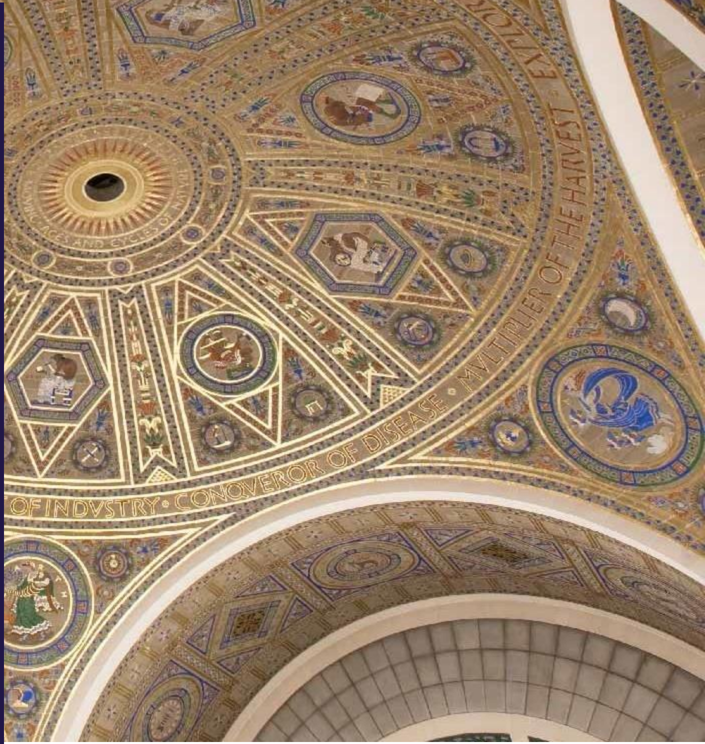
NATIONAL ACADEMIES *Sciences Engineering Medicine*

TRB TRANSPORTATION RESEARCH BOARD

TRB Webinar: Unleashing the Power of Fusion in Travel Demand Analysis

July 12, 2022

12:00 – 1:30 PM



PDH Certification Information

1.5 Professional Development Hours (PDH) – see follow-up email

You must attend the entire webinar.

Questions? Contact Andie Pitchford at TRBwebinar@nas.edu

The Transportation Research Board has met the standards and requirements of the Registered Continuing Education Program. Credit earned on completion of this program will be reported to RCEP at RCEP.net. A certificate of completion will be issued to each participant. As such, it does not include content that may be deemed or construed to be an approval or endorsement by the RCEP.

ENGINEERING RCEP

REGISTERED CONTINUING EDUCATION PROGRAM

AICP Credit Information

1.5 American Institute of Certified Planners Certification Maintenance Credits

You must attend the entire webinar

Log into the American Planning Association website to claim your credits

Contact AICP, not TRB, with questions

Purpose Statement

This webinar will showcase three state-of-the-art data fusion innovations for travel demand modeling. Presenters will address travel choice predictions by inferring missing alternative attributes, understanding electric vehicles (EVs) adoption patterns, and evaluating the impacts of the pandemic on passenger travel demands.

Learning Objectives

At the end of this webinar, you will be able to:

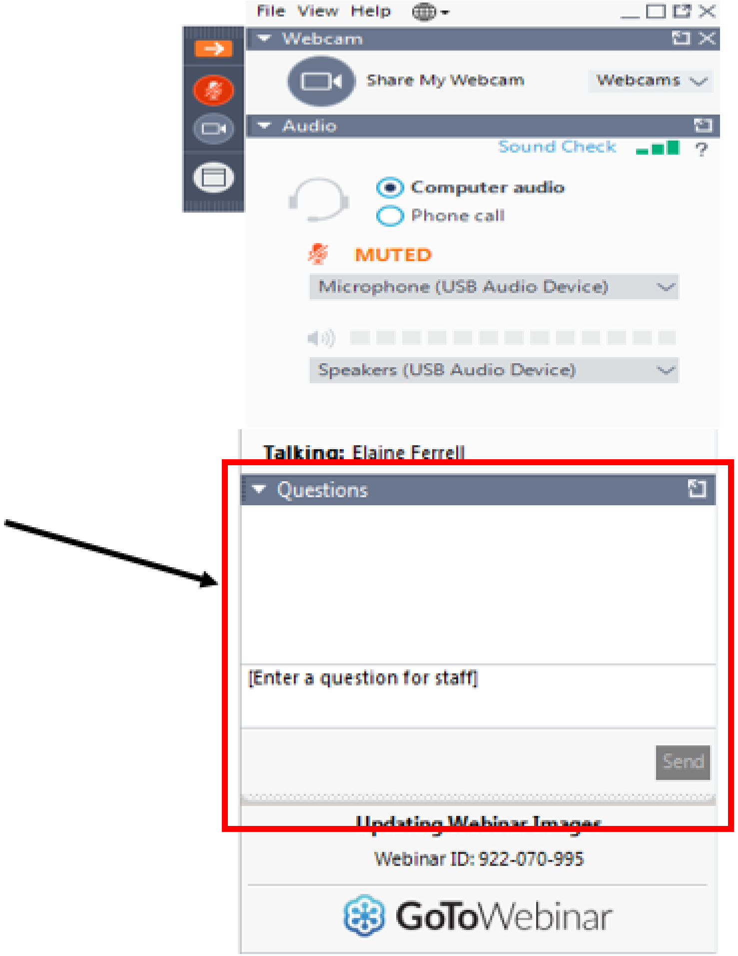
(1) Identify data sources for potential fusion

(2) Pool data efficiently for fusion exercises

(3) Determine the appropriate fusion methodology for modeling

Questions and Answers

- Please type your questions into your webinar control panel
- We will read your questions out loud, and answer as many as time allows



Today's Presenters

Abdul Rawoof Pinjari
abdul@iisc.ac.in
Indian Institute of Science

Naveen Eluru
naveen.eluru@ucf.edu
University of Central Florida

Sanjana Hossain
sanjana.hossain@mail.utoronto.ca
University of Toronto and
Northwestern University

Khandker Nurul Habib
khandker.nurulhabib@utoronto.ca
University of Toronto

Pooling Revealed Preference (RP) and Stated Preference (SP) Choice Data to Infer Missing Attributes

Abdul Rawoof Pinjari

Indian Institute of Science

Acknowledgements

- Co-investigators
  - Mehek Biswas (Indian Institute of Science)
  - Najeebul Feroz Malik (Indian Institute of Science)
  - Chandra Bhat (University of Texas at Austin)

Introduction: Random Utility Maximization (RUM)-based Choice Models

- RUM-based choice models: Workhorse for analysis and prediction of traveller choice and travel behaviour (e.g., mode choice, destination choice, route choice,…)

Utility function: $U_{qj} = \boldsymbol{\beta}'\boldsymbol{X}_{qj} + \varepsilon_{qj},\ \forall j \in \text{choice set } \mathbf{J}$

Alternative 'i' is chosen if it offers maximum utility

- What are the different sources of variability in choice models?
  - Assumptions:
    - Do people always follow utility maximization?
    - Are the choice sets, utility functions, and the stochastic structure (ε_{qj}) correctly specified?
  - The parameters ($\boldsymbol{\beta}$) vary in the population (random parameters)
  - How about uncertainty (and/or unavailability) of some explanatory variables ($\boldsymbol{X}_{qj}$) ?

Motivation: Missing data on alternative attributes in RUM Choice Models

- The analyst may not have complete and accurate information of all attributes ($\boldsymbol{X}_{qj}$) of choice alternatives (for one or more of the following reasons)
  - Omitted variables (e.g., transit crowding levels omitted from a mode choice model)
  - Missing data (e.g., crowding levels in transit are often not reported by non-transit users in surveys)
  - Analyst's errors in measuring $\boldsymbol{X}_{qj}$ (e.g., noisy travel time data, spatial aggregation/TAZs,…)
  - Travelers' perceptions of $\boldsymbol{X}_{qj}$ (which the analyst may not have accurate measurements of)
  - Inherent stochasticity (day-to-day variability in travel times)

- Mis-measured or unavailable attributes in $\boldsymbol{X}_{qj}$ should be specified as stochastic variables – to be estimated/inferred

Current Literature

I. Stochastic Variables in Choice Models

Nature of stochasticity in x_{qi}	Source of stochasticity	Existing studies	Modelling approaches
Inherent stochasticity	• Day-to-day variability in travel conditions on the network	Chen et al. (2011), Srinivasan et al. (2014), Biswas et al. (2019),...	• **Errors-in-variables** approach • **Integrated Choice and Latent Variable (ICLV)** models
Measurement errors in the data available to the analyst	• Use of spatially aggregate measures of level-of-service attributes instead of disaggregate measurements • Travel times from free flow speed assumptions (or using model skims) instead of actual measurements	Train (1978), Ortúzar & Ivelic (1987), Bhatta & Larsen (2011), Steimetz & Brownstone (2005), Walker et al. (2010)	
Perception errors of the traveller	• Analyst might not have access to accurate measures of the travelers' perceived travel times that actually influence their choices	Daly & Ortúzar (1990), Varotto et al. (2017)	

II. Unobserved Taste Heterogeneity of Individuals

Approach for incorporating stochasticity in β_q	Existing studies	Modelling approaches
Captured through random coefficients on alternative attributes such as travel time	Revelt and Train, 1998; Mc Fadden and Train, 2000; Bhat, 2001; Bhat, 2003; Hensher and Greene, 2003; Hess, Bierlaire, and Polak, 2005; Cirillo and Axhausen, 2006; Cherchi and Ortuzar, 2008; Milton et al., 2008; Guevara 2013; Bansal et al., 2019,…	Mixed logit
	Daganzo, 1979; Bunch, 1991; Keane 1992; Bhat, 2011; Bhat and Sidharthan, 2011; Dubey et al., 2020,….	Multinomial probit

Current Literature

III. Imputation of Missing Data

Nature of stochasticity in x_{qi}	Source of stochasticity	Existing studies	Modelling approaches
Missing data	• Missing income variable due to high rate of non-response in surveys	Sanko et al. (2014)	Integrated Choice and Latent Variable (ICLV) models
	• Missing parking prices of private-closed alternatives whereas data for private-open alternatives is available	Gopalakrishnan et al. (2020)	Multiple imputation method (Rubin, 1987), Control function approach
	• Missing travel time data beyond certain bounds	Steimetz and Brownstone (2005)	Multiple imputation method (Rubin, 1987) and Errors-in-variables approach
	• Missing salary and wage data	Greenlees et al. (2012)	Multiple imputation method (Rubin, 1987) and Stochastic Censoring

Gaps in Literature

- Despite a large body of literature in each of these areas, few studies attempt to simultaneously recognize and disentangle the two sources of variability
  - *stochasticity in* $\boldsymbol{X}_{qj}$ *– variables for alternative attributes (e.g., travel time, crowding)* AND
  - *unobserved heterogeneity in* $\boldsymbol{\beta}$ *– traveler sensitivity to those variables*

- How to infer inferring missing attributes (e.g., missing travel time or crowding level information) while also recognizing unobserved heterogeneity in response to those attributes?

- Typical choice models on typical choice data do not allow the simultaneous identifiability of…
  - both $\boldsymbol{X}_{qj}$ and $\boldsymbol{\beta}$ (let alone unobserved variability in $\boldsymbol{X}_{qj}$ and/or $\boldsymbol{\beta}$)

Objectives of this Research

- Research question: In mode choice models, can we infer mode-specific travel times, while recognizing variability in those attributes AND unobserved heterogeneity in response to the attributes?
- Method: Combine stated-preference (SP) & revealed-preference (RP) data to estimate both $\boldsymbol{X}_{qj}$ and β (and related stochasticity) using mixed logit models
  - Why does the method work?…
    - **SP data:** Typically, free of variability in $\boldsymbol{X}_{qj}$ values presented to the decision-maker. This makes heterogeneity in β an important source of variability in stated choices
    - **RP data:** Various sources of variability in $\boldsymbol{X}_{qj}$ (also, missing $\boldsymbol{X}_{qj}$) have a bearing on revealed choice data, as data are collected from the "field"
  - Scope:
    - Theoretical examination of identifiability
    - Simulation experiments
    - Empirical study

Integrated Mixed Logit Model for Pooled RP-SP Data

Utility function for alternative i on an SP choice occasion t for individual q:

carefully constructed travel time data; assumed to be free of variability

$$U_{qit} = \beta_{0qi} + \beta_{TT,q} TT_{qit} + \boldsymbol{\varphi}'\mathbf{z}_{qit} + \varepsilon_{qit}$$

random coefficient on travel time

Utility function for alternative i for an RP choice occasion t:

$$\tilde{U}_{qit} = \left(\tilde{\beta}_{0qi} + \beta_{TT,q} TT^{*}_{qit} + \boldsymbol{\varphi}'\mathbf{z}_{qit} + \tilde{\varepsilon}_{qit}\right) \times \lambda$$

stochastic travel time (to be estimated)

$TT^{*}_{qi} = d_{qi} \times \theta_{qi}$

d_{qi} = Trip distance for individual q

θ_{qi} = Inverse speed, or time required to traverse 1 kilometer (min/km), assumed random

$\boldsymbol{Z}_{qit}$ = Variables other than travel time (considered deterministic)

λ = Ratio of error term scales between SP and RP utilities

A Thought Experiment for Theoretical Analysis

- Assume, for now, that we estimate the model with only SP data
  - One can estimate the random coefficient ($\beta_{TT,q}$) on travel time (TT_{qit})

- Next, let's turn to RP data
  - We know random coefficient on travel time from SP data. Can we estimate travel time (TT^{*}_{qit}) ?
  - How many mode-specific travel times can be estimate?

