U.S. Department of Transportation

Evaluating Imputation Models for Totals Considering Missing Patterns

2023 FCSM Research & Policy Conference

Session E-6: The Missing Data Puzzle: Exploring Imputation Methods

2023-10-25

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

U.S. Department of Transportation

- INTRODUCTION
  - NCFO / Imputation Study
- DATA
  - Questionnaire / Released Tables
- METHODS
  - Analysis Flow / Individual Models / Combined Models
- RESULTS
  - Individual Models / Combined Models / Overall Best Models
- CONCLUSIONS

INTRODUCTION

U.S. Department of Transportation

National Census of Ferry Operators

U.S. Department of Transportation

- 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.

2020 NCFO Imputation Study

U.S. Department of Transportation

- **Background:**
  - ✓ Passenger and vehicle boarding counts are key information
  - ✓ 34% of ferry segments have missing passenger boarding in 2020 NCFO
  - ✓ Missing counts make it difficult to monitor changes in national totals

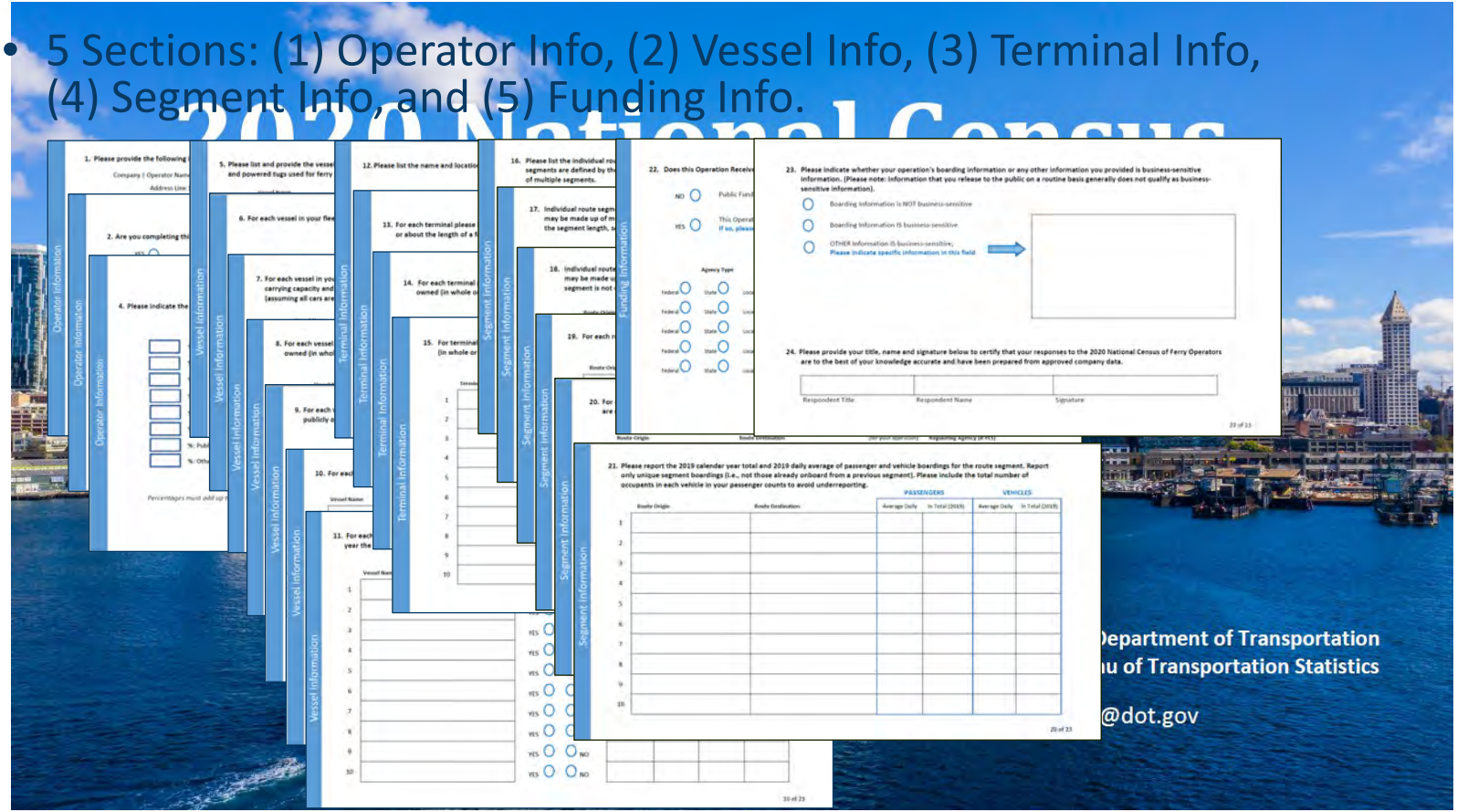
DATA

U.S. Department
of Transportation

2020 NCFO Questionnaire

U.S. Department of Transportation

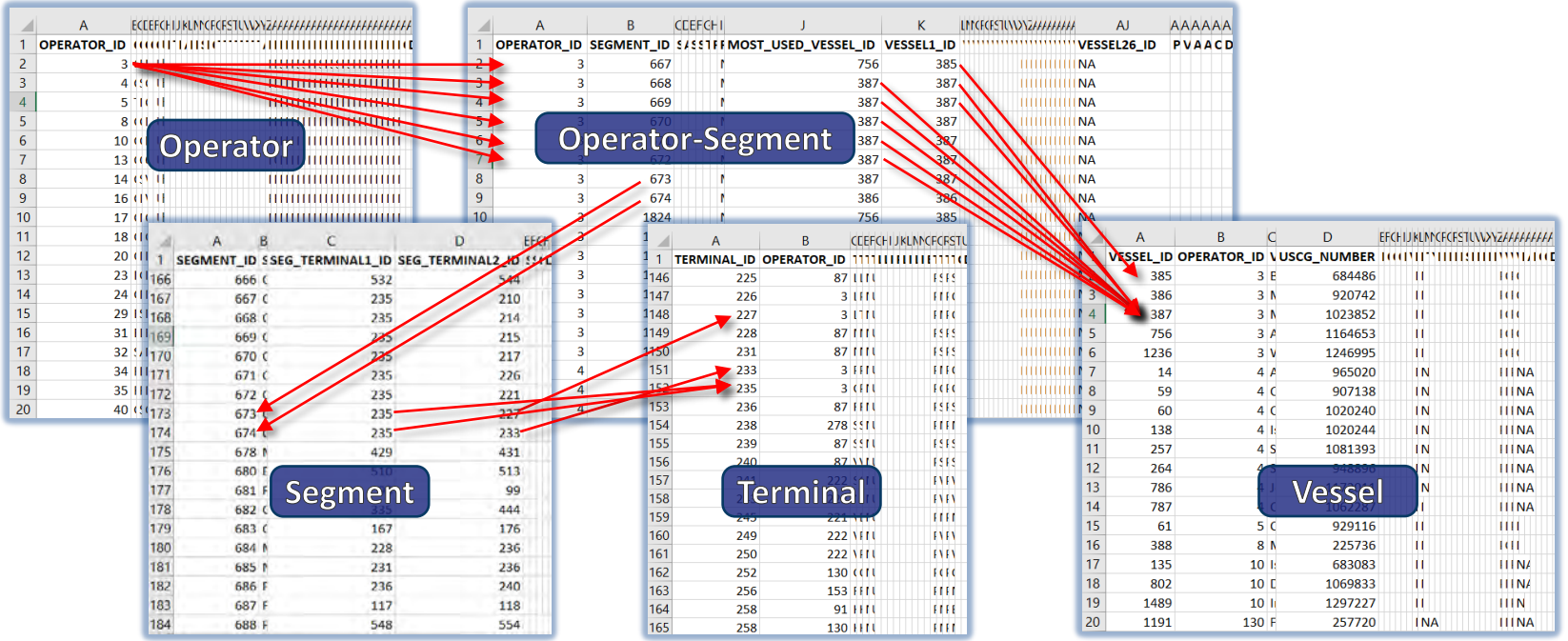
- 5 Sections: (1) Operator Info, (2) Vessel Info, (3) Terminal Info, (4) Segment Info, and (5) Funding Info.



