Nonresponse Bias Analysis of 2020 National Census of Ferry Operators

U.S. Department of Transportation
Office of the Secretary of Transportation

Bureau of Transportation Statistics

Recommended Citation

U.S. Department of Transportation, Bureau of Transportation Statistics. *Nonresponse Bias Analysis of 2020 National Census of Ferry Operators*. Washington, DC: 2024.

https://doi.org/10.21949/1531048

Acknowledgments

Bureau of Transportation Statistics

Patricia Hu
Director

Rolf Schmitt
Deputy Director

Produced under the direction of:

Cha-Chi Fan
Director, Office of Data Development and Standards

Project Manager

Clara Reschovsky
NCFO Program Manager

Visual Information Specialist

Alpha Wingfield

Major Contributors

Young-Jun Kweon

Clara Reschovsky

Aubrey Nguyen (formerly ORISE Fellow)

Other Contributors

Joseph McGill

Kenneth Steve

April Gadsby

Carl Cloyed

BTS information service contact information:

Ask-A-Librarian http://transportation.libanswers.com/

Phone 202-366-DATA (3282)

Quality Assurance Statement
The Bureau of Transportation Statistics (BTS) provides high quality information to serve government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. BTS reviews quality issues on a regular basis and adjusts its programs and processes to ensure continuous quality improvement.

Notice
This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for its contents or use thereof.

Table of Contents

List of Figures

List of Tables

Executive Summary

The Bureau of Transportation Statistics (BTS) conducts a biennial census of all ferry operators operating within the United States and its territories. To date, BTS has conducted the National Census of Ferry Operators (NCFO) in 2006, 2008, 2010, 2014, 2016, 2018, 2020, and 2023. The most recently released 2020 NCFO consists of five datasets that capture the 2019 ferry operational data: Operator, Operator Segment, Segment, Terminal, and Vessel. Efforts to enumerate ferry operations for the 2020 NCFO resulted in a frame of 246 operators for calendar year 2019. Among the 246 operators invited to the census, 164 operators (67 percent) participated by addressing at least one question in the NCFO. Of these 164 operators, 4 were deemed nonresponders for failing to answer enough questions to be counted among the unit responders—yielding a unit response rate of 65 percent.

According to the Office of Management and Budget's (OMB's) *Standards and Guidelines for Statistical Surveys* [2006], a nonresponse bias[1] analysis should be conducted when a unit response rate falls below 80 percent or an item response rate of a key item falls below 70 percent. Although the OMB Standards are for survey statistics, and NCFO is not a survey, BTS follows the recommendation and conducted a nonresponse bias analysis for the 2020 NCFO. Two goals of conducting the nonresponse bias study on the 2020 NCFO were to guide the improvement of data quality for future censuses and inform data users of potential bias in using the 2020 NCFO data. In this respect, the study performed three analytic tasks: (1) calculate four data quality metrics, (2) estimate nonresponse bias, and (3) identify variables influencing nonresponse.

Analysis results of nonresponse in the 2020 NCFO data led to the following conclusions:

- **Response rates varied across subgroups.** The overall unit response rate (URR) for the 2020 NCFO was 65 percent (160 out of 246 invited to the survey) but varied across subgroups according to reporting obligation and whether or not respondents accepted public funding. Operators reporting on behalf of governmentally owned and/or operated ferries or those accepting public funds had notably higher response rates than their counterparts by 21 and 15 percentage points, respectively. Consequently, data users focusing on a specific subgroup or making comparisons across subgroups should be cautious in interpreting the results.
- **A few large operators not responding to boarding counts were responsible for a large proportion of the bias in the total boarding counts of the respondents.** A ferry operator was considered large if they carried at least two million passengers or a half million vehicles annually. Nonresponse biases estimated an undercount of ~40 million passengers and ~2 million vehicles, which corresponds to 35 and 8 percent of the total observed boarding counts, respectively. Four large operators that failed to report their passenger boardings accounted for 60 percent of the total bias, underscoring the importance of obtaining boarding count data from large operators.