- We undertook theoretical explorations of parameter identifiability to answer the above questions (using principles of utility maximization, and rank and order conditions)

- Augmented the above explorations with simulation experiments using joint SP-RP models

Findings from Theoretical Explorations and Simulation Experiments

- Combining RP and SP data helps endogenously identify the distributions of mode-specific travel times in addition to stochasticity in the travel time coefficient – with the following nuances:

  - Of all J choice alternatives available, only J-1 means (for travel times) can be estimated
    - The mean travel time of at least one mode must be known exogenously
    - This is because only utility differences matter in RUM-based discrete choice models

  - J-1 variances (for travel times) can be estimated
    - Depending on the distributional assumptions for travel time, it may or may not be possible to estimate variance in travel time for the Jth mode

  - All of the above is possible even when the kernel error distribution is heteroscedastic

  - Difficult to identify stochasticity of multiple explanatory variables simultaneously (using RP-SP data)

Empirical Case Study: RP-SP Mode Choice Survey in Bengaluru, India

Scenario 1: Which of the following modes would you choose to travel between your routine origin and destination in a **non-pandemic situation**? Assume that all travellers are vaccinated and there is NO risk of COVID infection.

Bus	Metro/ Train	Own Two-wheeler	Ola/ Uber Car
○	○	◉	○
Total travel time spent while inside the vehicle(s)			
1 transfer 61 min	28 min	32 min	30 min
Total travel time spent outside vehicle(s)			
18 min	16 min	0 min	9 min
Possible delay due to traffic congestion or other uncertainties			
... up to 17 min more	... up to 2 min more	... up to 10 min more	... up to 12 min more
Total one-way cost of travel			
Rs. 27	Rs. 38	Rs. 36	Rs. 292
Extent of crowding in the vehicle			
Fully crowded or packed	Fully crowded or packed		
Service type			
Ordinary			

Full set of modes in RP data:

- Bus
- Metro
- Personal Car
- Personal Motorcycle
- Uber/Ola Car
- Uber/Ola Motorcycle
- Auto Rickshaw
- Bicycle/Walk

Sample size: 914 commuters

Empirical Case Study: Estimated Three Pooled RP-SP Choice Models

- Model 1: Pooled RP-SP mode choice model, with the distribution of mode-specific in-vehicle travel times/km estimated.
- Model 2: Pooled RP-SP mode choice model, with only means of mode-specific in-vehicle travel times/km estimated (variability in travel times is ignored).
- Model 3: Pooled RP-SP mode choice model, with exogenously obtained travel times from Google maps or speed assumptions.

- Assumptions
  - In all models: Metro mode travel times are known to the analyst and deterministic in all models
  - In all models: IVTT and OVTT coefficients are normal distributed. Cost coefficient is deterministic
  - In Model 1: mode-specific in-vehicle travel times (IVTT) are power log-normal distributed
  - All models recognize the panel (repeated observations) nature of empirical data – random effects on alternative specific constants

Empirical Results for Pooled RP-SP Models of Commute Mode Choice

Variable description	Model 1 (Means and variances of mode-specific travel times/km estimated)		Model 2 (Only means of mode-specific travel times/km estimated)		$t-stat\ for\ H_0 : \hat{\beta}_{Pooled\ Model\ I} = \hat{\beta}_{Pooled\ Model\ II}$	Model 3 (Exogenously obtained travel times used for RP data)	
	Parameter est.	t-stat.	Parameter est.	t-stat.		Parameter est.	t-stat.
Alternative-specific constant (Bus mode is the base)							
Car	-3.734	-2.64	-2.043	-3.38	1.09	-1.389	-2.22
Two-wheeler	-2.353	-2.14	-1.623	-2.44	0.56	-0.847	-1.35
Auto-rickshaw	-11.525	-2.44	-7.792	-3.33	0.70	-3.300	-3.95
Metro	-0.473	-2.12	-0.180	-2.10	1.22	-0.258	-3.19
Walk	-3.540	-1.15	-3.601	-1.04	0.01	-0.806	-0.40
Level-of-service variables (normally distributed coefficients on IVTT and OVTT/distance)							
Coefficient on IVTT – Mean parameter	**-0.061**	-2.72	-0.033	-3.90	1.28	-0.012	-5.01
Coefficient on IVTT – SD parameter	**0.063**	2.75	-0.046	-3.93	4.23	0.024	5.01
Coefficient on walk travel time	-0.112	-1.72	-0.101	-1.34	0.17	-0.065	-1.07
Coefficient on OVTT by distance – Mean parameter	**-0.910**	-2.63	-0.532	-3.56	1.01	-0.332	-4.03
Coefficient on OVTT by distance – SD parameter	**0.794**	4.07	0.076	2.70	3.63	0.079	2.21
Coefficient on travel cost	-0.009	-1.47	-0.003	-1.43	1.06	-0.004	-2.34
Parameters of variables in the RP data (inverse speeds are considered power lognormally distributed) (min per km)							
Car travel time per km – Mean parameter	2.580	15.36	0.950	11.48	2.35	--	--
Car travel time per km – SD parameter	0.440	1.66	--	--	--	--	--
Two-wheeler travel time per km – Mean parameter	1.460	11.56	0.380	12.33	3.60	--	--
Two-wheeler travel time per km – SD parameter	0.410	1.27	--	--	--	--	--
Auto-rickshaw travel time per km – Mean parameter	2.630	9.33	0.700	11.04	3.39	--	--
Auto-rickshaw travel time per km – SD parameter	0.740	1.27	--	--	--	--	--
Bus travel time per km – Mean parameter	2.820	18.21	0.920	21.12	4.39	--	--
Bus travel time per km – SD parameter	0.980	2.86	--	--		--	--
Metro travel time per km – Mean parameter	1.670	--	1.670	--	--	--	--
Metro travel time per km – SD parameter	0.000	--	--	--	--	--	--

Empirical Results for Pooled RP-SP Models of Commute Mode Choice (*t-stat<1)

Variable description	Model 1 (Means and variances of mode-specific travel times/km estimated)		Model 2 (Only means of mode-specific travel times/km estimated)		$t-stat\ for\ H_0 : \hat{\beta}_{Pooled\ Model\ I} = \hat{\beta}_{Pooled\ Model\ II}$	Model 3 (Exogenously obtained travel times used for RP data)	
	Parameter est.	t-stat.	Parameter est.	t-stat.		Parameter est.	t-stat.
Car	4.352	2.49	2.072	2.15	1.14	-1.877	-2.53
Two-wheeler	1.167	0.31	0.416	1.46	0.19	2.643	5.56
Auto-rickshaw	7.061	2.30	3.222	3.21	1.18	-0.480	-1.44
Bus and Metro	4.645	2.74	2.357	3.95	1.27	4.548	6.21
Walk	5.933	2.73	2.197	3.78	1.66	1.890	4.36
Gender (Female is the base, Car mode is the base)							
Two-wheeler	*	*	*	*	--	0.762	1.45
Auto-rickshaw	-6.060	-2.10	-4.742	-2.81	0.39	-2.149	-3.34
Bus	-3.991	-2.47	-2.201	-3.40	1.03	-1.851	-3.47
Metro	-4.617	-2.48	-2.320	-3.41	1.21	-1.780	-3.35
Walk	-4.497	-1.98	-3.389	-2.12	0.43	-2.751	-2.94
Income (Don't know and high income are the base categories, Car mode is the base)							
Coefficient on medium income- Bus	1.117	1.44	*	*	--	*	*
Coefficient on medium income- Metro	1.395	1.68	0.395	1.45	1.14	0.653	1.15
Coefficient on low income- Bus	2.457	4.11	1.066	3.04	1.11	0.945	1.40
Coefficient on low income- Metro	2.168	4.06	1.318	2.87	0.69	0.887	1.31
Age (Metro mode is the base; Age 19-25 years is the base)							
Coefficient on Age-26-45 years Car	1.910	1.99	1.202	2.30	0.65	0.924	1.51
Coefficient on Age-26-45 years- Two-wheeler	-0.498	-1.57	*	*	--	*	*
Coefficient on Age-26-45 years- Auto-rickshaw	3.514	1.67	2.301	1.95	0.51	1.090	1.46
Coefficient on Age-26-45 years – Bus	0.199	1.51	0.142	1.62	0.36	0.231	2.76
Coefficient on Age->45 years- Car	3.767	2.78	1.305	2.99	1.68	0.618	1.20
Coefficient on Age->45 years- Auto-rickshaw	1.872	1.22	*	*	--	0.778	1.17
Coefficient on Age->45 years- Bus	*	*	*	*	--	-0.176	-2.50
Scale (SP to RP)	0.170	2.591	0.095	3.62	1.04	0.463	5.63

Empirical Results for Pooled RP-SP Models of Commute Mode Choice

Model	Log-likelihood at convergence	AIC & BIC	Mc-Fadden's rho-square	Mean value of IVTT (in INR per hour)	Mean value of OVTT at avg. distance (in INR per hour)
Model 1 Pooled RP-SP mode choice model (with the distribution of travel time estimated)	-1,736.6	AIC: 3,552.5 BIC: 3,745.8	0.344	407	506
Model 2 Pooled RP-SP mode choice model (with only the mean travel times estimated)	-1,856.6	AIC: 3,807.8 BIC: 4,034.2	0.298	660	887
Model 3 Pooled RP-SP mode choice model with exogenously obtained (Google maps) travel times used for RP data	-1,950.2	AIC: 3,983.2 BIC: 4,180.8	0.263	180	415

- The model fit is superior for the proposed model than other two models
  - Inferring travel times appears to be better than using network skims or Google maps
  - Beneficial to recognize variability in the estimated attributes
- Values-of-time are substantially different across the three models

Empirical Results for Pooled RP-SP Models of Commute Mode Choice

Estimated average speeds of modes from the proposed model and other models

Model	Two-wheeler	Car	Auto-rickshaw	Bus
Model 1 Pooled RP-SP mode choice model (with the distribution of travel time/km estimated)	44 kmph	24 kmph	24 kmph	22 kmph
Model 2 Pooled RP-SP mode choice model (with only the mean travel time/km estimated)	40 kmph	27 kmph	30 kmph	24 kmph
Model 3 Pooled RP-SP mode choice model with exogenously obtained travel times for RP used as explanatory variables	34 kmph	25 kmph	22 kmph	21 kmph

Average speeds estimated from the pooled RP-SP model for mode choice are close to the actual travel speeds in Bengaluru

Speed estimates from model 2 not in expected order

Summary and Conclusions

- For RUM-based choice models, the analyst may not always have information on alternative attributes
- It is possible to combine stated- & revealed-preference (SP-RP) data sources to simultaneously identify alternative attributes ($\boldsymbol{X}_{qj}$) and response (β) to attributes, AND unobserved heterogeneity in them
- Simulation experiments and empirical case study of mode choice confirms the above hypothesis
- Mode-specific travel speeds inferred from the proposed method are reasonable in magnitude
- Inferring mode-specific travel times (and variability) using the proposed approach offers better fit than a model that uses exogenous information on travel times

Future Work

- Additional empirical case studies

- Can we infer transit crowding levels by combining RP and SP data? Crowding levels are often missing in RP datasets
  - Non-transit travelers may not respond to transit crowding questions
  - Non-transit travelers may not accurately report their perceptions of crowding levels

- SP attributes might, in some cases, be prone to perception errors of the respondent
  - We assumed that the survey respondents do not distort the attribute values presented to them

- The issue of attribute non-attendance was not considered in the current study.

Thank You!

Selected References

- Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H., & Rao, V. (1994). Combining Revealed and Stated Preferences Data, 5(4), 335–349.
- Bhat, C. R. (2000). Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. *Transportation Science*, 34(2), 228-238. Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35(7), 677-693.
- Biswas, M., Pinjari, A.R., & Dubey, S.K. Travel time variability and route choice: An integrated modelling framework. In 2019 11th International Conference on Communication Systems & Networks (COMSNETS), 2019 January, 737-742.
- Biswas, M., Pinjari, A. R., Bhat. C. R., & Ghosh, S. Choice Models with Stochastic Variables and Random Coefficients, 2023. Working paper. Indian Institute of Science.
- Bhatta, B. P., & Larsen, O. I. (2011). Errors in variables in multinomial choice modeling: A simulation study applied to a multinomial logit model of travel mode choice. *Transport Policy*, 18(2), 326-335.
- Díaz, F., Cantillo, V., Arellana, J., & de Dios Ortúzar, J. (2015). Accounting for stochastic variables in discrete choice models. *Transportation Research Part B: Methodological*. 78, 222-237.
- Revelt, D., & Train, K. (1998). Mixed logit with repeated choices: households' choices of appliance efficiency level. Review of economics and statistics, 80(4), 647-657.
- Sanko, N., Hess, S., Dumont, J., Daly, A. 2014. Contrasting imputation with a latent variable approach to dealing with missing income in choice models. *Journal of Choice Modelling* 12, 47-57.
- Walker, J., Li, J., Srinivasan, S., & Bolduc, D. (2010). Travel demand models in the developing world: Correcting for measurement errors. *Transportation Letters*, 2(4), 231-243.

