U.S. Department of Transportation

Imputing Missing Ferry Boarding Counts for Estimating National Totals

2022 FCSM Research & Policy Conference

Session C-6: Lifting the Fog: Editing and Imputation

2022-10-25

Young-Jun Kweon
Bureau of Transportation Statistics

Disclaimer

U.S. Department of Transportation

This study was performed under the sponsorship of the Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for its contents or use thereof.

Outline

- INTRODUCTION
  NCFO / Imputation Study
- DATA
  Questionnaire / Released Tables
- METHODS
  Prep ADS / MI Models / Comparison Metrics / Model Estimation
- RESULTS
- CONCLUSIONS

INTRODUCTION

U.S. Department
of Transportation

National Census of Ferry Operators

- The Safe, Accountable, Flexible Efficient, Transportation Equity Act–A Legacy for Users (SAFETEA-LU) of 2005 (P.L. 114-94) requires BTS to maintain a national ferry database.
- BTS conducts a biennial census of all ferry operators in the U.S. and its territories.
- Who should be included?
  - Ferry operators providing itinerant, fixed route, common carrier passenger/vehicle roll-on, roll-off (RoRo) ferry service, and railroad car float operations
- Who should NOT be included?
  - Non-itinerant operations (e.g., cruise-to-nowhere services)
  - Excursion (e.g., whale watches, casino boats, dinner cruises, etc.)
  - Passenger only water taxi services not operating on a fixed route
  - LoLo (Lift-on/Lift-off) freight/auto carrier services
  - Long distance passenger only cruise ship services

2018 NCFO Imputation Study

- **Background:**
  - ✓ About 5% are missing in passenger and vehicle boarding counts.
  - ✓ With the missing data, trend analysis is challenging.
- **Purpose:**
  - ✓ To impute missing annual boarding counts
- **2018 NCFO**
  - ✓ 238 operators invited (frame)
  - ✓ 181 operators participated