[1] Nonresponse bias occurs when individuals who do not respond to a survey or census are different from those who do respond in a way that the difference skews the results of the analysis of data collected.

- **A probable range of the total bias for the entire ferry population was estimatable with assumptions.** Assumptions were required in estimating the probable boarding counts for nonparticipating ferry operators. With these assumptions, the nonresponse bias estimated an undercount of 48 million to 78 million passengers from the total observed passenger boarding count for the entire ferry population of an estimated at 114 million passengers. Thus, the estimated national boarding counts in 2019 were between 162 million and 187 million passengers. The nonresponse bias in total vehicle boarding led to an undercount of 3.9 million to 9.2 million vehicles in comparison to the observed total of 26.6 million vehicles, yielding estimated boarding counts ranging from 31 million to 36 million vehicles.
- **Ferry operators in a specific state with certain characteristics were less likely to respond.** Based on conditional inference tree analysis, nonresponding operators were located in a specific state, and their terminals were located in heavily populated areas. Targeting these operators for outreach and follow-up could increase participation and response in future censuses.

The study findings and conclusions led to the following recommendations:

- **Develop a process to track response on each of the key items.** The process will identify nonresponding ferry operators during the census so that BTS can quickly follow up with these operators. This recommendation would increase the unit response rate and assist in collecting quality data.
- **Develop a list of ferry operators grouped by historical boarding counts.** The list will help BTS identify which ferry operators would be critical in obtaining boarding count data so that BTS can prioritize follow-up contacts based on the list. This recommendation would increase the unit response rate and the item response rates for the boarding count items.
- **Develop a process to identify abnormal changes in three key items (passenger boarding, vehicle boarding, and segment length).** This process will help to identify responding ferry operators whose values could suffer from input errors, such as accidentally adding or excluding a digit in boarding counts. When an abnormal change is identified, BTS will follow up with the corresponding operator to verify the veracity of its input, and in the case an error is found, the operator could correct it in a timely fashion. This recommendation would improve the quality of these three key data items.
- **Consider adding a question to the Segment Information section of the census questionnaire asking which cargo types (passengers, vehicles, and freight) are included at a segment level.** The proposed question in Segment Information section would improve data quality of the two boarding count items (passenger and vehicle boarding) and facilitate imputation of missing boarding counts by easily verifying zero vehicle boarding counts. This recommendation would improve the quality of data of the boarding counts.

1. Introduction

1.1. NATIONAL CENSUS OF FERRY OPERATORS (NCFO)

The Bureau of Transportation Statistics (BTS) conducts a biennial census of all ferry operators operating within the United States and its territories—the National Census of Ferry Operators (NCFO). Fixing America's Surface Transportation Act (P.L. 114-94, sec. 1112) requires BTS to maintain a database of existing ferry operations across the United States. In 2000, the Federal Highway Administration (FHWA) commissioned the Volpe National Transportation Center to survey all known ferry operations, leading to development of a national ferry database. Since then, BTS has conducted a data collection of all ferry operators and maintained the national ferry database. This database is an important source of information for various organizations and has been a key data source for developing BTS' Intermodal Passenger Connectivity Database. The database has also been a key input for the FHWA to allocate federal funds through the Ferry Boat Program. The FHWA uses segment length and boarding counts to allocate federal funds to operators who operate on regulated segments between at least one publicly owned terminal or in a publicly owned vessel.

To date, BTS has conducted the NCFO in 2006, 2008, 2010, 2014, 2016, 2018, 2020, and 2023. The 2020 NCFO was the most recently released census, consisting of five datasets capturing the 2019 ferry operational data: Operator, Operator Segment, Segment, Terminal, and Vessel. Efforts to enumerate ferry operations for the 2020 NCFO resulted in a frame of 246 operators for calendar year 2019. Of the 246 operators, 164 operators participated in the census; participation means an operator submitted response to at least one item in the NCFO. It should be noted the year of the NCFO indicates the year data was collected to estimate the previous year's ferry activities. For example, the 2020 NCFO represents ferry operations from all of 2019. However, from the 2022 NCFO forward, the year will indicate the year of ferry operations and not the year the data was collected. This change was made to align the NCFO with other BTS data products, such as the Commodity Flow Survey (CFS) and the Tank Car Report.