Fusion by design: Evaluate the impacts of COVID-19 on passenger travel demands

Sanjana Hossain, Patrick Loa, Kaili Wang, Sk. Md. Mashrur, Alireza Dianat, Khandker Nurul Habib

July 12, 2023

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

TDMG
TRAVEL DEMAND MODELLING GROUP

Overview

1. Background
2. Study design
3. Data fusion method
4. Results
5. Application of fused data for choice modelling
6. Conclusion

Background

Attitudes and perceptions are important predictors of travel choices

Traditional household travel surveys rarely include attitudinal questions

Specialized surveys can collect detailed attitudinal data but suffer from limited sample sizes

Need a method to fuse “core” household travel survey with “satellite” surveys that collect attitudinal data

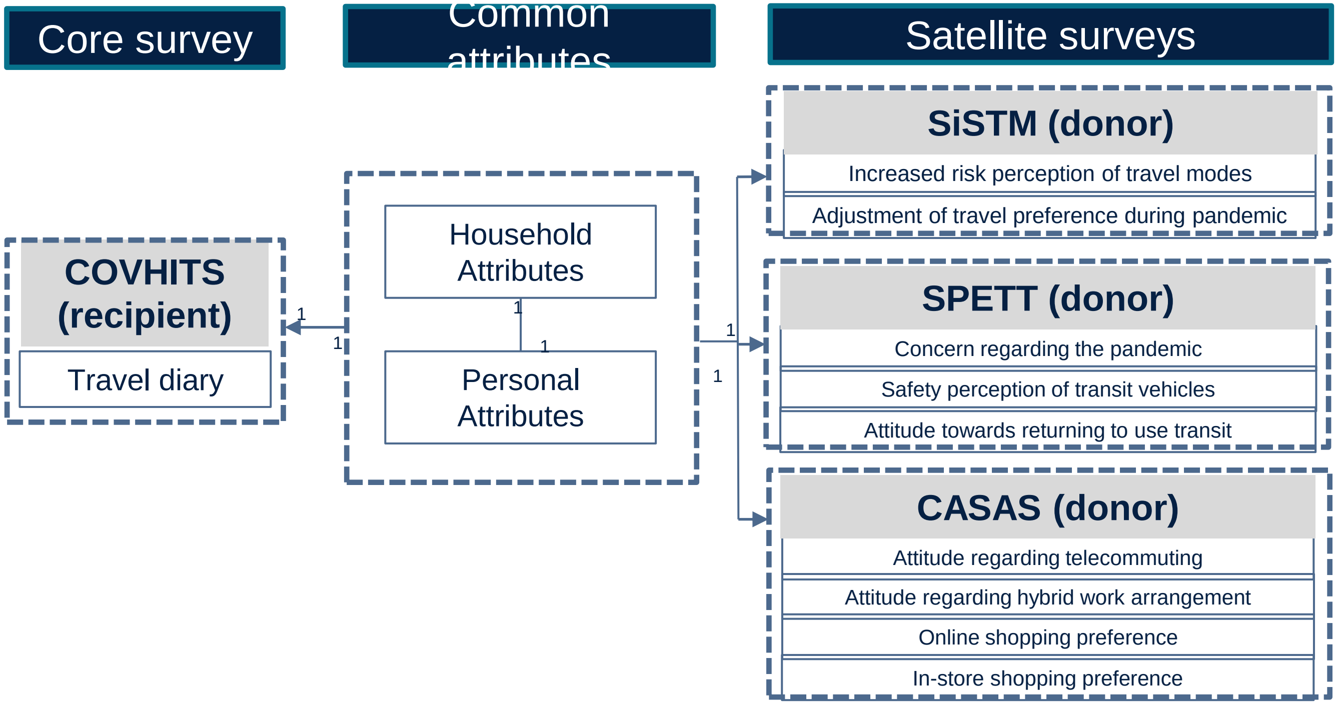
Research Objective

Propose a data fusion method to enrich a “core” household travel survey by linking it to three “satellite” surveys that

Study area: Greater Toronto Area

	Survey	Description	Study period	Sample size
Core	COVHITS	**COV**id-19 influenced **H**ouseholds’ **I**nterrupted **T**ravel **S**chedule	Oct - Nov’ 21	8,911 individuals
Satellites	SiSTM	**S**tudy into the use of **S**hared **T**ravel **M**odes	Jul’ 21	767 individuals
	SPETT	**S**tated **P**reference **E**xperiment on **T**ravel mode and especially **T**ransit choice behavior	Jul’ 21	849 individuals
	CASAS	**C**ovid **A**ctivity **S**cheduling and **A**daptation **S**urvey	Jul’ 21	860 individuals

Core & satellite survey design



Data fusion method

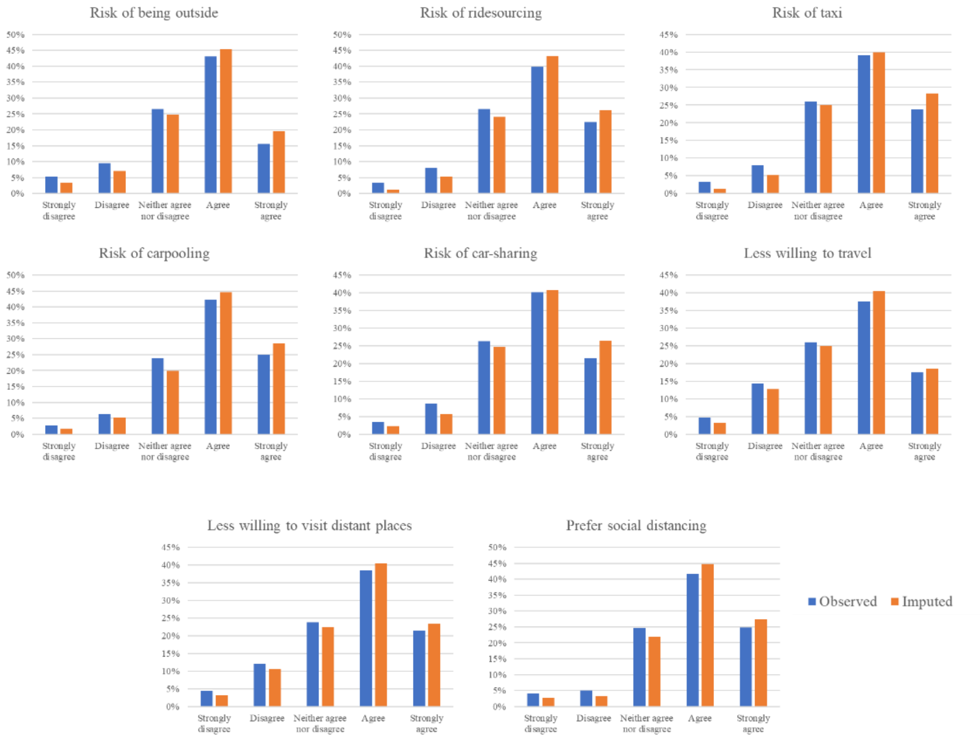
Step 1: Harmonisation and reconciliation of sources
- Make common variables coherent in attribute levels
- Ensure that the samples refer to the same population

Step 2: Analysis of the explanatory power for common variables
- Cramer’s V used to analyze association of common variables with target variables
- Common variables with good explanatory power selected as matching variables

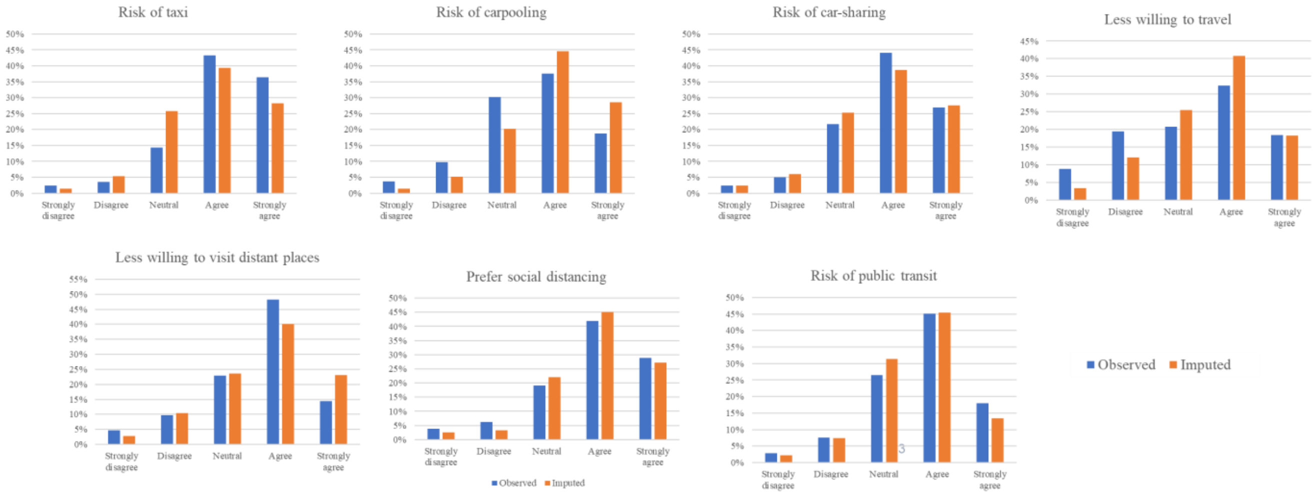
Step 3: Matching method
- k-NN type extension of the hot-deck imputation technique
- Generate multiple instances of fused dataset using Monte Carlo draws from the k NN
- Gower’s dissimilarity coefficient used as the measure of distance

Step 4: Quality evaluation
- Compare marginal distributions of target variables in donor and the fused datasets
- Compare true (observed) values of target variables of the actual recipient units in the core survey with their imputed values

Preservation of marginal distributions in satellite (donor) and fused data



Validation of data fusion outputs



- Fusion replicates general trend of most attitudinal responses
- For some variables like “risk of carpooling” and “less willing to travel”, the distributions are somewhat different, indicating that people are gradually getting used to the pandemic

Data fusion results – socioeconomic attributes

Imputed attitudinal variables meet a priori expectations regarding socioeconomic status of respondents

Age is a significant factor affecting individuals' perception of risks and adjustment to travel during the pandemic

Older respondents have higher risk perception, are more concerned about the pandemic, perceive public transit as less safe, and have a greater preference for in-store shopping

Data fusion results – travel behaviour

Individuals with higher imputed levels of perceived risks made fewer trips

Individuals who agree with the advantages of telecommuting completed fewer work trips per day

Individuals who prefer online grocery shopping made fewer shopping trips, and vice versa

Data fusion results – travel behaviour

Individuals with higher perceived risk of pandemic rely more on driving and avoid public transit

Similarly, individuals who adjusted their travel patterns during this period rely more on driving

Among the different types of transit vehicles, bus/streetcar is perceived to be the least safe

Application of fused data for choice modelling

- Empirical investigation conducted with the synthetic fused data
- Demonstrate how to use the fusion outputs for subsequent modelling
- Hybrid commute mode choice model estimated with the fused data
  - a subset of the travel diary data representing commuting trips from the core survey
  - the socio-demographic information of the respondents
  - their attitudinal statements imputed from the satellite surveys