DATA

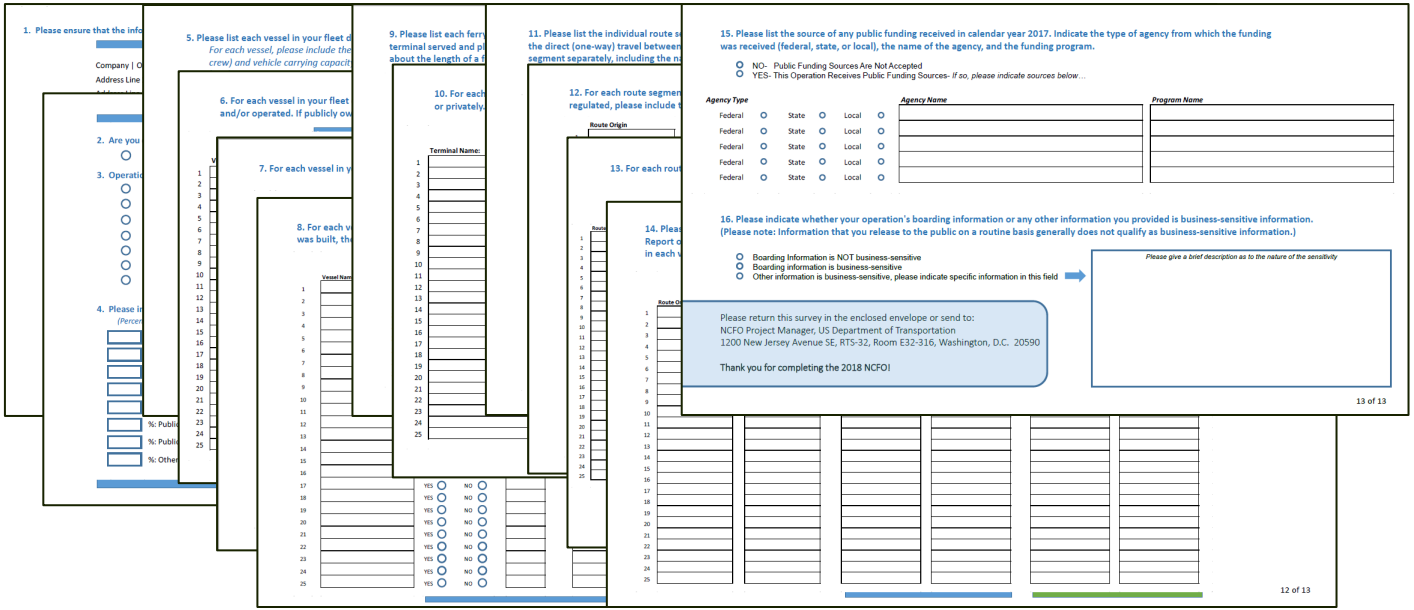
U.S. Department of Transportation

2018 NCFO Questionnaire

U.S. Department of Transportation

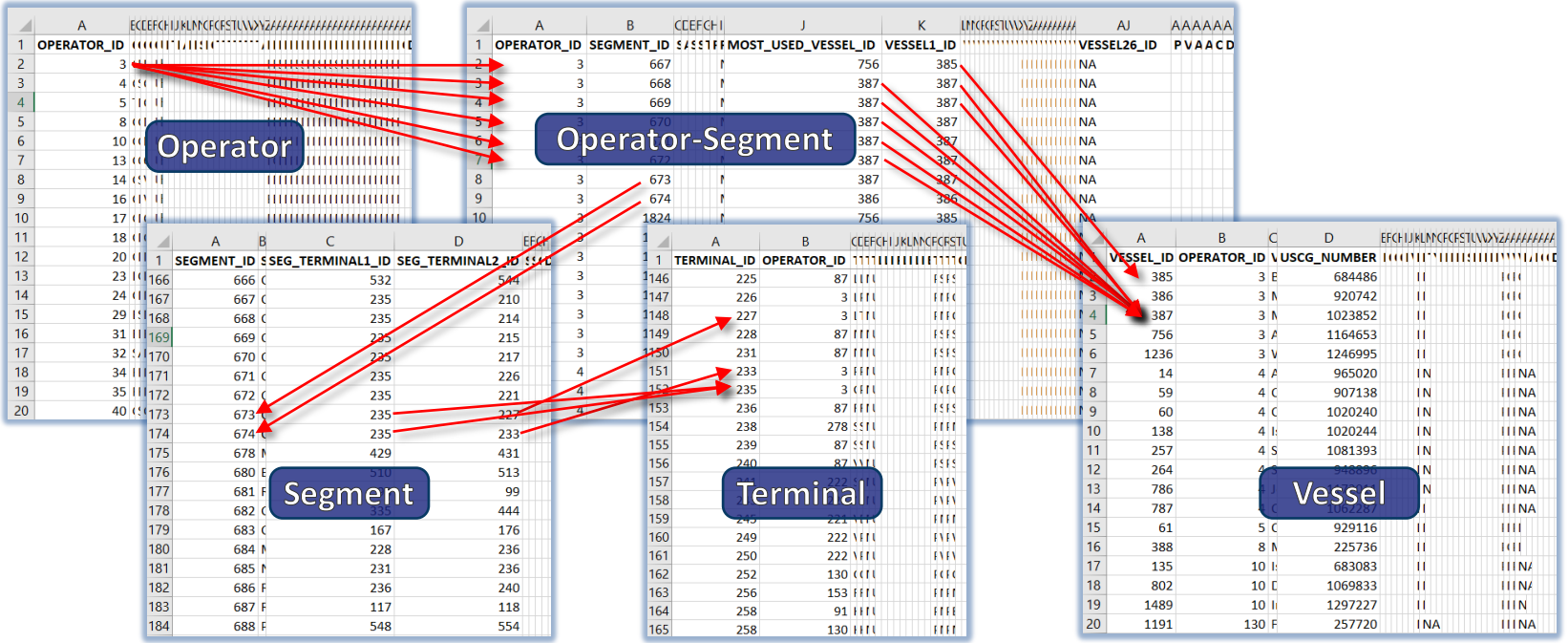
DATA

- 5 Sections: (1) Operator Info, (2) Vessel Info, (3) Terminal Info, (4) Segment Info, and (5) Funding Info.