2020 NCFO Data Release

U.S. Department of Transportation

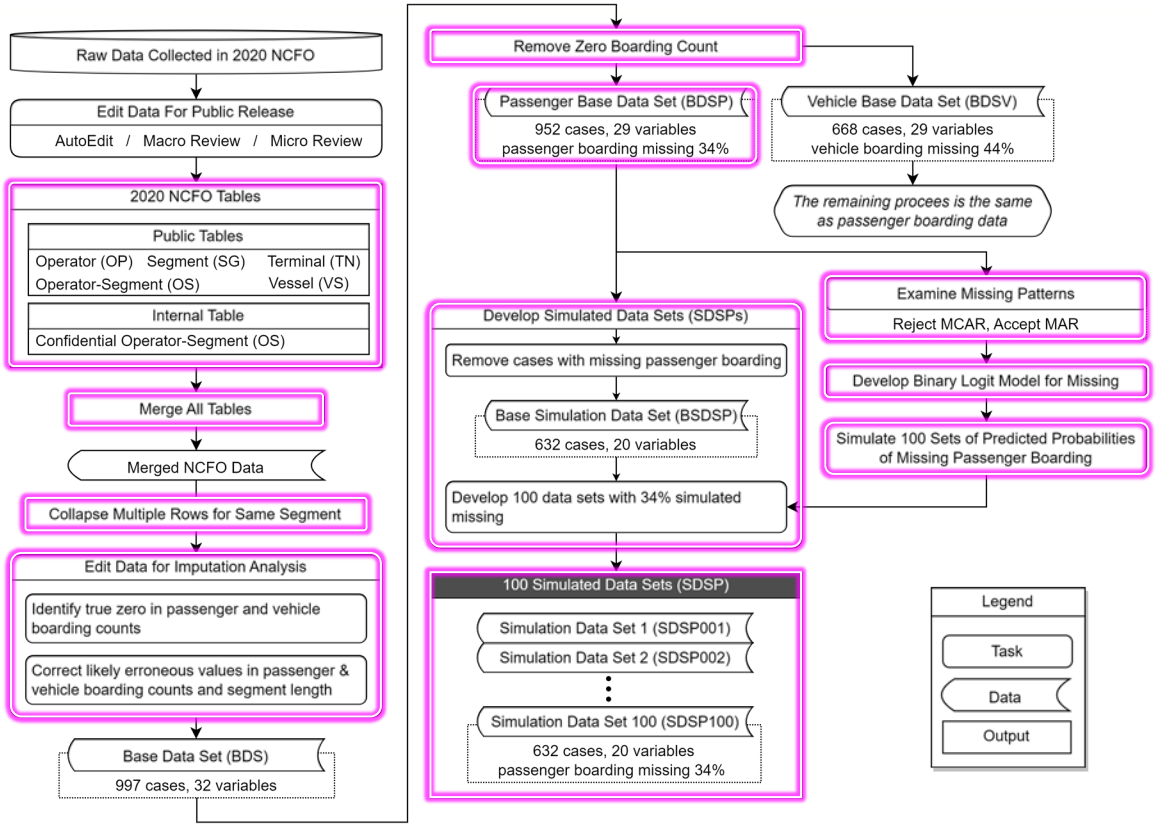
5 Tables	Operator	Operator-Segment	Segment	Terminal	Vessel
152 Variables	50	41	7	20	34