Hybrid choice model estimation

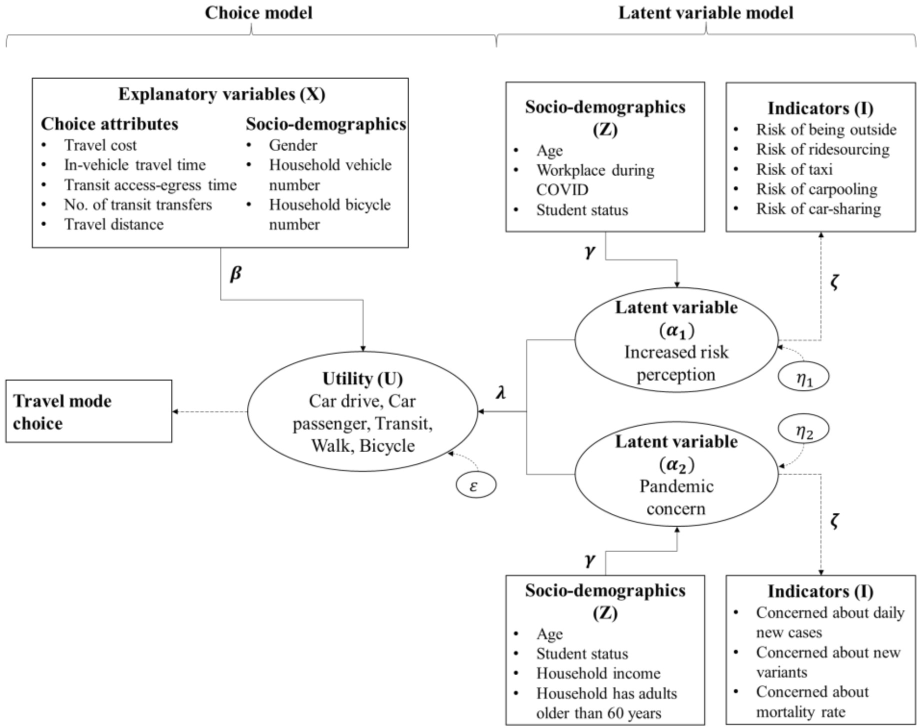
- Five major commute modes: car drive, car passenger, transit, walk, and bicycle
- Transportation level-of-service (LOS) attributes
  - Travel time generated using Google directions API
  - Auto cost generated using cost matrices widely used for transportation planning in the study region
  - Transit fare generated a calibrated Deterministic User Equilibrium traffic assignment model of the study area called the GTA model was used
- Model estimated using each of the synthetic fused datasets

Hybrid choice model specification: Factor analysis

- Factor analysis: to identify latent factors based on the imputed attitudinal questions
- Consistent findings obtained using two factors (with loadings larger than 0.4)

Latent Construct	Observed indicator	Factor Loading
Perception of increased risk during the pandemic	I believe there are more risks associated with leaving my home than before the pandemic	0.402
	I believe there is more risk associated with using ride-sourcing services than before the pandemic	0.400
	I believe there is more risk associated with using taxi services than before the pandemic	0.494
	I believe there is more risk associated with carpooling than before the pandemic	0.445
	I believe there is more risk associated with using car-sharing services (e.g., Zipcar, Communauto) than before the pandemic	0.436
Concerns regarding the pandemic	I am concerned about the number of daily new cases in Ontario, Canada	0.479
	I am concerned about the emergence of the new variant of COVID-19	0.483
	I am concerned about the mortality rate of the disease which is causing the pandemic	0.445

Final hybrid choice model specification



Choice model results

LOS attributes (travel cost, trip length, different travel time components, number of transit transfers) have –ve signs

Females are less likely to cycle than males

Household vehicle and bicycle ownership positively affect car use (car drive and car passenger) and bicycle use

Choice model results – latent attitudes

“Increased risk perception” has +ve effect on car drive mode and negative effect on shared ride mode

Individuals who have higher “pandemic concern” are less likely to choose transit for commuting

Structural and measurement models results

Older respondents and respondents who had to be physically present in their workplace during the pandemic have higher risk perceptions

Respondents who had to be physically present in their workplace during the pandemic have higher risk perceptions

Older respondents and respondents who lived with senior household members have increased pandemic concern

Individuals whose household income is below $60,000 are less likely to be concerned about the pandemic than higher-income individuals

Key findings

- The study presents a proof of concept of how the implicit data fusion method may be used to integrate multiple travel survey data
- The fused data can be reliably used for much more complex and stable investigations than would be possible individually with either the core or the satellite survey data
- Imputing multiple fused datasets helps reduce potential biases that can affect subsequent analyses using the data
- Ideal satellite design should ensure
  - Comprehensive set of consistent, coherent common variables that are well associated with the target variables
  - Same survey conduction period (to control for any external effects)

Questions?

sanjana.hossain@mail.utoronto.ca

Maximum-likelihood-based data fusion approach: Application to EV adoption

TRB Webinar: Unleashing the Power of Fusion in Travel Demand Analysis

Dr. Naveen Eluru
Professor, Civil, Environmental and Construction Engineering
Program Director, MS in Travel Technology and Analytics
University of Central Florida
Orlando, FL, USA

Presentation Outline

Behavioral Data Fusion Approach

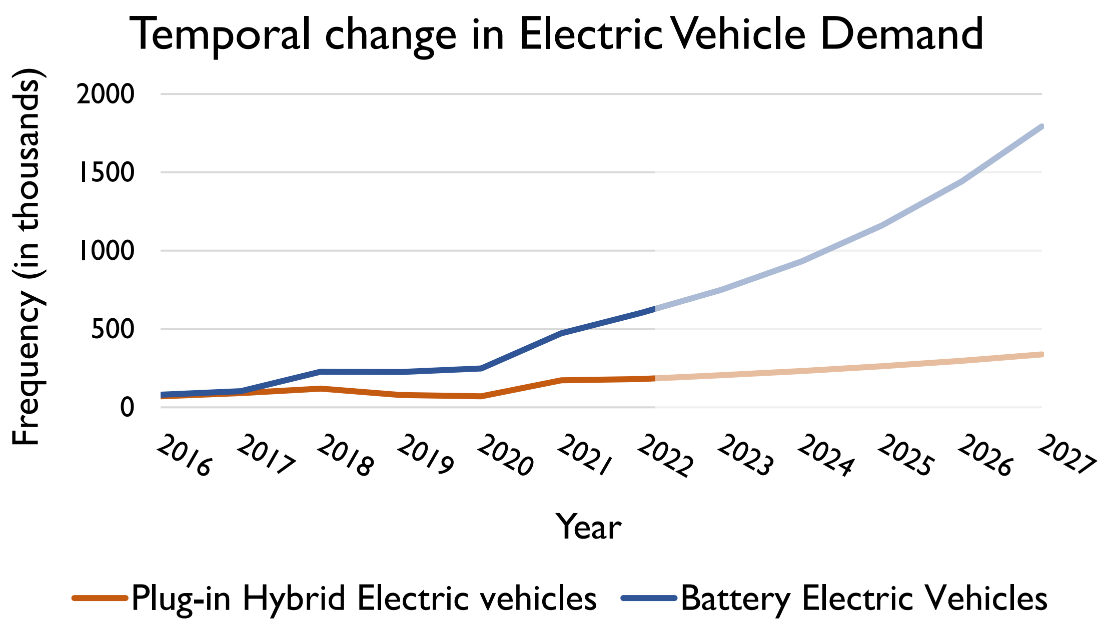
- Idea
- Implementation

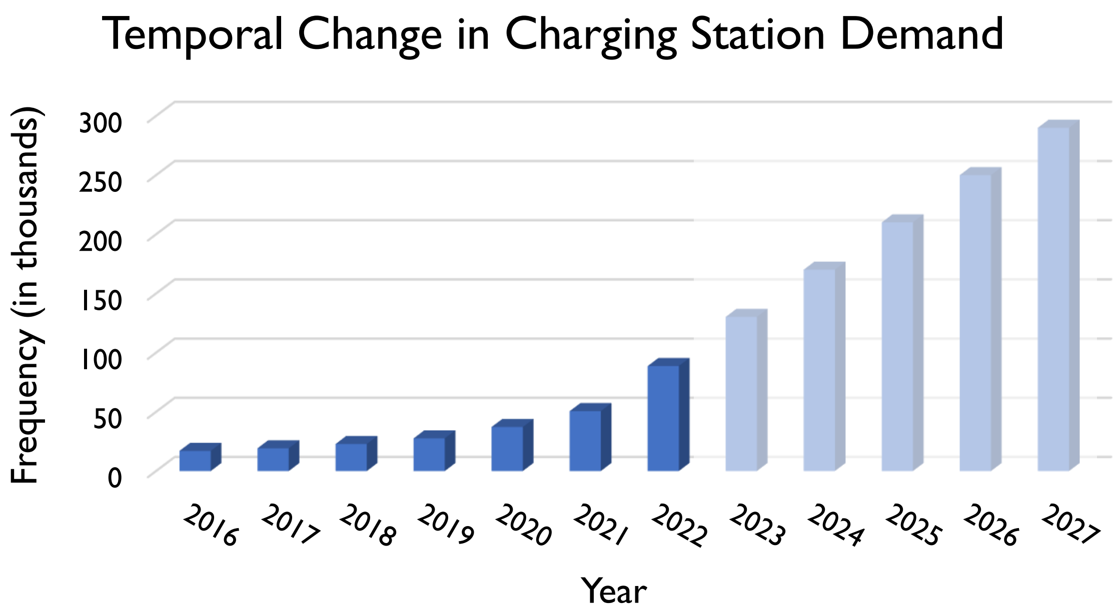
Application to EV Adoption

- Experimental Design
- Results

Background

- Electric Vehicles (EV) sales were close to 0.8 million vehicles in 2022 and is projected to reach 2.1million by 2027.
- EV sales increases demand of charging stations