The NCFO is a census of all known ferry operations within the United States and its territories, encompassing the 50 States, Puerto Rico, the U.S. Virgin Islands, American Samoa, Guam, and the Commonwealth of the Northern Mariana Islands. In addition to ferry operations providing domestic service within the United States and its territories, operations providing services to or from at least one U.S. terminal are also included. Ferry operations included within the scope of the NCFO are those providing itinerant, fixed-route, common carrier passenger and/or vehicle ferry service. Railroad car float operations are also included within the scope of the NCFO. Details about the NCFO are found at https://www.bts.gov/NCFO.

1.2. NONRESPONSE BIAS STUDY

According to the Office of Management and Budget's (OMB's) *Standards and Guidelines for Statistical Surveys* [2006], a nonresponse bias analysis should be conducted when response rates fall below the following thresholds:

- 80 percent for a unit response rate,
- 70 percent for an item response rate of a key item, or
- 70 percent for a total quantity response rate.

According to Standard 3.2 of the OMB's Guidelines, "agencies must appropriately measure, adjust for, report, and analyze unit and item nonresponse to assess their effects on data quality and to inform users. Response rates must be computed using standard formulas to measure the proportion of the eligible sample that is represented by the responding units in each study, as an indicator of potential nonresponse bias." Although OMB Standards are for survey statistics and NCFO is not a survey, BTS followed the recommendation and conducted a nonresponse bias analysis for the 2020 NCFO to improve future NCFOs. It should be noted that a unit was a ferry operator, and an item is synonymous with a variable or a data element in this document.

1.3. STUDY PURPOSE & SCOPE

Two goals of conducting the nonresponse bias analysis on the 2020 NCFO were to guide the improvement of data quality for future censuses and inform data users of potential bias in using 2020 NCFO data. In this respect, the study performed three analytic tasks: (1) calculated four data quality metrics, (2) estimated nonresponse bias, and (3) identified variables influencing nonresponse. The scope is limited to available data used in the study: 2014, 2016, 2018, and 2020 NCFO data. For estimating the nonresponse bias in the total boarding counts, 2018 NCFO data were used to impute missing values of 2020 NCFO nonrespondents. To identify which variables influenced responses, 2014, 2016, and 2018 NCFO data were used to impute missing values of 2020 NCFO nonrespondents on several time-invariant variables. It should be noted that although NCFO data prior to 2014 were available, this study did not to include those data due to data quality concerns from those years and, instead, focused on the 2014, 2016, 2018, and 2020 data because they were consistent over the time period.

2. Data Description

Two data sources are used in this study, the NCFO and American Community Survey (ACS), and they are described separately here.

2.1. NATIONAL CENSUS OF FERRY OPERATORS (NCFO)

After data collection was completed, the data editing process was performed to assess validity of data, identify and correct erroneous data, and impute missing data. For the 2020 NCFO, the data year was 2019 and the data collection year was 2020. The editing process included several layers of extensive data review: automated data edits developed for the 2020 NCFO data to identify and remove duplicate or non-existent vessels, terminals, and segments [Nguyen 2022]; manual edits based on analysts' reviews and verification using external data sources;[2] and descriptive and distributional analysis for identifying abnormality[3] (e.g., Kweon [2021]); and ad-hoc analysis/review. After the collected data were processed and scrutinized, they were released in five separate files, called tables, through the BTS website, https://www.bts.gov/NCFO. The five tables of the 2020 NCFO include a total of 152 variables, some of which are found in more than one table, such as OPERATOR_ID found in four tables, and they are described in Table 1. Each table contains at least one indexed ID variable, which are used to link the tables for further analysis. The definitions of the variables are found at the 2020 NCFO Data Dictionary webpage.[4]