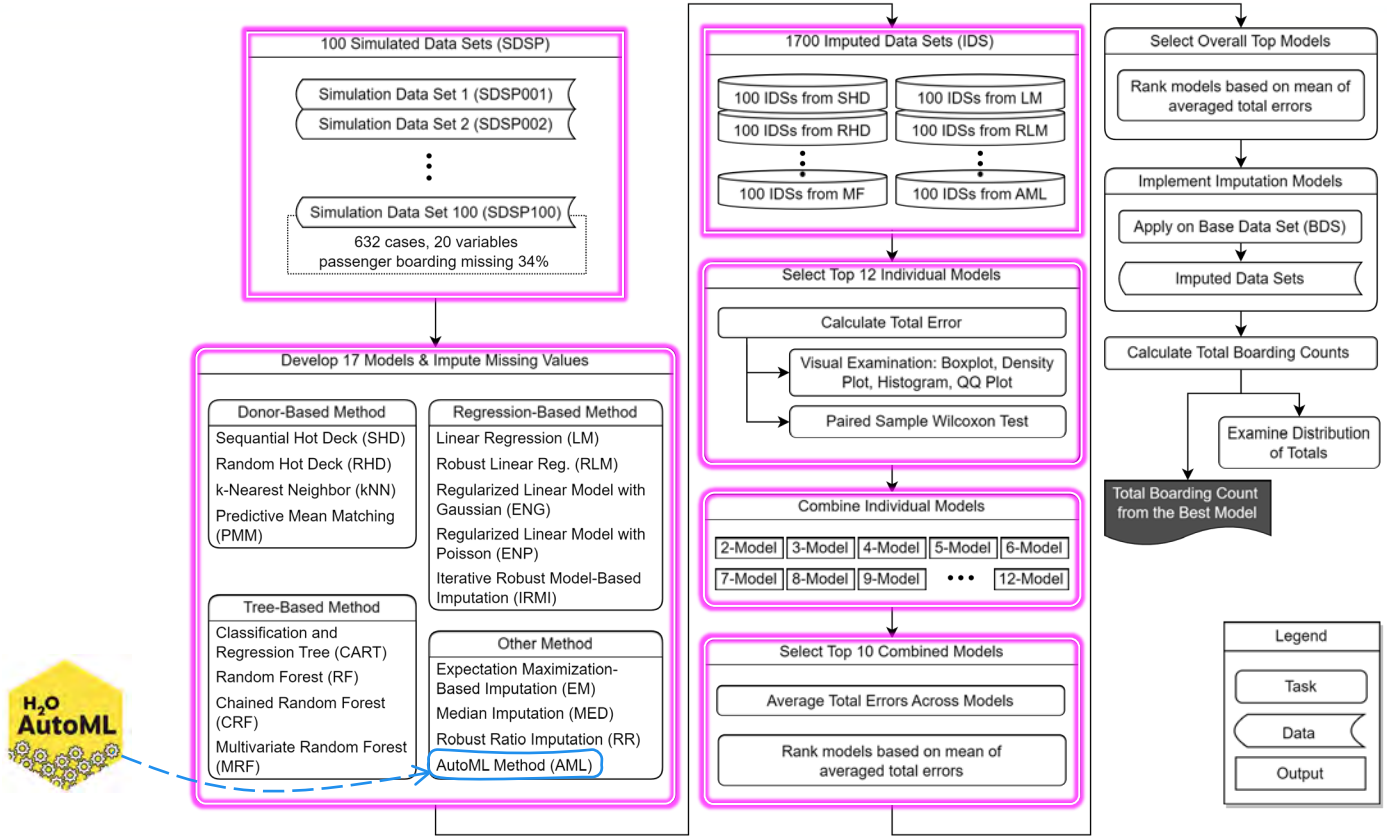
METHODS

U.S. Department of Transportation

Analysis Flow: Data Prep



Analysis Flow: Modeling



Individual Models

U.S. Department of Transportation

Donor-Based Method
Sequantial Hot Deck (SHD)
Random Hot Deck (RHD)
k-Nearest Neighbor (kNN)
Predictive Mean Matching (PMM)

Regression-Based Method
Linear Regression (LM)
Robust Linear Reg. (RLM)
Regularized Linear Model with Gaussian (ENG)
Regularized Linear Model with Poisson (ENP)
Iterative Robust Model-Based Imputation (IRMI)

Tree-Based Method
Classification and Regression Tree (CART)
Random Forest (RF)
Chained Random Forest (CRF)
Multivariate Random Forest (MRF)

Other Method
Expectation Maximization-Based Imputation (EM)
Median Imputation (MED)
Robust Ratio Imputation (RR)
AutoML Method (AML)