2018 NCFO Data Release

5 Tables	Operator	Operator-Segment	Segment	Terminal	Vessel
155 Variables	47	48	7	19	34

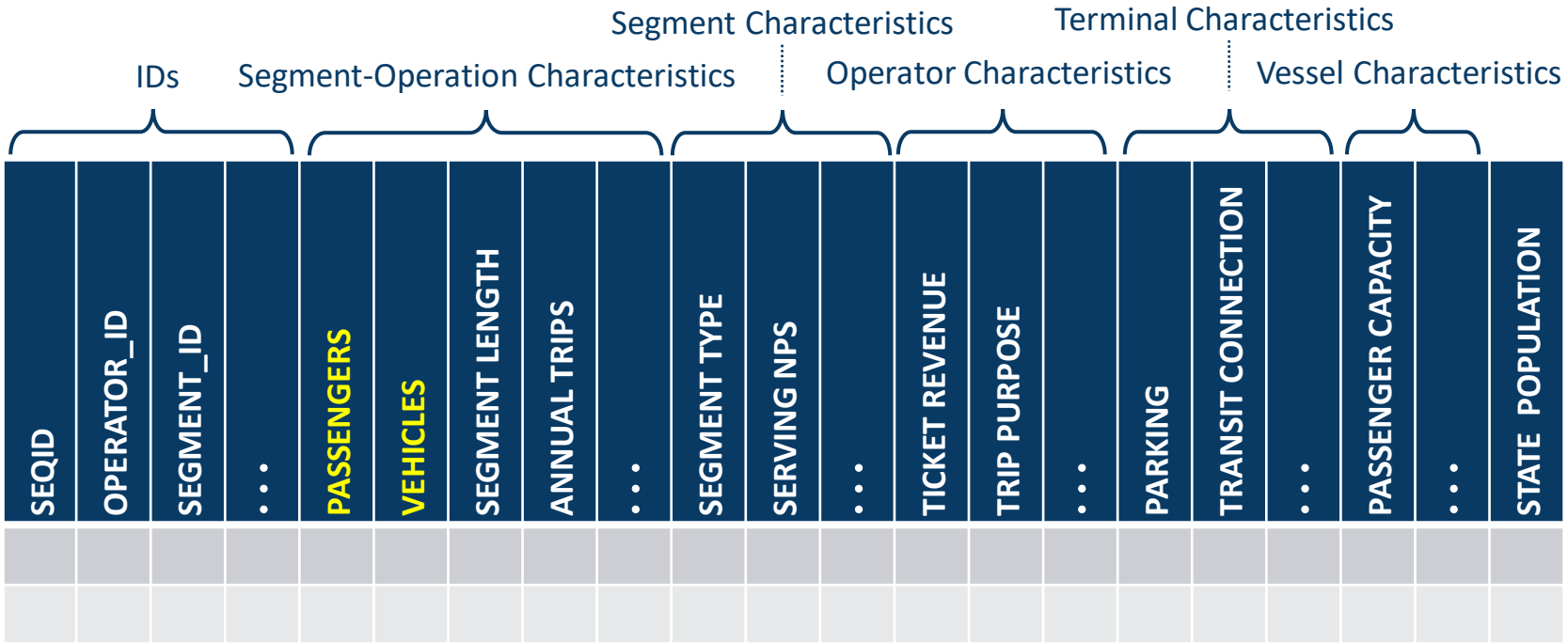


METHODS

U.S. Department of Transportation

Prep Analysis Data Set

- 5 NCFO tables + State Population



MI Models

- **6 Combinations of Methods/Restrictions & 8 Variable Sets: A total of 48 Model Specifications**