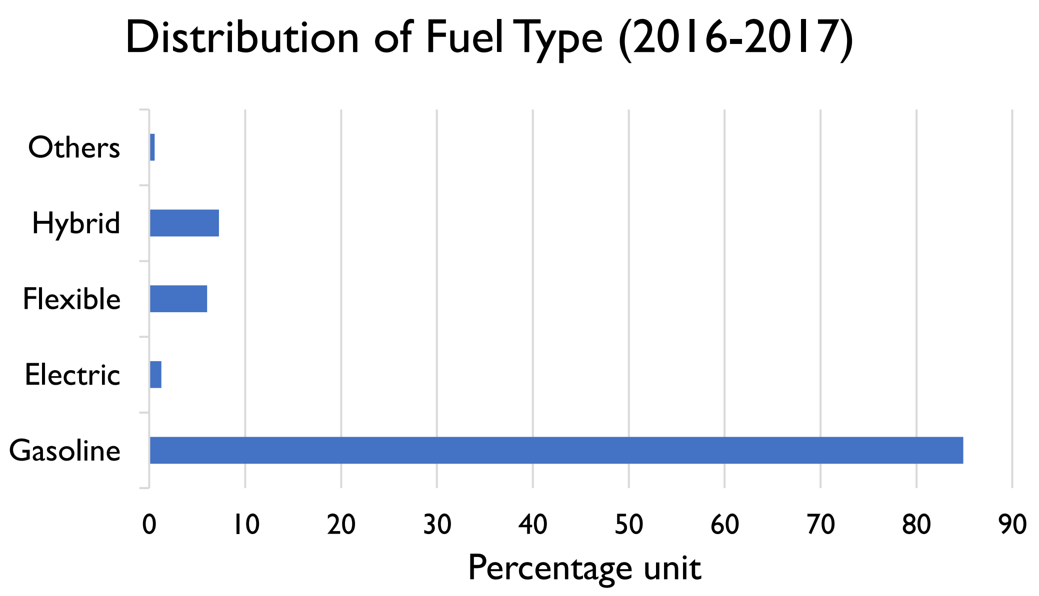


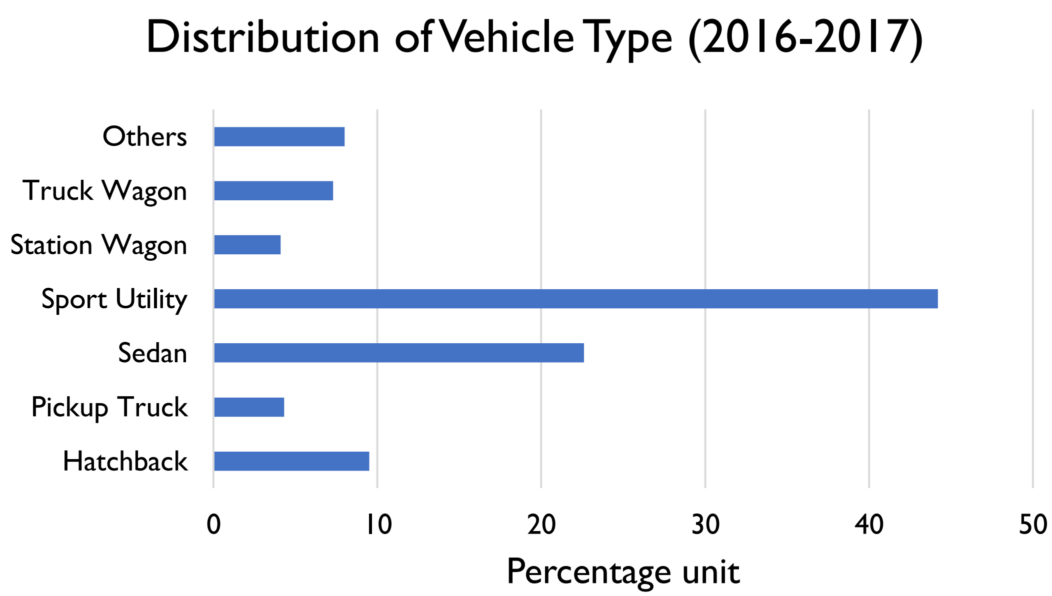


Research Objectives

Objective 1: Evaluation of the factors influencing the purchase of a new vehicle

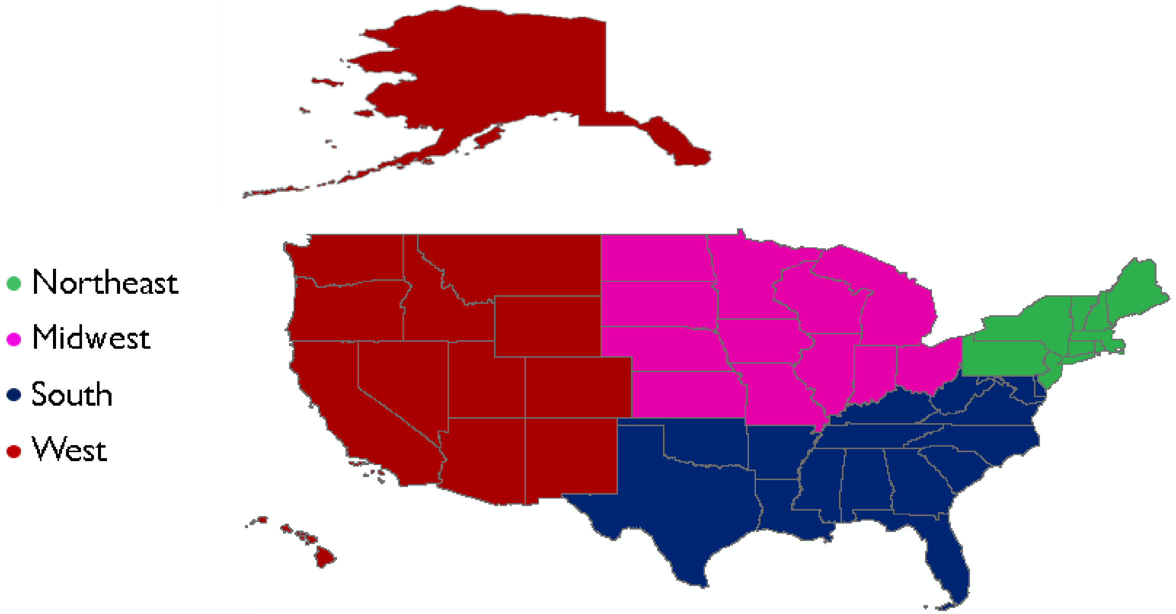
Objective 2: Evaluation of consumers' choice regarding fuel type and vehicle type

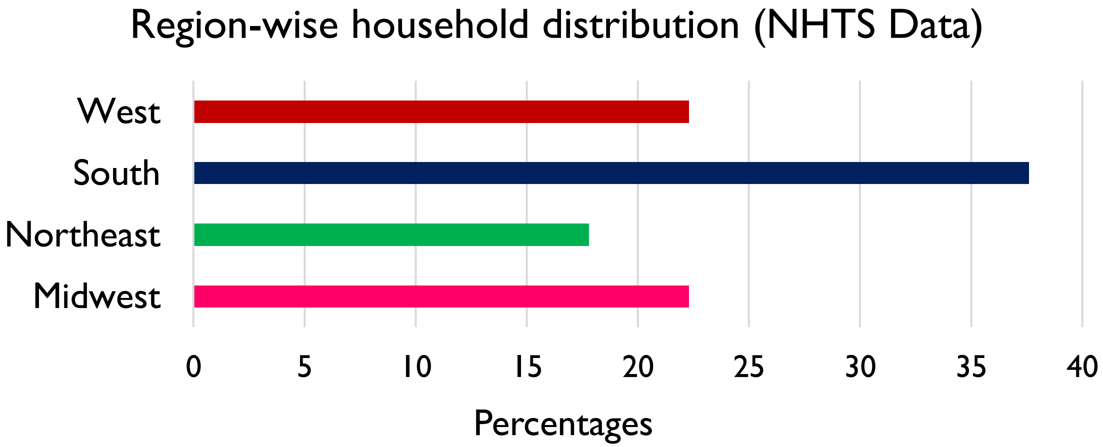




Data Source

- ***MaritzCX Data***
  - Vehicle purchase information
  - Purchase of 133,672 new vehicles in 2016-17
- ***National Household Travel Survey (NHTS) 2017 Data***
  - Consumer's information
  - 129,696 households, representing the properties of 118,208,251 households in the entire USA





Idea

Maritz CX

Challenges

Sample Size limited

Representativeness

Demographics
Vehicle Purchase Decision

Purchase Planning

Idea

Fusion

National Household Travel Survey Data

Create a hybrid Fused Data

Can get additional important variables from NHTS

Idea

Fusion

National Household Travel Survey Data

There is no common identifier

Create a hybrid Fused Data

Idea

Gender
Race
Home ownership

Demographics

Location (Urban/ Rural)
Age category
Income category
State

Common Attributes

Vehicle Purchase Decision: Vehicle type and Fuel type

HH Level Data

HH size
Vehicle availability
Number of workers
Number of drivers

Fused Data

Fusion Example

MaritzCX Data				
Household Id	Age	State	Vehicle type	Fuel type
1	23	CA	Sedan	Gasoline
2	37	FL	SUV	Gasoline
3	59	NY	Truck	Gasoline
4	43	GA	SUV	Electricity

NHTS Data					
Household Id	Person Id	Age	State	HH Size	HH Status
1	1	28	FL	2	Rent
1	2	37	FL	2	Rent
2	1	56	GA	2	Own
2	2	68	GA	2	Own
3	1	37	FL	1	Rent

Fusion Example

MaritzCX Data

Household Id	Age	State	Vehicle type	Fuel type
1	23	CA	Sedan	Gasoline
2	37	FL	SUV	Gasoline
3	59	NY	Truck	Gasoline
4	43	GA	SUV	Electricity

NHTS Data

Household Id	Person Id	Age	State	HH Size	HH Status
1	1	28	FL	2	Rent
1	2	37	FL	2	Rent
2	1	56	GA	2	Own
2	2	68	GA	2	Own
3	1	37	FL	1	Rent

Fused Data

Household Id	Age	State	Vehicle type	Fuel type	HH Size	HH Status
2	37	FL	SUV	Gasoline	2	Rent
2	37	FL	SUV	Gasoline	1	Rent

Fusion Example

Location
Age
Income
State

Equal weight **0.5 (1/NF)**

Function :
Other matching variables

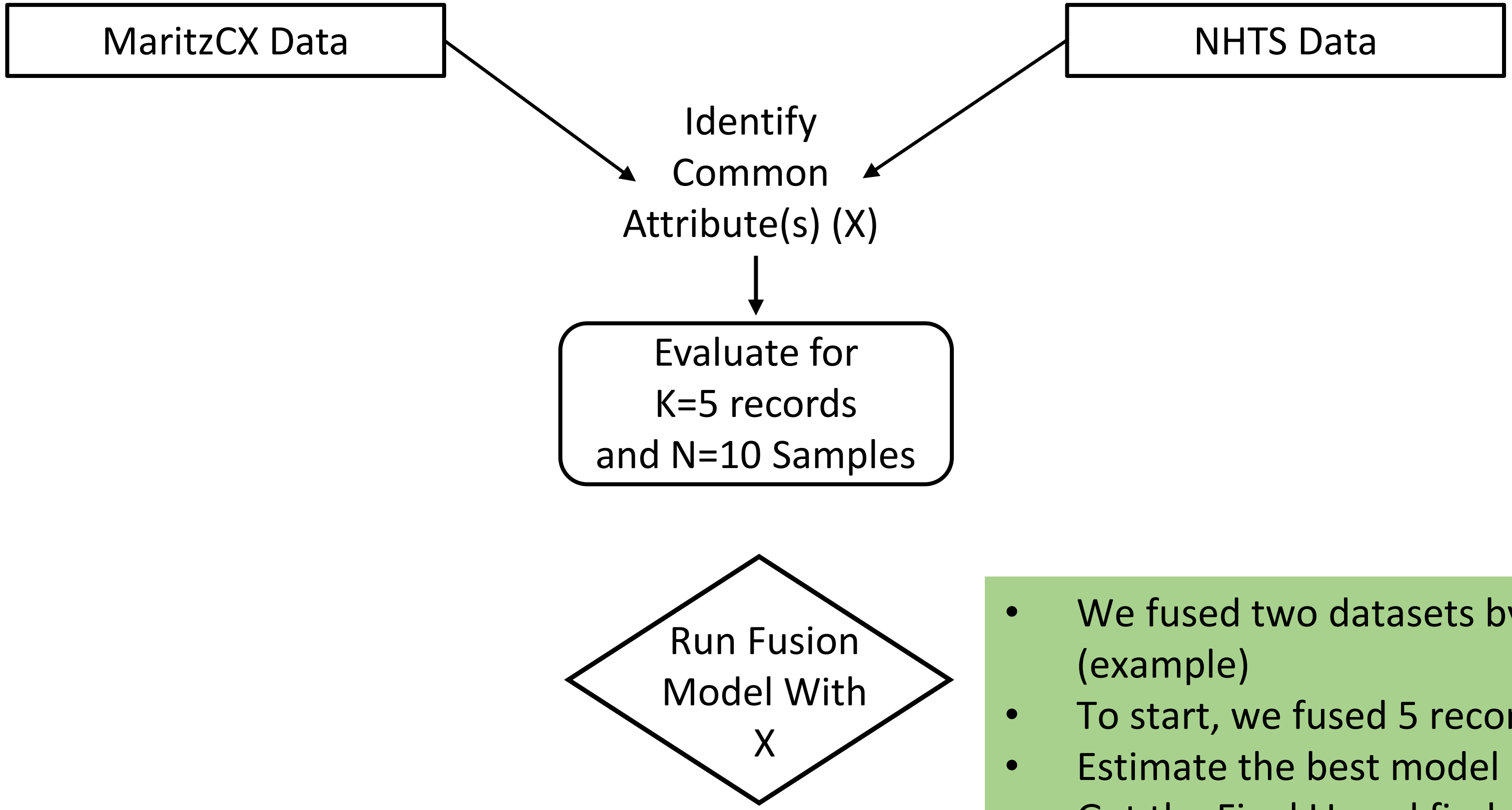
Yes → **Higher weight**

No → **Lower weight**

Fused Data

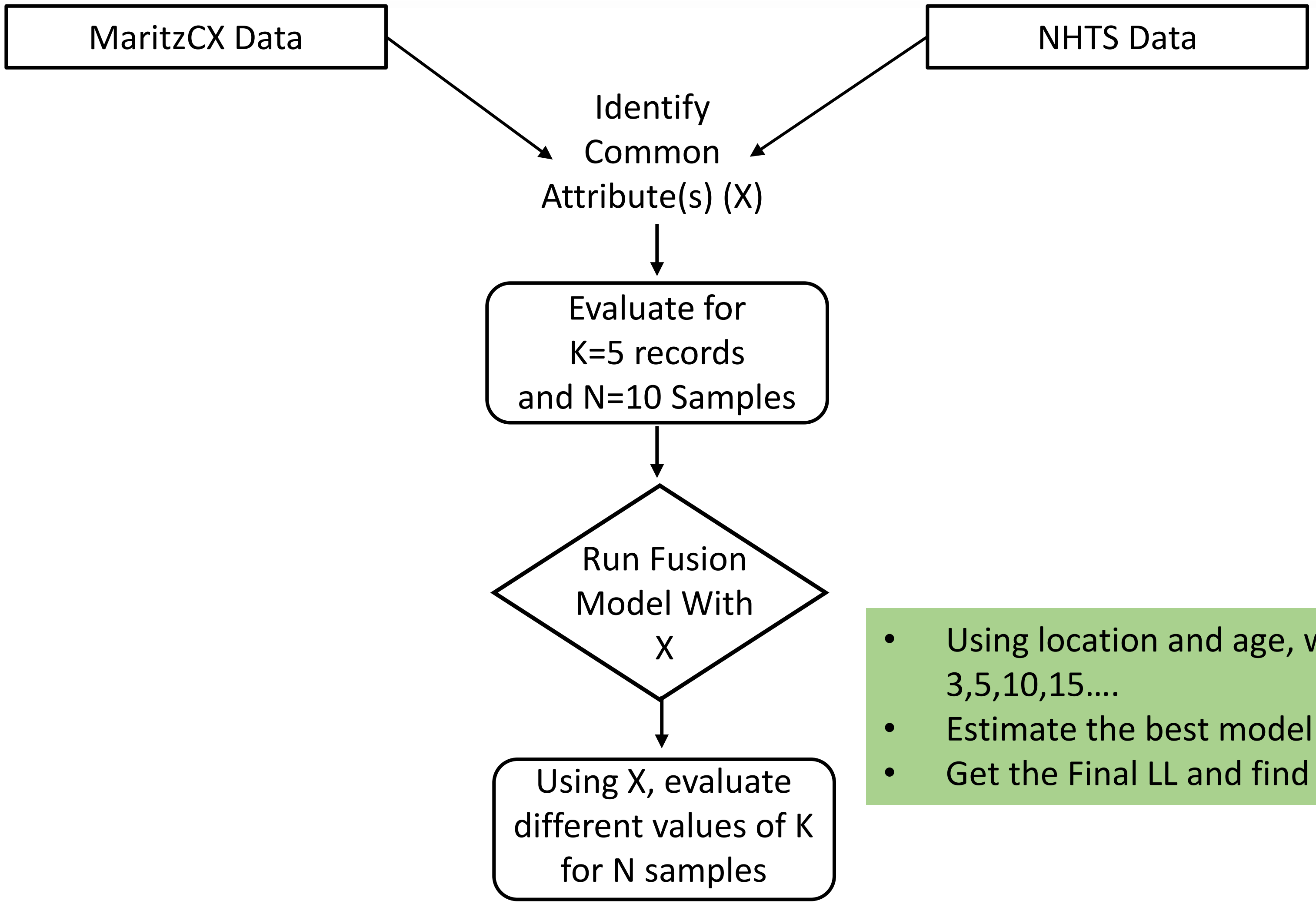
Household Id	Age	State	Vehicle type	Fuel type	HH Size	HH Status
2	37	FL	SUV	Gasoline	2	Rent
2	37	FL	SUV	Gasoline	1	Rent