Table 1. Five Tables of the 2020 NCFO

Table	Description	Number of Variables[a]
Operator	Contains information about ferry operators and details about their operation.	50
Operator-Segment	Contains information related to route segments, such as segment length, average trip time, passenger boarding count, and season start and end dates.	41
Segment	Contains information about each route segment, such as terminal connections, type of geographic area served, and whether or not a National Park Service location is served.	7
Terminal	Contains information about ferry terminals, access mode(s), and operation entity.	20
Vessel	Contains information about ferry vessels, such as the passenger and/or vehicle capacity, speed, and fuel type.	34

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Data Dictionary*, available at https://cms.bts.gov/surveys/national-census-ferry-operators-ncfo/2020-ncfo-data-dictionary as of April 2024.

[a] The number of variables changes across NCFOs because new variables were added and existing variables were removed as the NCFO has evolved over the years (e.g., 2018 NCFO tables contain a total of 158 variables). Some variables (e.g., breadth, depth, and length of a vessel in the Vessel table) were created by BTS using data provided by ferry operators and external data, such as the U.S. Coast Guard (USCG) vessel database.

---

2 For example, when a reported city was found outside of a reported state, the reporting ferry operator's website was examined to verify the correct state and city.

3 For example, when a segment length changed from the previous census by more than 5 percent or a passenger boarding count changed by more than 10 percent, the corresponding record was flagged for further examination.

4 https://cms.bts.gov/surveys/national-census-ferry-operators-ncfo/2020-ncfo-data-dictionary

2.2. AMERICAN COMMUNITY SURVEY (ACS)

There is a perceived connection between population estimates and the urban/rural or metropolitan area status of the areas where ferry terminals are located and nonresponse in the NCFO. For example, ferry operators whose terminals are located in areas with large, dense populations might be less likely to respond to the census than their counterparts in areas with smaller populations or vice versa. Location-based population estimates are available in ACS data, and an ACS data file containing 5-year estimates from 2014 to 2019 was obtained. The 2019 county-level population counts, except for the Canadian (provincial counts), BVI (island counts), and AS/USVI (island counts) terminals where only other population counts were available, were used in the study. The 2019 data file was loaded into ArcGIS and matched with location information from the terminal table of the 2020 NCFO.[5] Populations and classifications of metropolitan areas were brought into the study.

[5] 2020 NCFO was conducted in 2020 to collect 2019 data.

3. Methods

3.1. DEVELOPING ANALYSIS DATA SETS

There are five NCFO tables publicly available at the BTS website (Table 1) and there is an internal table containing confidential data, such as passenger boarding counts for ferry operators that requested nondisclosure. These six NCFO tables were processed separately (Section 3.1.1. Processing NCFO Data Sets), and ACS population data were processed for merging with the NCFO data (Section 3.1.2. Processing ACS Data Set). After all individual data sets were processed, they were merged into analysis data sets (Section 3.1.3. Merging NCFO Data Sets). R language was used to execute the data processing.

3.1.1. Processing NCFO Data Sets

Each of the six NCFO data sets, also called tables, were processed separately. Specifically, variables deemed not useful were removed, "NA" was replaced with numeric 0 for some variables where "NA" is inferred as numeric 0,[6] and variable types were changed for further processing and analysis (e.g., character to numeric type). All ID variables, such as OPERATOR_ID, SEGMENT_ID, TERMINAL_ID, and VESSEL_ID, were preserved for data merging. New variables were created to indicate whether a valid value was recorded for a specific variable of interest, such as passenger boardings and segment length. Also, a variable to indicate a unit response status was created based on the definition established for the 2020 NCFO, which is described in the Section 3.2. Determining Unit Response.