	MCMC		FCS			
Variable Set 1	111	112	121	122	123	124
Variable Set 2	211	212	221	222	223	224
Variable Set 3	311	312	321	322	323	324
	411	412	421	422	423	424
	511	512	521	522	523	524
	611	612	621	622	623	624
	711	712	721	722	723	724
Variable Set 8	811	812	821	822	823	824

Model 322

- FCS method with Variable Set 3 (12 variables)
- Box-Cox transformation for all continuous variables
- No min/max/round imposed on categorical variables

Comparison Metrics & Simulated Data

- **Root Mean Squared Error (RMSE)** = $\sqrt{\frac{\sum(Actual-Imputed)^2}{N}}$

Data 1

Obs No.	PASSENGERS	:	:	:	:	:	:
1	#	#	#	#	×	#	#
2	#	#	#	#	#	#	#
3	×	#	×	#	#	#	#
4	#	×	#	#	#	×	#
5	#	#	#	#	#	#	#
6	#	#	×	×	#	#	#
7	×	#	#	#	#	#	#
:	:	:	:	:	:	:	:
939	#	#	#	#	×	#	#
940	×	#	#	#	#	#	#
941	#	×	#	#	#	#	×
942	×	#	#	×	#	#	#
943	×	#	×	×	#	×	#
:	×	:	:	:	:	:	:

Data 2

Obs No.	PASSENGERS	:	:	:	:	:	:
1	×	#	#	#	×	#	#
2	#	#	#	#	#	#	#
3	#	#	×	#	#	#	#
4	×	×	#	#	#	×	#
5	×	#	#	#	#	#	#
6	#	#	×	×	#	#	#
7	#	#	#	#	#	#	#
:	:	:	:	:	:	:	:
939	#	#	#	#	×	#	#
940	#	#	#	#	#	#	#
941	#	×	#	#	#	#	×

Data 3

Obs No.	PASSENGERS	:	:	:	:	:	:
1	#	#	#	#	×	#	#
2	#	#	#	#	#	#	#
3	#	#	×	#	#	#	#
4	×	×	#	#	#	×	#
5	#	#	#	#	#	#	#
6	#	#	×	×	#	#	#
7	#	#	#	#	#	#	#
:	:	:	:	:	:	:	:
939	×	#	#	#	×	#	#
940	#	#	#	#	#	#	#
941	×	×	#	#	#	#	×

...

Data 30

Obs No.	PASSENGERS	:	:	:	:	:	:
1	#	#	#	#	×	#	#
2	×	#	#	#	#	#	#
3	#	#	×	#	#	#	#
4	#	×	#	#	#	×	#
5	#	#	#	#	#	#	#
6	×	#	×	×	#	#	#
7	#	#	#	#	#	#	#
:	:	:	:	:	:	:	:
939	#	#	#	#	×	#	#
940	×	#	#	#	#	#	#
941	#	×	#	#	#	#	×