Experimental Design for Fusion



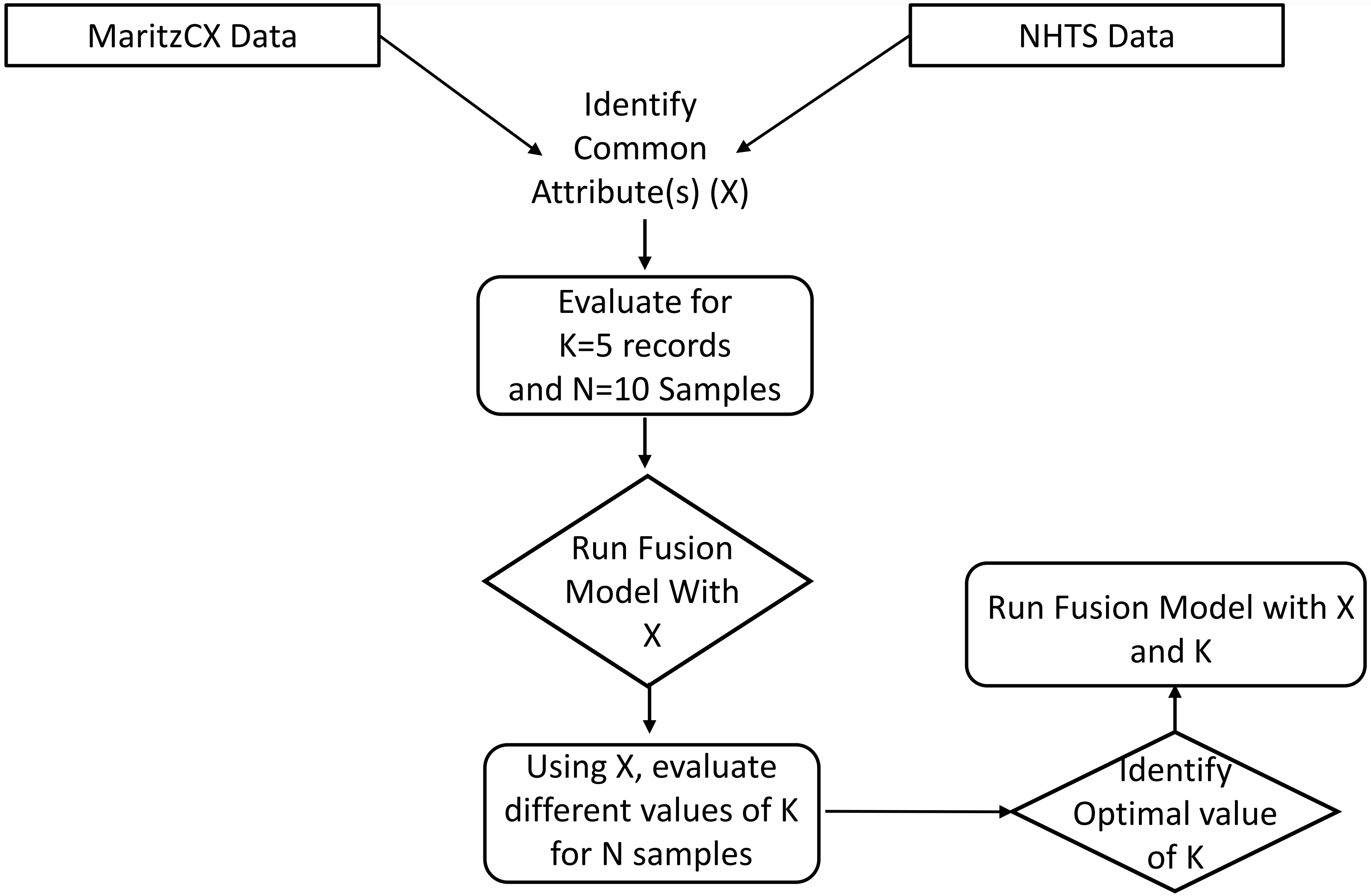
- We fused two datasets by matching Location and age (example)
- To start, we fused 5 records
- Estimate the best model N times
- Get the Final LL and find out the avg. improvement

Experimental Design for Fusion



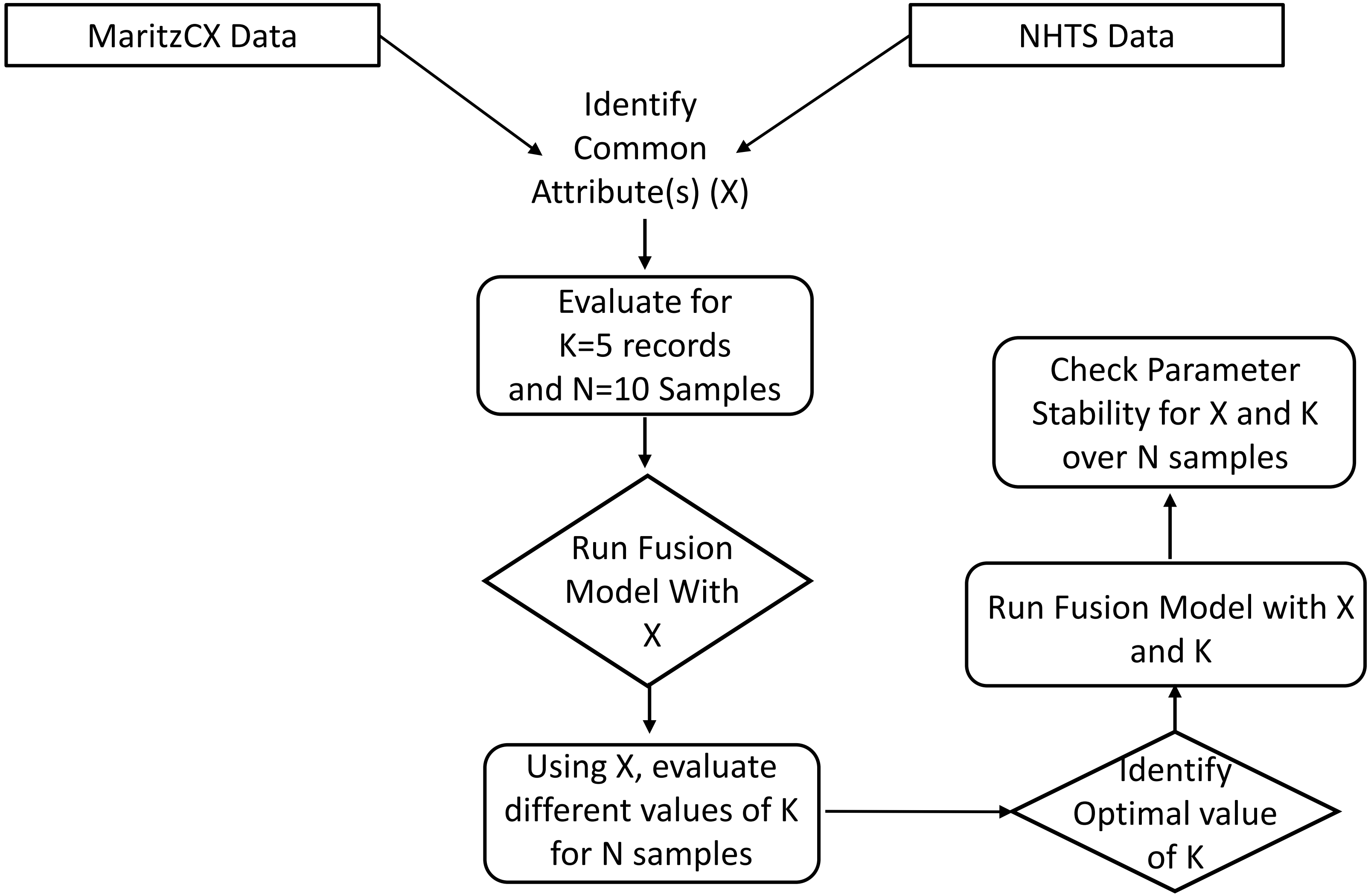
- Using location and age, we tried different number of fusion 3,5,10,15….
- Estimate the best model for N samples
- Get the Final LL and find out the avg. improvement

Experimental Design for Fusion



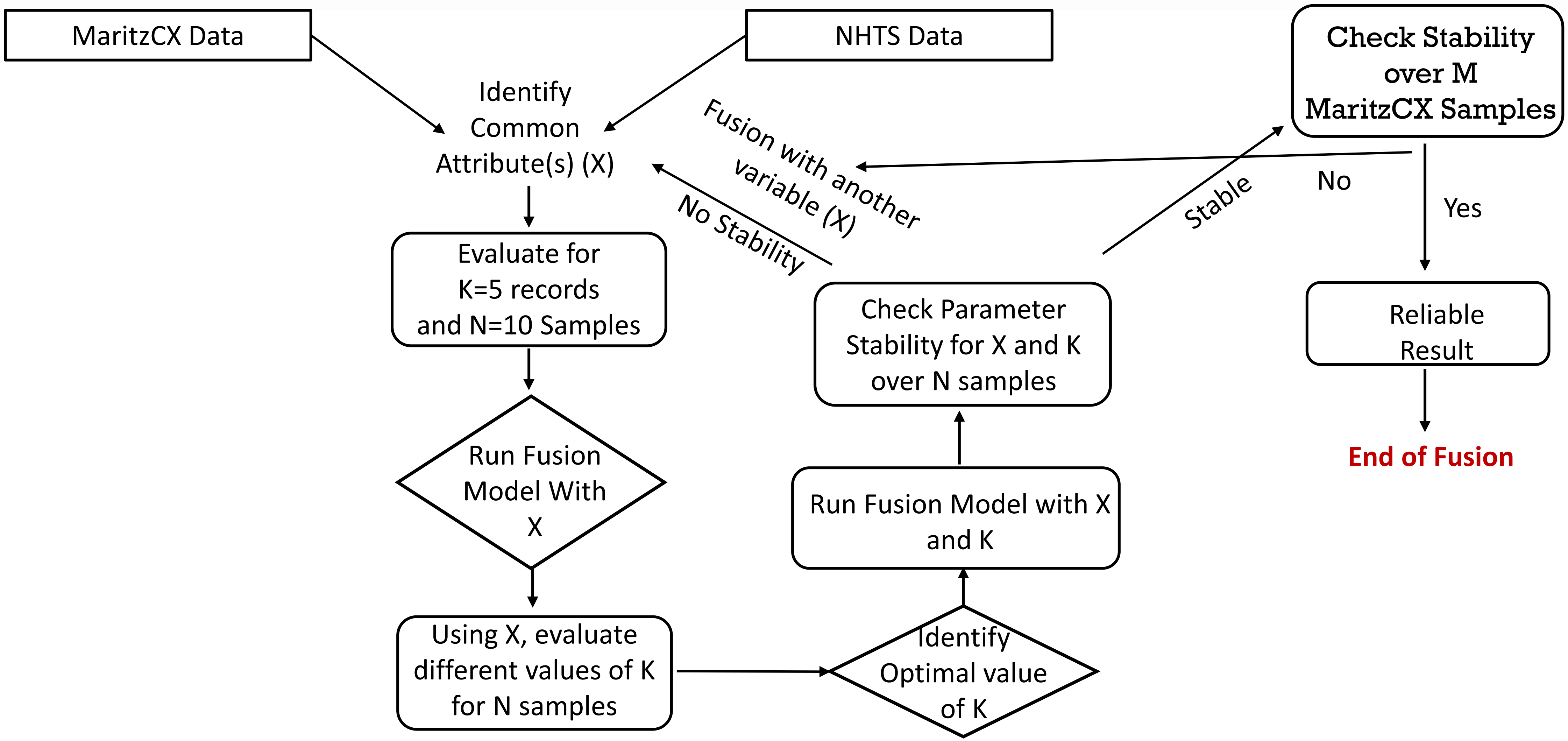
Run for Different X and K Identify optimal X and K

Experimental Design for Fusion



Over 10 samples, we see the difference in parameter estimates

Experimental Design for Fusion



Econometric methodology

Multinomial logit fusion Model (Vehicle choice model):