3.1.2. Processing ACS Data Set

The 2014–2019 American Community Survey 5-year data file B01003 for Total Population for all counties in the United States was obtained from Census Bureau's website on April 28, 2022, and loaded into ArcGIS for aggregating to match terminal locations from the 2020 NCFO. Additionally, population was obtained for U.S. territories from the 2020 Decennial Census population counts for the Virgin Islands' and American Samoa's Districts. Puerto Rico's municipalities are included in ACS. Population and land area were obtained from the 2021 Canadian census for the provinces that contain ferry terminals with segments that connect to the United States. The U.S. land areas of counties were obtained from the Census Bureau's TigerLine shapefiles. In ArcGIS, for counties (or equivalent) that contain one or more ferry terminals, the population density was calculated, and the county was matched to a metropolitan/ micropolitan flag and an urban/rural flag.

---

[6] For example, there were eight revenue variables corresponding to eight revenue categories, such as ticket sales, advertisement, and federal fund. A respondent was asked to enter percentages of its total revenue across the eight categories. Some respondents entered nonzero numbers in some categories and left a blank in other categories. These blank cases were recorded as "NA" in the survey instrument and assumed to imply 0 percent. For example, if a respondent entered 100 in ticket sales category and left the remaining revenue categories blank, this means the revenue of its ferry operation came entirely from ticket sales (i.e., 100 percent for the ticket sales category) and there were no other revenue sources (i.e., 0 percent for the other revenue categories).

The processed population data set includes FIPS code, land area (squared miles), total population, and population density (population per square mile). A rural/urban categorical variable with three levels was created based on the 2010 urban/rural criteria[7] of the U.S. Census Bureau as follows:

- Urbanized Areas (UAs) of 50,000 or more people.
- Urban Clusters (UCs) of at least 2,500 and fewer than 50,000 people.
- Rural areas of all areas not included in an urban area.

3.1.3. Merging NCFO Data Sets

When individual data sets were processed as described in the previous sections, the individual NCFO data sets were merged to produce analysis data sets. The data sets were prepared at two different levels, operator and operator-segment-terminal-vessel. The operator level means each row in the data set corresponded to a ferry operator so that the prepared data set had a total of 164 rows. Meanwhile, the operator-segment-terminal-vessel level means each row of the data corresponded to a segment[8] where an operator used a vessel to ferry between two terminals and the prepared data set had 1,005 rows. Among these rows, 42 rows had at least 1 unmatching component. For example, 34 rows had no reported segments of 17 ferry operators.[9]

To prepare the operator-level analysis data sets, individual NCFO data sets were aggregated at operator level and a new variable recording the number of segments for each operator was created during the aggregation. A new variable recording the number of vessels serving each segment was created in Operator-Segment data set by counting nonmissing values across the 26 vessel ID variables (i.e., VESSEL1_ID through VESSEL26_ID) for each row.

Individual data sets were merged using ID variables. Operator-Segment data set was merged with Operator data set using OPERATOR_ID. The resulting data set was then merged with Segment data set using SEGMENT_ID. Terminal data set was then merged with the resulting data set twice, first for the origin terminal (Terminal 1) and then for the destination terminal (Terminal 2). Vessel data set was then merged using VESSEL_ID mapped to VESSEL1_ID in the resulting data set. Finally, the population data set was merged by terminal location information. It should be noted that there were ferry operators who did not report all of their segments, terminals, and/or vessels—meaning the final merged data set had missing values in some of these ID variables. For example, when an operator did not report one of its two segments but reported terminals not mapped to the reported segment in Terminal data set, the final data set had missing values in SEGMENT_ID.

---

7 https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html
8 A segment is a pair of origin and destination terminals.
9 These rows were found with only one terminal reported.

3.2. DETERMINING UNIT RESPONSE

Of the 246 ferry operators sent surveys, 164 operators (67 percent) were counted as "participating" in the 2020 NCFO by addressing one or more census questions. But only 160 of these operators (65 percent) addressed questions at a level deemed acceptable for inclusion among those operators delivering (unit) responses.