Combined Models

2-Model | 3-Model | 4-Model | 5-Model | 6-Model

7-Model | 8-Model | 9-Model | • • • | 12-Model

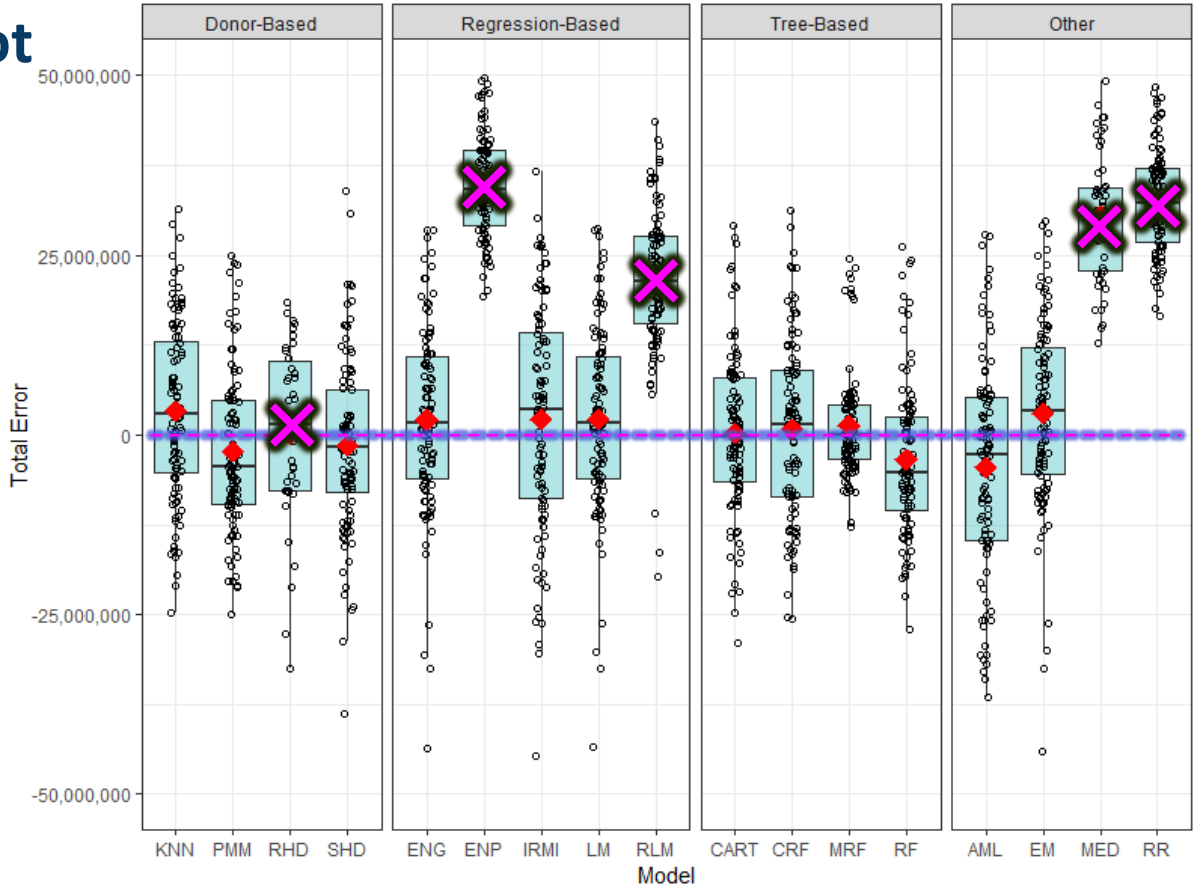
U.S. Department of Transportation

RESULTS

U.S. Department
of Transportation

Individual Models

- Boxplot



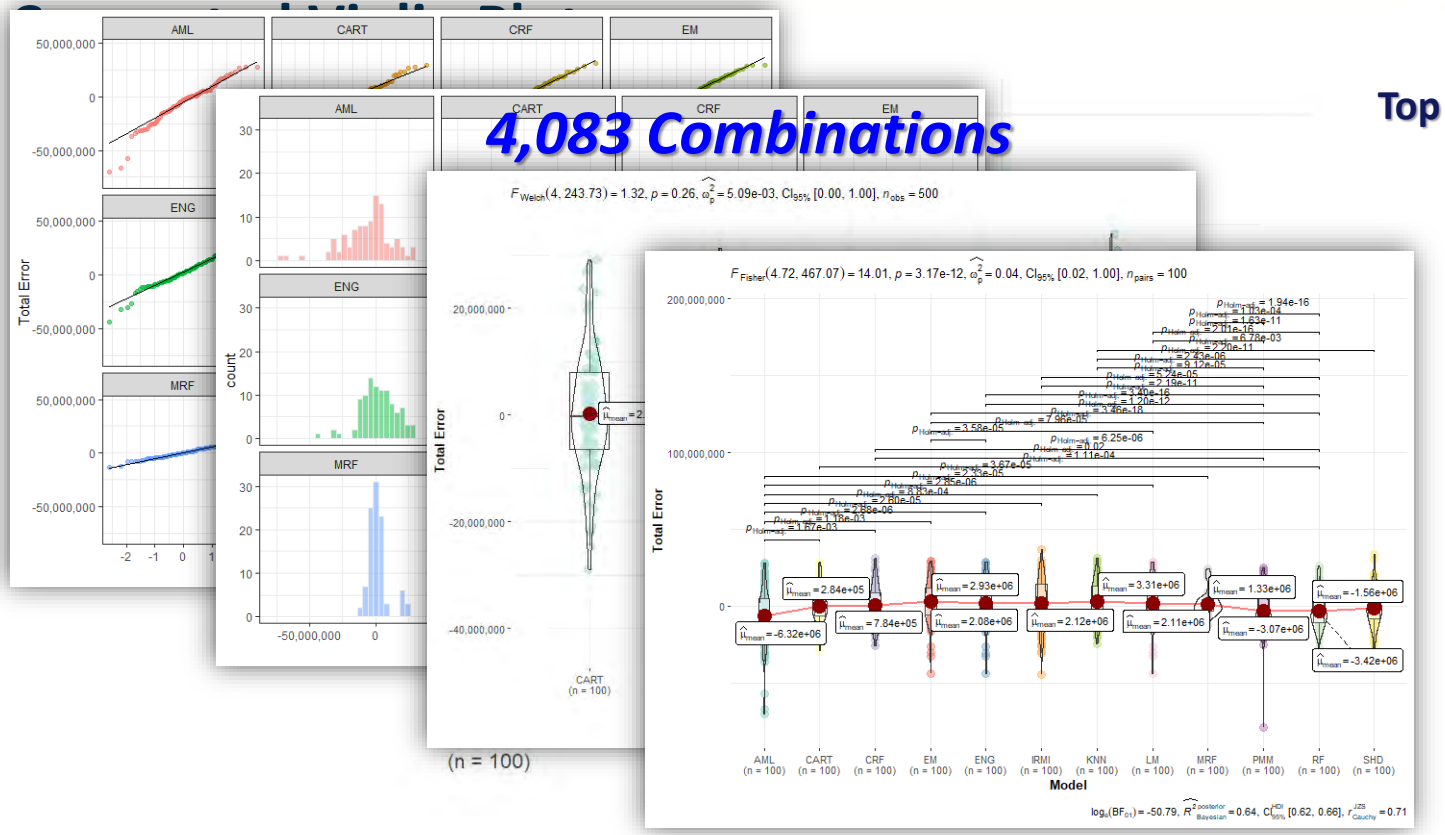
Top 12 Individual Models

1 CART
2 CRF
3 MRF