47 forced missing cases in each data

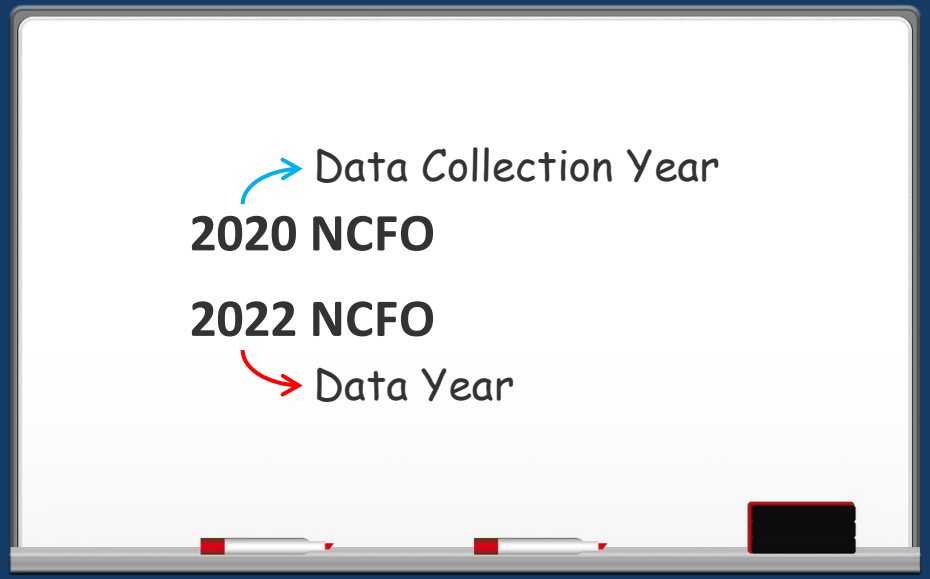
Model Estimation

- **Among 48 models, 21 models were estimated and included for model comparison**

111	112	121	122	123	124
211	212	221	222	223	224
311	312	321	322	323	324
411	412	421	422	423	424
511	512	521	522	523	524
611	612	621	622	623	624
711	712	721	722	723	724
811	812	821	822	823	824

: Produced same results

: Failed producing imputed values in at least one of 30 data sets



RESULTS

U.S. Department of Transportation

Results: All 30 Data Sets

- RMSE(Passenger) = $\sqrt{\frac{\sum(PASS_{Actual} - PASS_{Imputed})^2}{47 \times 30}}$

Top 10 Models: All 30 Data Sets

Rank	Model	RMSE
1	121	294,912
2	111	297,203
3	712	299,134
4	812	305,134
5	623	310,022
6	512	312,194
7	622	314,309
8	624	314,309
9	522	315,487
10	524	315,487

Average of Passengers in 2018 NCFO: 131,204 persons per segment

Results: By Data Set

- RMSE(Passenger) = $\sqrt{\frac{\sum(PASS_{Actual} - PASS_{Imputed})^2}{47}}$

Top 5 Models by Data Set

Data 1

Rank	Model	RMSE
1	812	116,687
2	712	120,990
3	322	124,631
4	324	124,631
5	622	125,002

Data 2

Rank	Model	RMSE
1	712	133,719
2	121	134,371
3	812	135,959
4	622	138,663
5	624	138,663

Data 3

Rank	Model	RMSE
1	111	336,627
2	712	342,691
3	121	351,350
4	812	356,842
5	622	359,865

Data 4

Rank	Model	RMSE
1	111	470,599
2	121	487,447
3	812	505,332
4	412	516,106
5	712	516,111

Data 5

Rank	Model	RMSE
1	121	424,625
2	623	443,567
3	522	451,134
4	524	451,134
5	512	451,662

Data 6

Rank	Model	RMSE
1	712	210,778
2	111	212,953
3	622	215,804
4	624	215,804
5	121	219,367

...

Data 30

Rank	Model	RMSE
1	111	239,243
2	323	246,124
3	522	248,447
4	524	248,447
5	412	251,256

Comparison: Quasi-Nat's Total

❶ Quasi-Nat'l **Obs. Total** = $\sum_{All\ Cases} PASS_{Actual}$

❷ Quasi-Nat'l **Est. Total** = $\sum_{Non-Missing} PASS_{Actual} + \sum_{Forced\ Missing} PASS_{Imputed}$

Data 1

Before forced missing

Obs No.	PASSENGERS	...	...	...	...	...	...
1	#	#	#	#	×	#	#
2	#	#	#	#	#	#	#
3	#	#	×	#	#	#	#
4	#	×	#	#	#	×	#
5	#	#	#	#	#	#	#
6	#	#	×	×	#	#	#
7	#	#	#	#	#	#	#
:	:	:	:	:	:	:	:
939	#	#	#	#	×	#	#
940	#	#	#	#	#	#	#
941	#	×	#	#	#	#	×