Decision Component

$$v^*_{nm,jk} = \alpha X_{n,jk} + \beta' S_{nm,jk} + \varepsilon_{nm,jk}, v_{nm,jk} = 1 \; if \; v^*_{nm,jk} > 0; v_{nm,jk} = 0 \; otherwise$$

Attributes in the base (MaritzCX) data

Attributes in the fused (NHTS) data

Propensity of n[th] individual for m[th] fused data choosing fuel type j, and vehicle type k

$$P_{nm,jk} = \frac{exp\left(v^*_{nm,jk}\right)}{\sum_{k=1}^{K}\sum_{j=1}^{J} exp\left(v^*_{nm,jk}\right)}$$

Weight Component

$$w_{nm} = \frac{exp(\Psi Z_{nm})}{\sum_{m=1}^{M} exp(\Psi Z_{nm})}$$

Econometric methodology

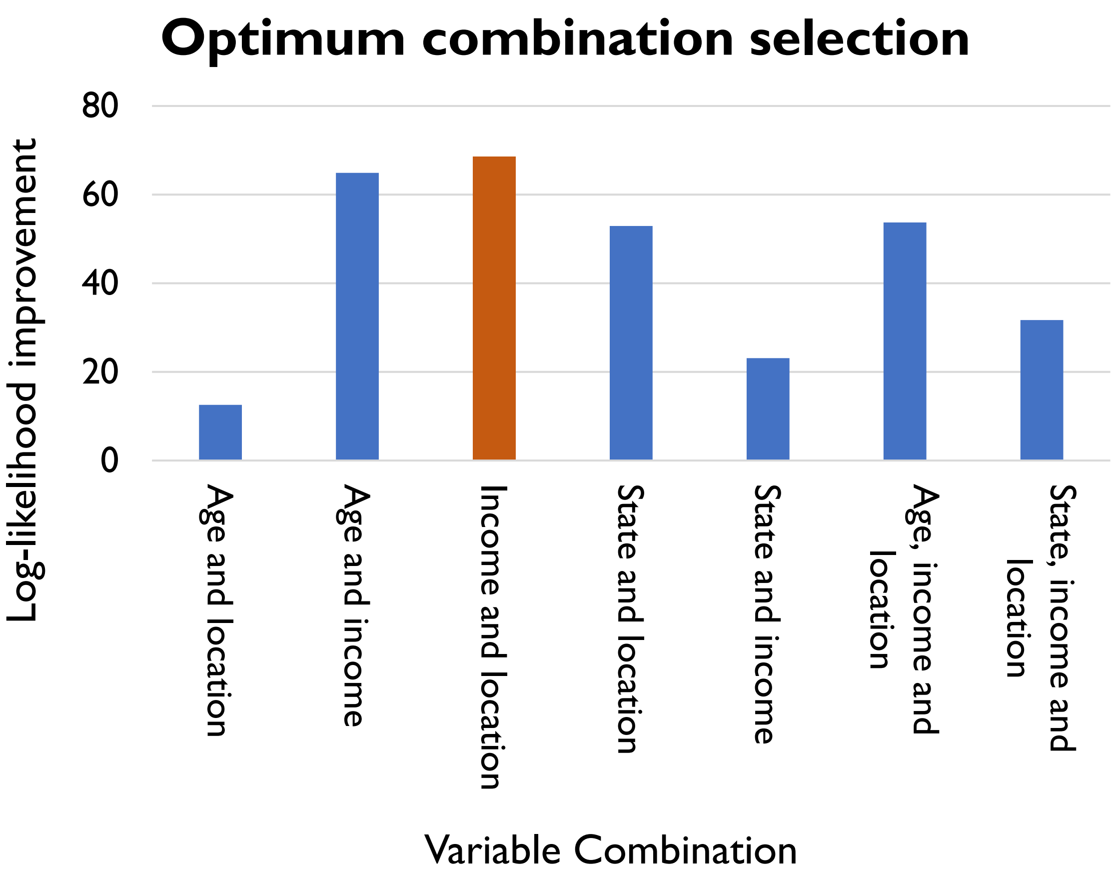
$$Q_n = \sum_{k=1}^{K} w_{nm} * P_{nm,jk}$$

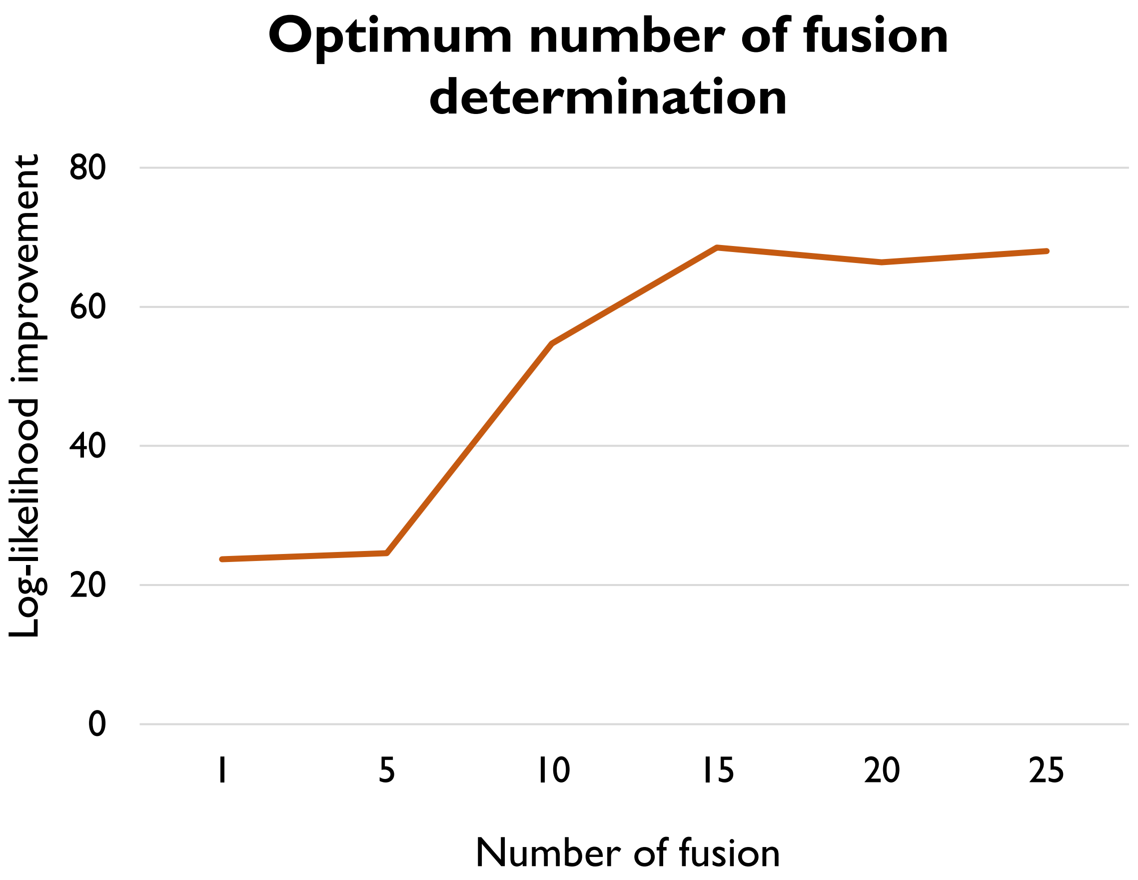
This matching, when executed, will provide us a relationship between the MaritzCX and NHTS datasets

Finally, the log-likelihood function for the fused dataset is defined as:

$$LL = \sum_{n=1}^{N} \log(Q_n)$$

Optimum Model

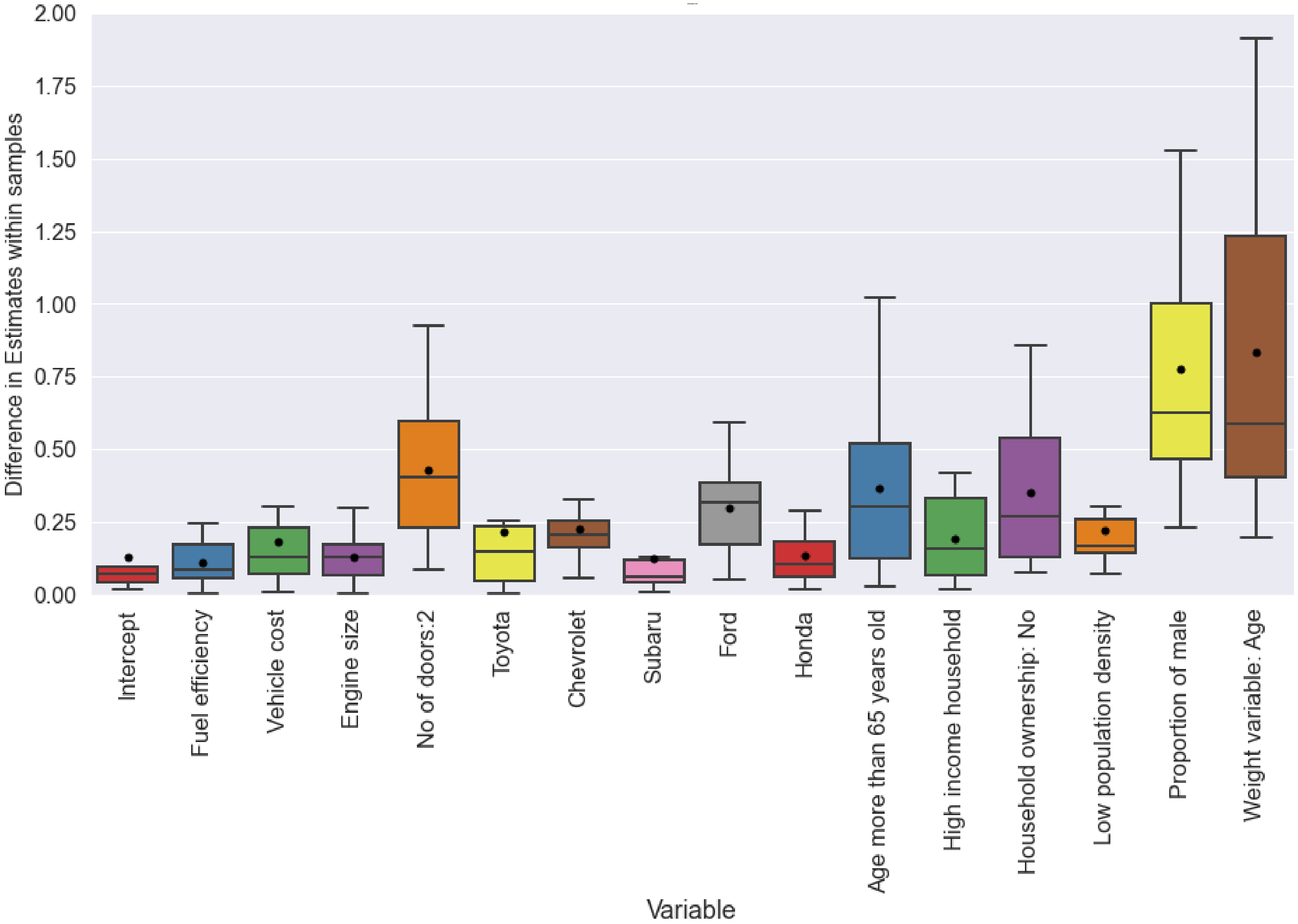




Model Stability Test

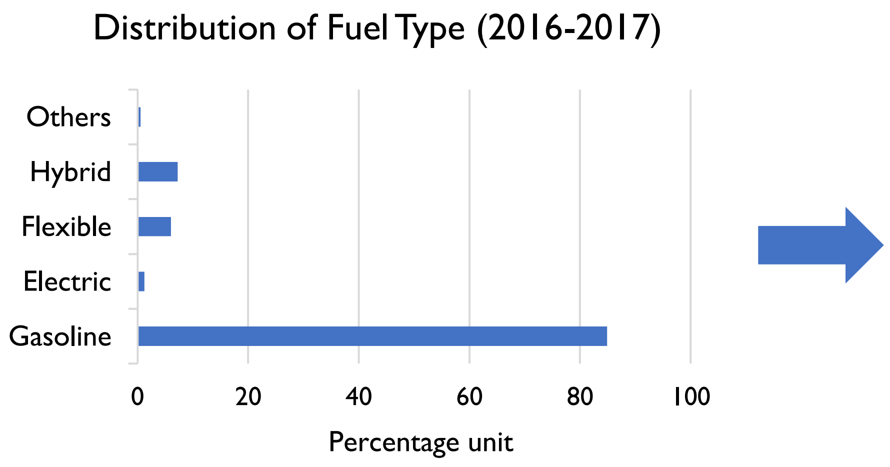
Parameter Test Statistic

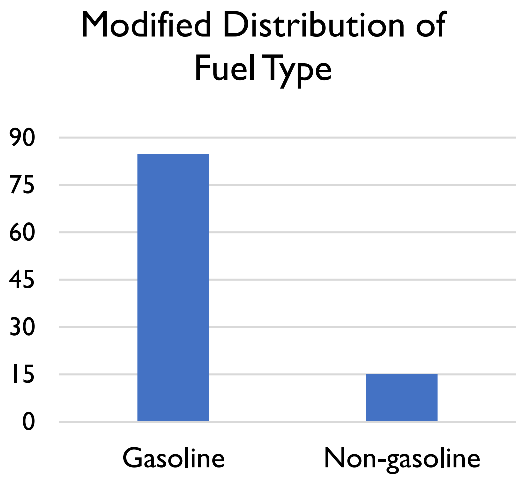
$$= abs\left[\frac{Sample\ parameter - Population\ benchmark}{\sqrt{SE_{sample}^2 + SE_{population}^2}}\right]$$