4 SHD
5 ENG
6 LM
7 IRMI
8 EM
9 PMM
10 KNN
11 RF
12 AML

Combined Models ❶

U.S. Department of Transportation

- Segmented Violin Plots



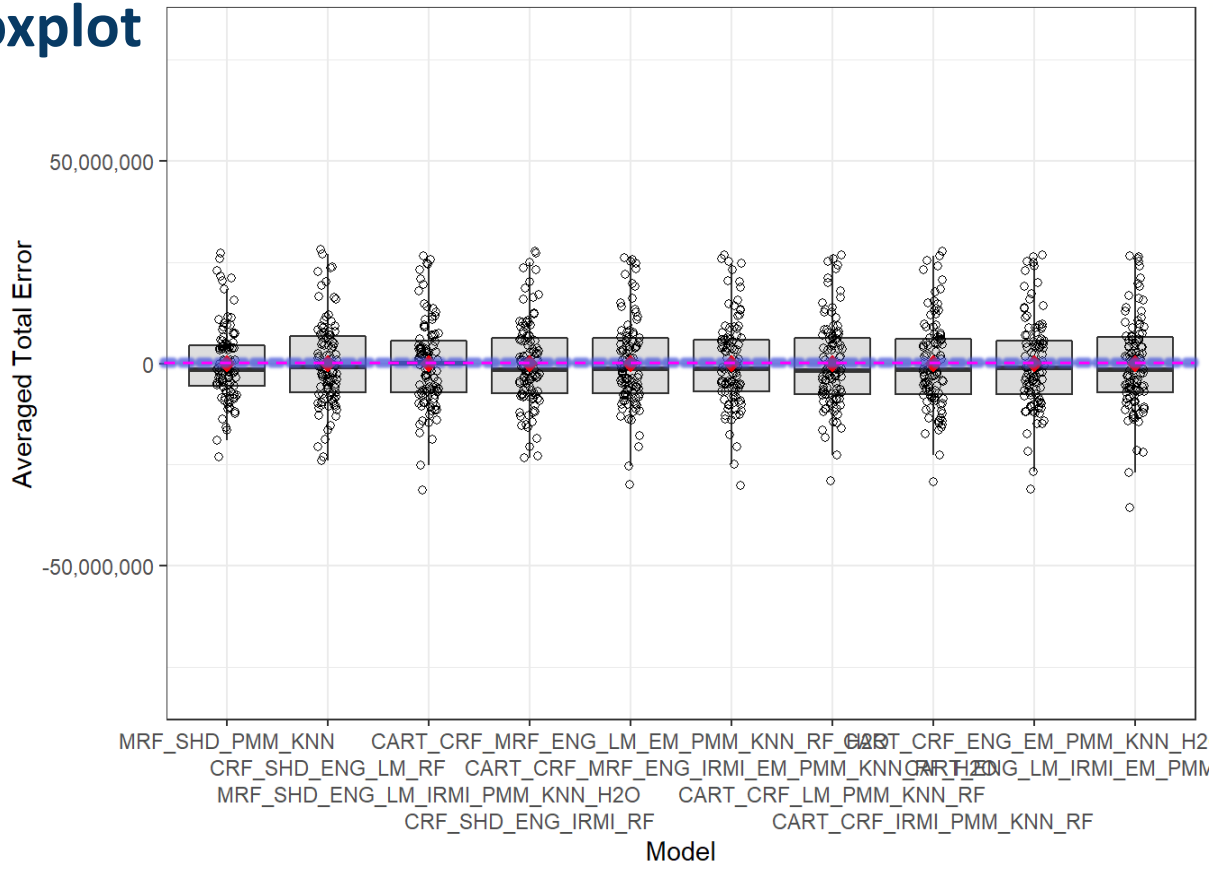
Top 12 Individual Models

1 CART
2 CRF
3 MRF
4 SHD
5 ENG
6 LM
7 IRMI
8 EM
9 PMM
10 KNN
11 RF
12 AML

Combined Models ❷

U.S. Department of Transportation