Prior to the 2020 NCFO, all ferry operators who submitted data to the census, regardless of how many questions were addressed, were determined as (unit) "responses," and thus were counted in calculating the response rate of the census. The 2020 NCFO makes a distinction between those who provided data at an acceptable level and those who did not. It should be noted a (ferry) operator and a respondent are used interchangeably throughout this document. However, there are few cases where an operator means a company/organization operating a ferry service, not a respondent, and this distinction is clear in the contexts where the term, operator, is used.

As for the 2020 NCFO, data deemed to be at an "acceptable" level is defined as including at least two segments and at least one vessel: Accordingly, an operator is determined to have delivered a "response" when the following three criteria are met:

- A respondent should enter the operator's name.[10]
- A respondent should report a minimum of two segments.
- A respondent should report a minimum of one vessel.

In the 2020 NCFO questionnaire, a respondent is first asked to provide a ferry operator's name (Figure 1-B). The respondent is asked to list all vessels in its fleet in Question 5 (Figure 1-B), and to provide up to three of the most used vessels for each segment in Question 19 (Figure 1-C). To satisfy the three criteria for "response," a respondent should provide an operator's name in Question 1 and list at least one vessel in Question 5 and two segments in Question 19. As long as the respondent fills in "Company | Operator Name", in Question 1 and "Vessel Name" in Question 5 for at least one vessel and "Route Origin" and "Route Destination" in Question 19 for at least two segments, the operator is determined to be "response." This means even if the respondent provides no information other than Operator Name, Vessel Name, Route Origin, and Route Destination, that information is still determined a "response."

[10] Operator name is prefilled for a ferry operator included in the frame. However, a new ferry operator may have come into existence unbeknownst to BTS and thus was not included in the frame. For such an operator, the operator's name is blank and must be entered.

Figure 1. Three Questions Used To Determine Incidence of an Operator Response

A. Question 1 in Operator Information section

1. Please provide the following information on your operation.

Company | Operator Name:	
Address Line 1:	
Address Line 2:	
City:	
State/Territory/Province:	
Zip Code:	

B. Question 5 in Vessel Information section

5. Please list and provide the vessel number and call sign for each vessel in your fleet during calendar year 2019 (include unpowered barges and powered tugs used for ferry service).

	Vessel Name	USCG Vessel Number	MMSI Number	Call Sign
1				
2				
3				
4				

C. Question 19 in Segment Information section

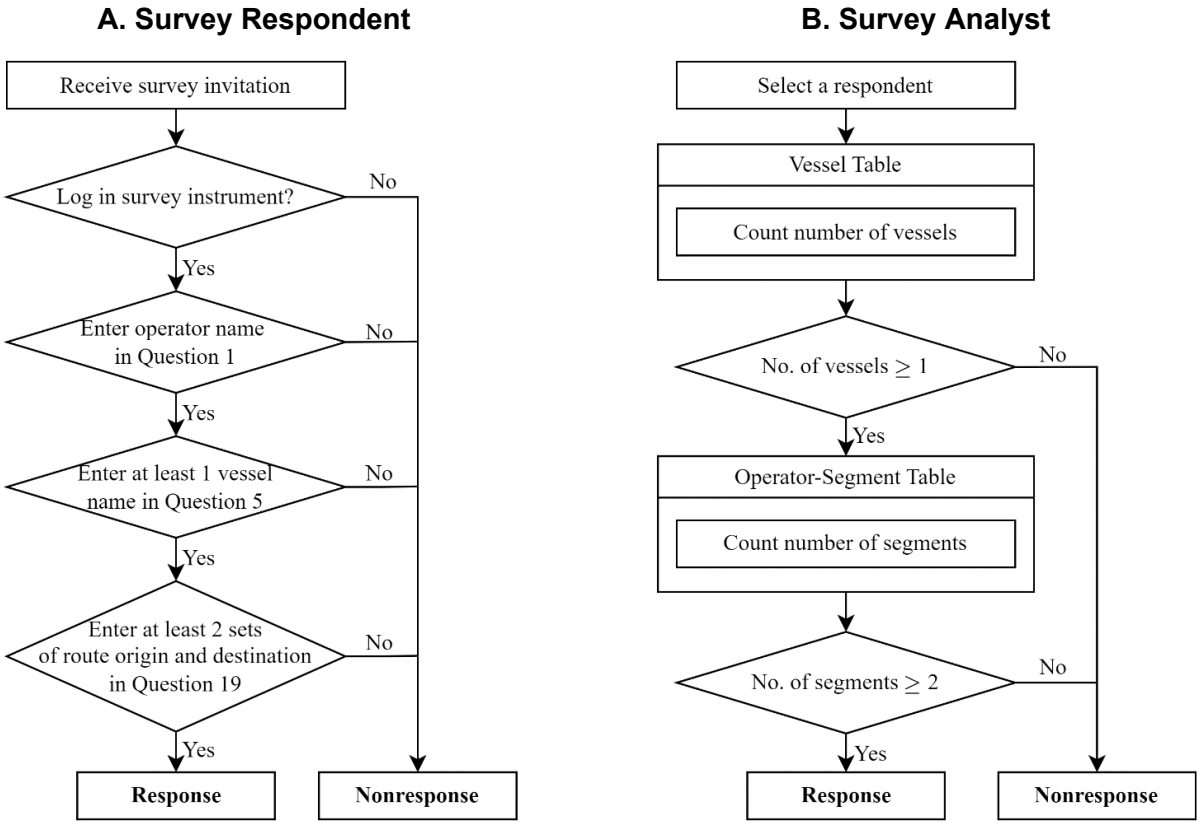
19. For each route segment, please list the name of the vessels MOST used on the segment in calendar year 2019.