❶

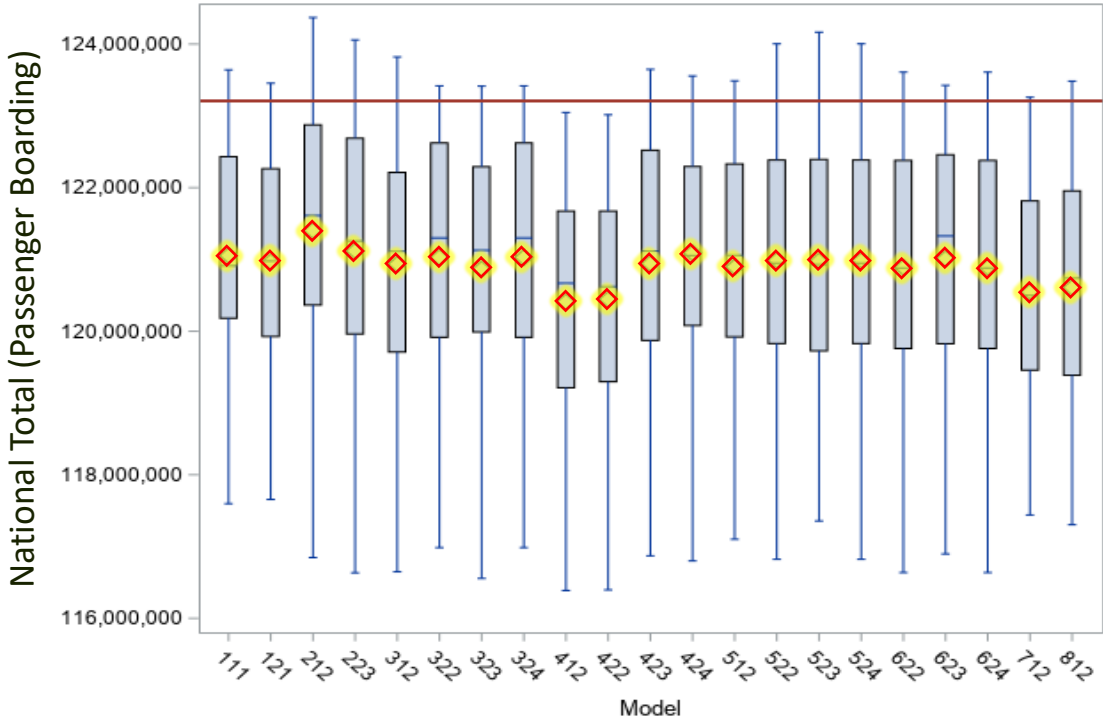
***After** forced missing*

Obs No.	PASSENGERS	...	...	...	...	...	...
1	#	#	#	#	×	#	#
2	#	#	#	#	#	#	#
3	×	#	×	#	#	#	#
4	#	×	#	#	#	×	#
5	#	#	#	#	#	#	#
6	#	#	×	×	#	#	#
7	×	#	#	#	#	#	#
:	:	:	:	:	:	:	:
939	#	#	#	#	×	#	#
940	×	#	#	#	#	#	#
941	#	×	#	#	#	#	×