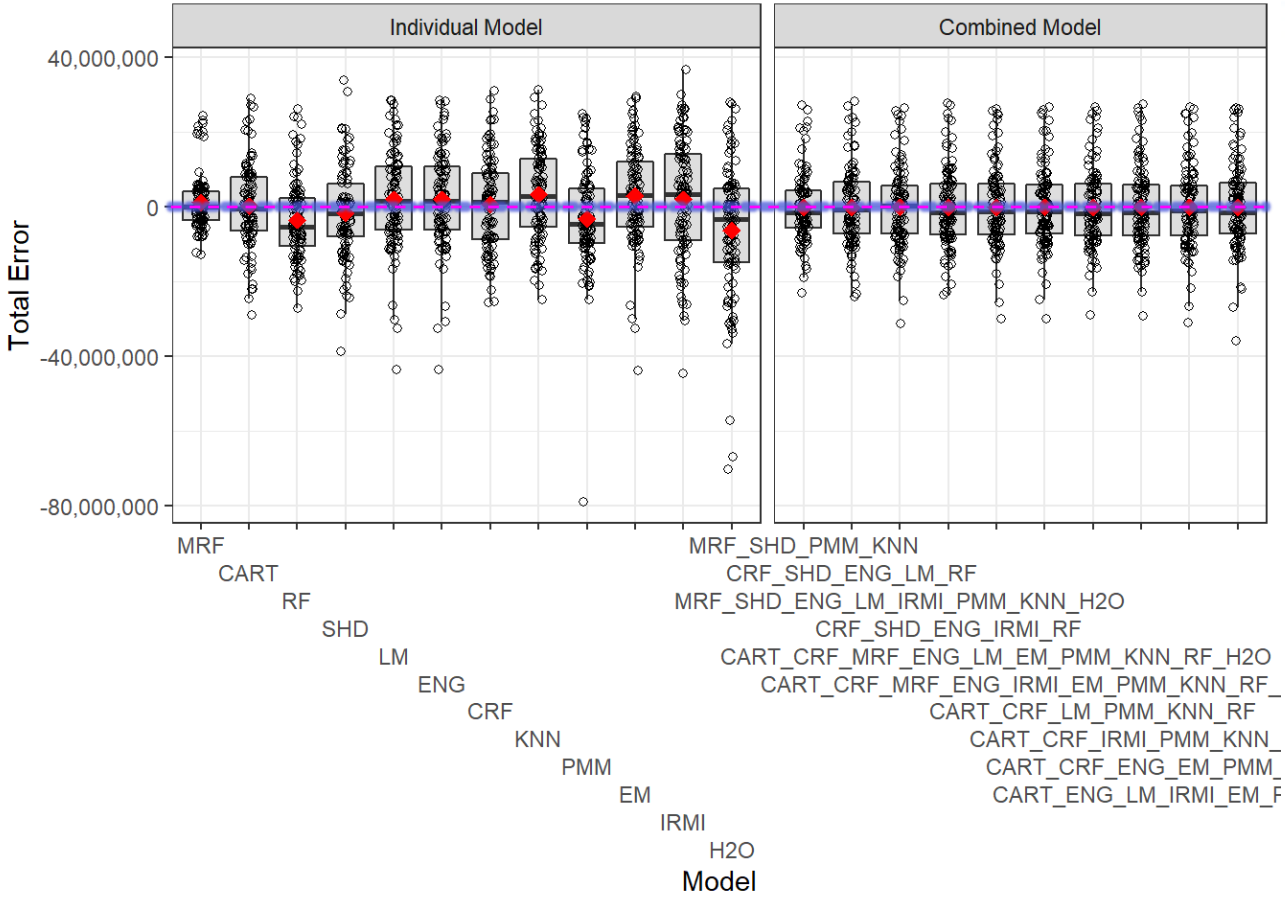
- **Boxplot**



Top 10 Models

1 MRF_SHD_PMM_KNN
2 CART_CRF_IRMI_PMM_KNN_RF
3 CART_CRF_LM_PMM_KNN_RF
4 CRF_SHD_ENG_IRMI_RF
5 CART_CRF_ENG_EM_PMM_KNN_H2O
6 CART_CRF_MRF_ENG_LM_EM_PMM_KNN_RF_H2O
7 CRF_SHD_ENG_LM_RF
8 CART_CRF_MRF_ENG_IRMI_EM_PMM_KNN_RF_H2O
9 MRF_SHD_ENG_LM_IRMI_PMM_KNN_H2O
10 CART_ENG_LM_IRMI_EM_PMM_KNN_RF_H2O