	Route Origin	Route Destination	Vessel 1 (most used vessel)	Vessel 2 (2nd most used vessel)	Vessel 3 (3rd most used vessel)
1					
2					
3					
4					

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *Questionnaire of 2020 National Census of Ferry Operators*.

Figure 2 visualizes how "response" was determined from two perspectives: (a) a survey respondent and (b) a survey analyst. Figure 2-A shows the process of determining "response" for each respondent from a standpoint of a survey respondent filling out the NCFO questionnaire. Meanwhile, Figure 2-B shows the process from a standpoint of a survey analyst calculating a unit response rate of the NCFO. From the survey respondent's standpoint (Figure 2-A), the respondent first logs into the survey instrument after receiving the survey invitation. When the respondent provides the ferry operator's name in Question 1, at least one vessel name in Question 5, and at least two sets of origin and destination in Question 19, it was determined as "response."

Figure 2. Determining Unit Response for the 2020 NCFO



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, 2024.
Note: The order of vessel and segment elements can switch in part B.

From the survey analyst's standpoint (Figure 2-B), the process was based on the tables released to the public, specifically two tables, Vessel and Operator-Segment tables. Since an operator name in Operator table always existed, the step corresponding to Question 1 was not needed and was not shown in Figure 2-B. In the Vessel table, the number of reported *vessels* for each operator was calculated. In the Operator-Segment table, the number of reported *segments* for each operator was calculated. An operator was determined to submit a "response" only when the respondent reports at least one vessel and at least two segments. It should be noted that the order of two of the diamond decision boxes can be switched— the number of segments can precede the number of vessels.

It is also noteworthy that there might be a case where the vessels and segments reported by a respondent may not be matched. Even in such a case, the operator was determined to be a "response" as long as the three criteria are met. For example, a certain respondent may report two vessels and two segments in operation. If the most used vessel reported for the two segments is not matched with any of the two reported vessels, the operator is still determined to provide a "response" since it satisfies the three criteria for "response" in the NCFO.

3.3. SELECTING 19 KEY VARIABLES

In order for BTS to fulfill the core mission of the NCFO, a suitable level of responses in the data is required. To determine what constitutes that level for the 2020 NCFO, 19 key variables were selected (Table 2). There are cases where an operator was determined to supply a "response" but may not have provided enough valid information on all 19 variables for that response to rise to a suitable level.