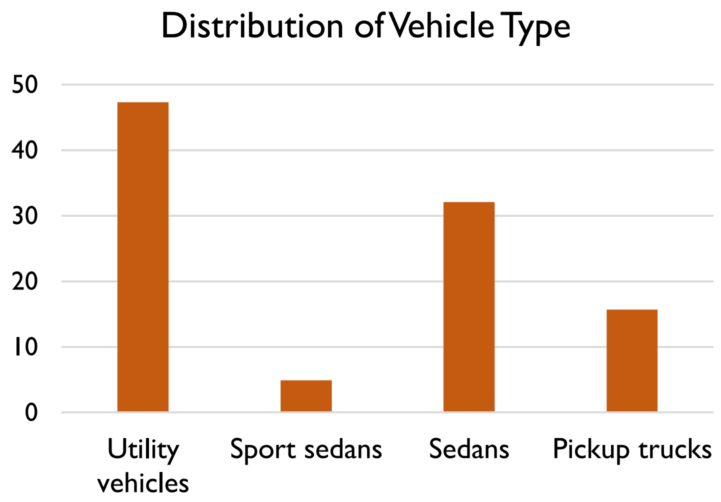


Dependent Variables Considered

- Vehicle purchase model (Binary logit model):
  - Purchase a new vehicle (vehicles of model year 2016 and 2017, and acquired in 2017)
- Vehicle choice model (Multinomial logit fusion model):
  - Fuel type
  - Vehicle type







Exogenous Variables Considered

Exogenous Variables

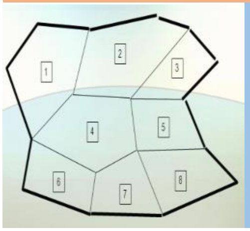
Vehicle Characteristics

- Fuel efficiency
- Engine size
- Vehicle price

Household Economic Characteristics

- Household income
- Vehicle availability
- House ownership

Spatial Factors



- Northeast region
- Midwest region
- South region
- West region

Household Demographics

- Proportion of HH members at different age/ gender
- Ethnicity
- Household size
- Number of workers
- Number of adults

Model Coefficients for Vehicle Choice Model

Variable	Base: Gasoline	Base: Utility vehicles		
	Non-gasoline	Sport sedans	Sedans	Pickup trucks
Intercept	▼	▼	▲	▼
MaritzCX Data				
Vehicle Characteristics				
Fuel efficiency (mpg)	▲	--	--	--
Vehicle price (in thousands)	--	▲	▼	▲
Engine size (liters)	--	--	▼	--
Number of doors (base: 4-door car)				
2-door car	--	--	▼	▲
Vehicle brands (Base: Other brands)				
Toyota	▲	▼	--	▲
Chevrolet	--	--	▼	▲
Subaru	--	▼	▼	
Ford	▼	--	--	▲
Honda	▼	--	▼	▲

Model Coefficients for Vehicle Choice Model

Variable	Base: Gasoline	Base: Utility vehicles		
	Non-gasoline	Sport sedans	Sedans	Pickup trucks
Economic Characteristics				
Household income (base: Low and medium income)				
High income	--	--	--	▼
Demographic Characteristics				
Age (Base: Less than 65 years old)				
Age more than 65 years old	--	▼	▲	--
NHTS Data				
Economic Characteristics				
Household ownership (Base: Own a house)				
Rent a house	--	▲	--	▼
Demographic Characteristics				
Population density (Base: High and medium density)				
Low population density	▲	--	▼	--

Model Coefficients for Vehicle Choice Model

Variable	Base: Gasoline	Base: Utility vehicles		
	Non-gasoline	Sport sedans	Sedans	Pickup trucks
Demographic Characteristics				
Gender (Base: Proportion of female)				
Proportion of male	--	▼	▼	▲
Weight variable				
Age	▲			

Elasticity Analysis

Parameter		Gasoline utility vehicles	Gasoline sport sedans	Gasoline sedans	Gasoline pick-up trucks	Non-gasoline utility vehicle	Non-gasoline sport sedans	Non-gasoline sedans	Non-gasoline pick-up trucks
Fuel efficiency	Non-gasoline sport sedans	-0.20	-0.63	-0.12	-0.04	-0.20	**46.05**	-0.08	-0.04
	Non-gasoline sedans	-1.17	-0.89	-0.87	-0.87	-1.36	-1.16	**16.85**	-1.35
	Non-gasoline pick-up trucks	-0.45	-0.19	-0.26	-1.93	-0.77	-0.14	-0.29	**41.61**
	Non-gasoline utility vehicle	-2.91	-2.11	-1.79	-2.15	**40.11**	-2.89	-1.34	-3.46
Vehicle cost	Gasoline sport sedans	-3.25	**54.47**	-1.79	-1.72	-2.52	-10.39	-0.99	-1.35
	Gasoline sedans	1.75	1.47	**-4.56**	0.87	1.67	1.75	0.79	0.96
	Gasoline pick-up trucks	-4.66	-2.33	-2.22	**32.87**	-4.72	-1.36	-1.80	-17.27
	Non-gasoline sport sedans	-0.33	-1.06	-0.20	-0.07	-0.32	**75.35**	-0.12	-0.07
	Non-gasoline sedans	0.25	0.18	0.18	0.17	0.29	0.25	**-3.49**	0.27
	Non-gasoline pick-up trucks	-0.62	-0.27	-0.34	-2.69	-1.04	-0.20	-0.38	**57.09**
Engine size	Gasoline sedans	11.33	9.46	**-29.53**	5.68	10.79	11.24	5.07	6.36
	Non-gasoline sedans	1.46	1.10	1.08	1.06	1.73	1.48	**-20.96**	1.72
Number of doors:2		-24.27	10.95	-65.38	267.07	-24.69	8.86	-49.99	269.35

Concluding Thoughts

Proposed a novel behavioral data fusion algorithm that can augment available information in one dataset [funded by NSF]

The efficacy of the proposed fusion method is rigorously tested with a well-crafted experimental design

We have implemented this method for Evacuation Decisions Analysis using Social media data and NHTS data

Proposed fusion algorithm can be widely applied across various energy and transportation sectors location-based smartphone data and smart sensor data

Acknowledgements

UCF

Md Istiak Jahan and Tanmoy Bhowmik

University of Maryland

Sachraa G Borjigin, Jiehong Lou, Nneoma M. Ugwu and Prof.

Deb A. Niemeier (Clark Distinguished Chair Professor)

Thank you!

Today’s Presenters

Abdul Rawoof Pinjari
abdul@iisc.ac.in
Indian Institute of Science

Naveen Eluru
naveen.eluru@ucf.edu
University of Central Florida

Sanjana Hossain
sanjana.hossain@mail.utoronto.ca
University of Toronto and
Northwestern University

Khandker Nurul Habib
khandker.nurulhabib@utoronto.ca
University of Toronto

Upcoming events for you

July 23-26, 2023

TRB International Conference on Low Volume Roads

Aug 8 - 10, 2023

Sustainment Management System (SMS) Summit 2023

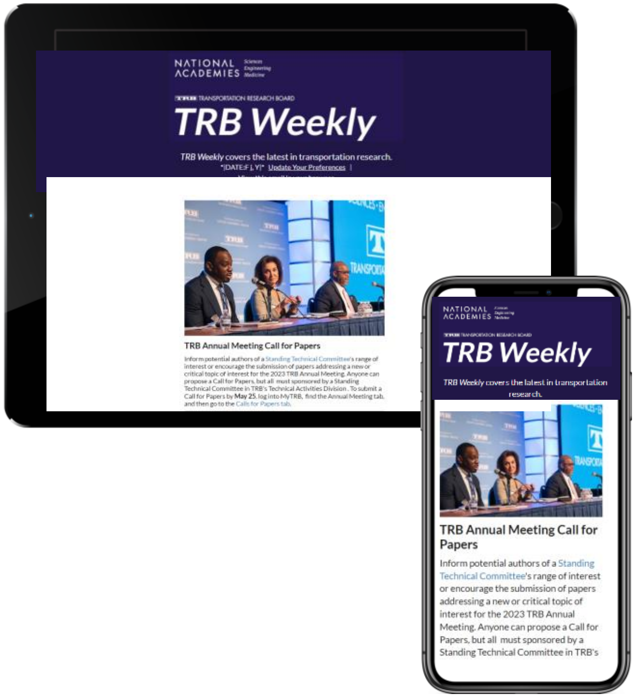
https://www.nationalacademies.org/trb/events

Subscribe to *TRB Weekly*

If your agency, university, or organization perform transportation research, you and your colleagues need the *TRB Weekly* newsletter in your inboxes!

Each Tuesday, we announce the latest:

- RFPs
- TRB's many industry-focused webinars and events
- 3-5 new TRB reports each week
- Top research across the industry



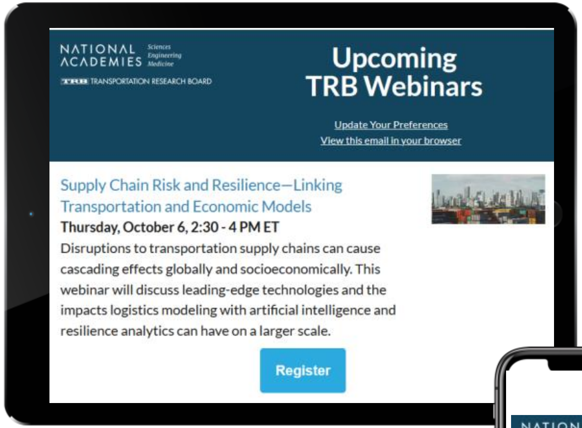
Spread the word and subscribe!
https://bit.ly/ResubscribeTRBWeekly

Discover new TRB Webinars weekly

Set your preferred topics to get the latest listed webinars and those coming up soon every Wednesday, curated especially for you!

https://mailchi.mp/nas.edu/trbwebinars

And follow #TRBwebinar on social media



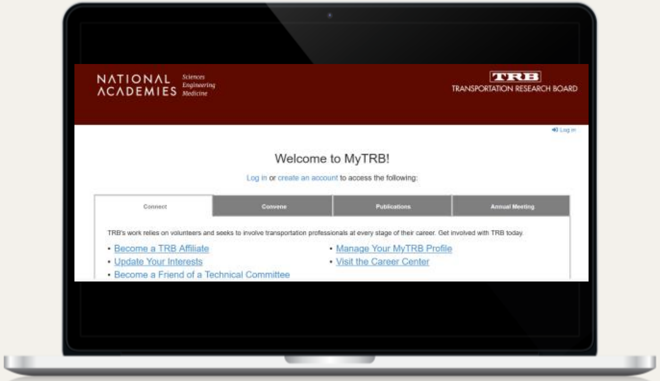
@NASEMTRB

@NASEMTRB

Transportation Research Board

Get involved

https://www.nationalacademies.org/trb/get-involved



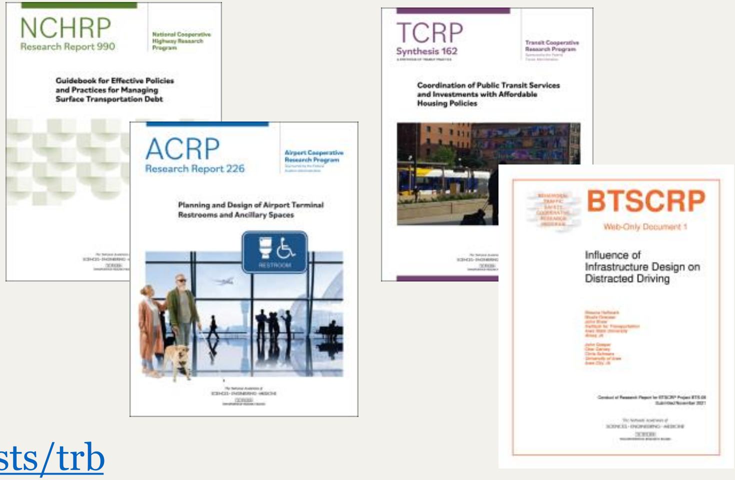
- **Become a Friend of a Standing Technical Committee**

  Network and pursue a path to Standing Committee membership

- **Work with a CRP**

- **Listen to our podcast**

Apple Podcasts | Spotify | Google Podcasts | Overcast | Castbox | Pocket Casts | RSS feed



https://www.nationalacademies.org/podcasts/trb

We want to hear from you

- **Take our survey**

- **Tell us how you use TRB Webinars in your work at trbwebinar@nas.edu**