu, X., MacKenzie, D., 2021. Assessing the VMT effect of ridesourcing services in the US. Transportation Research Part D: Transport and Environment 94, 102816.. https://doi.org/10.1016/j.trd.2021.102816

Xue, C., Wu, Q., Sun, M., Bai, P., Chen, Y., 2021. The Interaction between E-Shopping and Shopping Trips: An Analysis with 2017 NHTS. Complexity 2021, 8247158.. https://doi.org/10.1155/2021/8247158

Yang, Y., Diez Roux, A.V., Bingham, C.R., 2011. Variability and seasonality of active transportation in USA: evidence from the 2001 NHTS. International Journal of Behavioral Nutrition and Physical Activity 8, 96.. https://doi.org/10.1186/1479-5868-8-96

Zeileis, A., Kleiber, C., Jackman, S., 2008. Regression Models for Count Data in R. Journal of Statistical Software 27.. https://doi.org/10.18637/jss.v027.i08

Zhang, Y., Zhang, Y., 2018a. Associations between Public Transit Usage and Bikesharing Behaviors in The United States. Sustainability 10, 1868.. https://doi.org/10.3390/su10061868

Zhang, Y., Zhang, Y., 2018b. Examining the Relationship between Household Vehicle Ownership and Ridesharing Behaviors in the United States. Sustainability 10, 2720.. https://doi.org/10.3390/su10082720

Zhang, Y., Zhang, Y., 2018c. Exploring the Relationship between Ridesharing and Public Transit Use in the United States. International Journal of Environmental Research and Public Health 15, 1763.. https://doi.org/10.3390/ijerph15081763

Zhou, Y., Wang, X.(., 2014. Explore the relationship between online shopping and shopping trips: An analysis with the 2009 NHTS data. Transportation Research Part A: Policy and Practice 70, 1–9.. https://doi.org/10.1016/j.tra.2014.09.014