Top Individual & Combined Models



CONCLUSIONS

U.S. Department
of Transportation

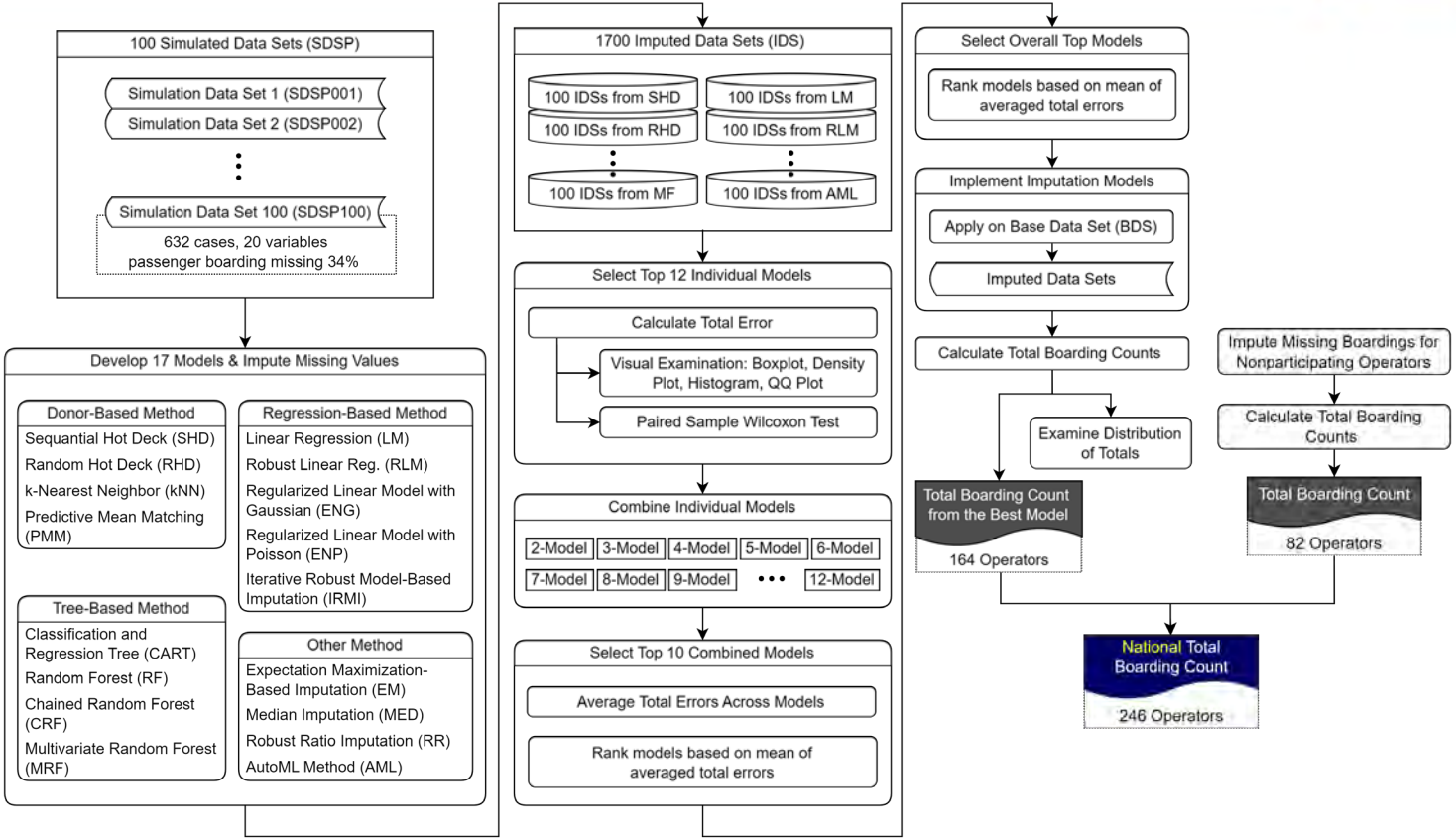
Conclusions

U.S. Department of Transportation

- Imputation for estimating **totals**
- Limited to this study

- Some models show very poor imputation performance: ENP, RLM, RR.
  - Attributable to the evaluation metric
- AutoML does not live up to expectation
- Averaging results from individual models stabilizes the distribution of total errors. But, majority of variation stays
- Thus, the best model cannot be determined statistically

What's Next?


100 Simulated Data Sets (SDSP)
Simulation Data Set 1 (SDSP001)
Simulation Data Set 2 (SDSP002)
Simulation Data Set 100 (SDSP100)
632 cases, 20 variables
passenger boarding missing 34%
Develop 17 Models & Impute Missing Values
Donor-Based Method
Sequantial Hot Deck (SHD)
Random Hot Deck (RHD)
k-Nearest Neighbor (kNN)
Predictive Mean Matching (PMM)
Regression-Based Method
Linear Regression (LM)
Robust Linear Reg. (RLM)
Regularized Linear Model with Gaussian (ENG)
Regularized Linear Model with Poisson (ENP)
Iterative Robust Model-Based Imputation (IRMI)
Tree-Based Method
Classification and Regression Tree (CART)
Random Forest (RF)
Chained Random Forest (CRF)
Multivariate Random Forest (MRF)
Other Method
Expectation Maximization-Based Imputation (EM)
Median Imputation (MED)
Robust Ratio Imputation (RR)
AutoML Method (AML)
1700 Imputed Data Sets (IDS)
100 IDSs from SHD
100 IDSs from LM
100 IDSs from RHD
100 IDSs from RLM
100 IDSs from MF
100 IDSs from AML
Select Top 12 Individual Models
Calculate Total Error
Visual Examination: Boxplot, Density Plot, Histogram, QQ Plot
Paired Sample Wilcoxon Test
Combine Individual Models
2-Model
3-Model
4-Model
5-Model
6-Model
7-Model
8-Model
9-Model
12-Model
Select Top 10 Combined Models
Average Total Errors Across Models
Rank models based on mean of averaged total errors
Select Overall Top Models
Rank models based on mean of averaged total errors
Implement Imputation Models
Apply on Base Data Set (BDS)
Imputed Data Sets
Calculate Total Boarding Counts
Examine Distribution of Totals
Total Boarding Count from the Best Model
164 Operators
Impute Missing Boardings for Nonparticipating Operators
Calculate Total Boarding Counts
Total Boarding Count
82 Operators
National Total Boarding Count
246 Operators

Acknowledgement

- Clara Reschovsky, NCFO Program Manager/Survey Statistician, USDOT-BTS
- Cha-Chi Fan, Office Director, USDOT-BTS

U.S. Department of Transportation

Contact

Young-Jun Kweon

young-jun.kweon@dot.gov

NCFO

ferry@dot.gov

Questions?

U.S. Department of Transportation