❷

Results: Quasi-Nat'l Total

❶ Observed Total vs. ❷ Estimated Total

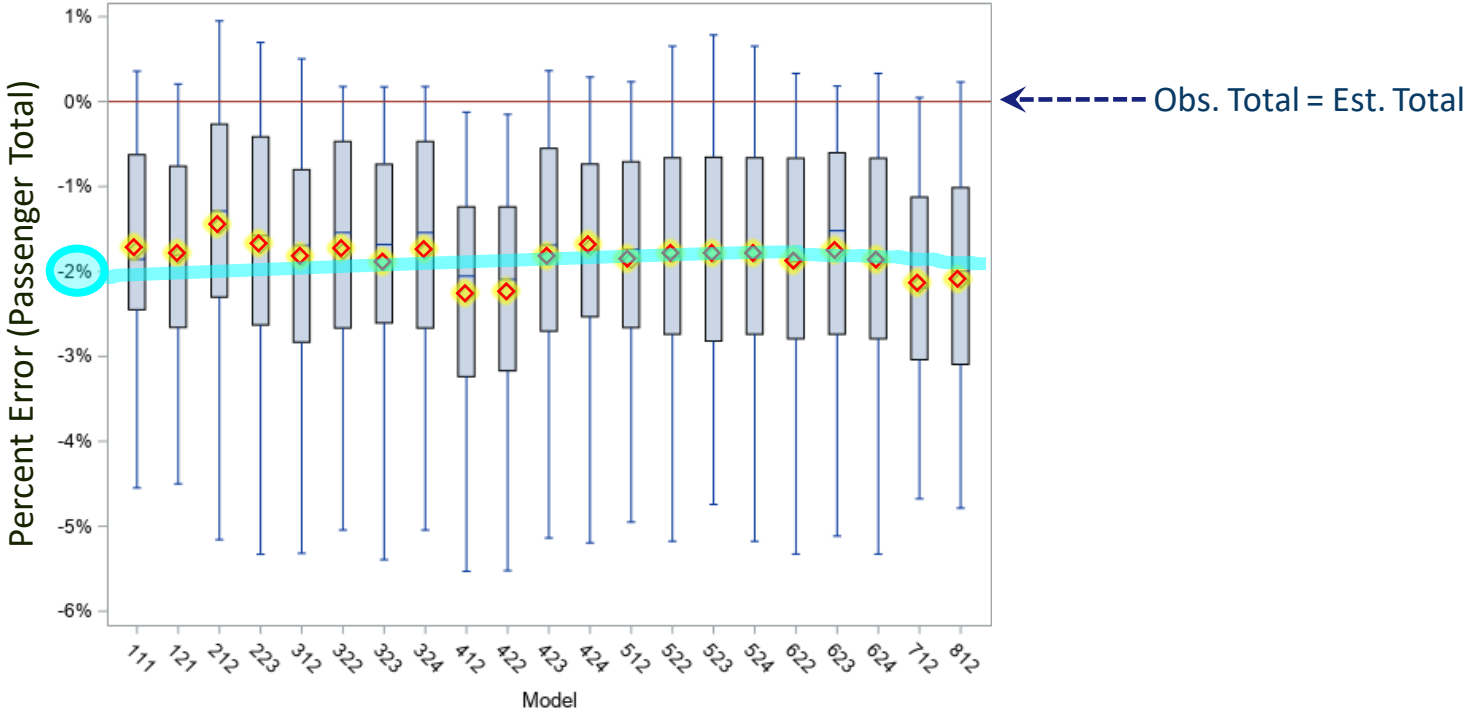


❶ Observed Total

❷ *Each box plot includes 30 estimated totals, one for each data set.*

Results: % Error in Nat'l Total

$$\text{Percent Error \%} = \frac{(Est.Total - Obs.Total)}{Obs.Total} \times 100\%$$



CONCLUSIONS

U.S. Department
of Transportation

Conclusions

- MI models developed are not acceptable for boarding count imputation at an individual segment level.
- They may be acceptable for estimating national total boarding count.
  - ✓ An individual model would be not appropriate. Instead, several MI models and averaging should be used for estimating national total.

Moving Forward

- Analyzing 2020 & 2018 NCFO together
- Applying logical imputation on cases where statistical imputation is not likely to work
- Employing ML methods (e.g., random forest and Ensemble)
- Estimating boarding counts of operators nonresponding to both 2020 and 2018 censuses

Acknowledgement

- Aubrey Nguyen, IT Auditor, U.S. Government Accountability Office (GAO)
  ORISE Fellow, USDOT (Formerly)
- Clara Reschovsky, NCFO Program Manager/Survey Statistician, USDOT

U.S. Department of Transportation

Contact

Young-Jun Kweon

young-jun.kweon@dot.gov

NCFO

ferry@dot.gov

Thank You!

Questions?

U.S. Department of Transportation