Table 2. 19 Key Response Variables of 2020 NCFO

Key Response Variable	Definition	Table
Accept public funding	Indicate whether the ferry operator accepts public funding	Operator
Operator city	City for the ferry operator mailing address	
Operator state	State for the ferry operator mailing address	
Operator name	The complete company name of the ferry operator	
Average trip time	Average trip time of a segment in minutes	Operator-Segment
Most used vessel	Vessel most often used for the route	
Passenger boarding	Total passenger boardings for the census year	
Segment length	Segment length in nautical miles	
Season end date	Seasonal service end date	
Season start date	Seasonal service start date	
Vehicle boarding	Total vehicle boardings for the census year	
Segment name	Name of the segment	Segment
Terminal 1 city	City in which the ferry origin terminal is located	Terminal
Terminal 2 city	City in which the ferry destination terminal is located	
Terminal 1 state	State in which the ferry origin terminal is located	
Terminal 2 state	State in which the ferry destination terminal is located	
Terminal 1 name	Name of the ferry origin terminal	
Terminal 2 name	Name of the ferry destination terminal	
Vessel name (of most used vessel)	Name of the vessel indicated most used in Operator-Segment table	Vessel

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, 2020 NCFO Data Dictionary, available at https://cms.bts.gov/surveys/national-census-ferry-operators-ncfo/2020-ncfo-data-dictionary as of April 2024.

3.4. MEASURING DATA QUALITY

Four rates were used to measure different aspects of data quality for the 2020 NCFO: (1) Unit Response Rate (URR), (2) Item Response Rate (IRR), (3) Total Item Response Rate (TIRR), and (4) Modified Quantity Response Rate (MQRR). The URR and IRR were from American Association for Public Opinion Research's (AAPOR's) Standard Definitions [AAPOR 2016] and their technical definitions were adapted from those in *U.S. Census Bureau Statistical Standards* [U.S. Census Bureau 2022]. The four rates were explained and defined in subsections.

It is worth noting that a unit was a ferry operator, and an item is synonymous with a variable or a data element in this document. Typically, an item also corresponded to a question. But, as seen in Figure 1, one question can correspond to several variables. For example, Question 19 (Figure 1-C) is mapped to five variables in the public tables: SEG_TERMINAL1_ID (Route Origin), SEG_TERMINAL2_ID (Route Destination), VESSEL1_ID (Vessel 1), VESSEL2_ID (Vessel 2), and VESSEL3 (Vessel 3).

3.4.1. Unit Response Rate (URR)

The Unit Response Rate (URR)[11] is the primary data quality metric of the NCFO and the most frequently used rate for a survey or census. It is typically expressed as a percentage ranging from 0 to 100 percent. OMB's *Standards and Guidelines for Statistical Surveys* [2006] uses the URR as one of the metrics for recommending a nonresponse bias study.

The URR is the proportion of units that were eligible and responded to the survey (expressed as a percentage) and is computed as follows:

$$URR = \left(\frac{R}{E + U}\right) \times 100\% \quad (1)$$

Where:

- *R* (Response) = the number of units that were eligible for data collection and determined to constitute a "response." A unit was determined as a response when it satisfied the three criteria noted in Section 3.2. Determining Unit Response and Figure 2 visualizes the determining process.
- *E* (Eligible) = the number of units that were eligible for data collection. These included chronic refusal units (e.g., eligible reporting units having notified BTS that they do not participate in the census).
- *U* (Unknown Eligibility) = the number of units for which eligibility could not be determined. For example, the email inviting to the census that was sent to a ferry operator was bounced back with a mail delivery failure notice such as "address not found" and BTS was not able to reach the operator in a follow-up contact via email and/or phone call. In such a case, the eligibility of the operator could not be determined. Also, the eligibility of a ferry operator for the NCFO could not be determined when an invitation email was successfully delivered but the respondent did not log into the online census instrument.

3.4.2. Item Response Rate (IRR) & Total Item Response Rate (TIRR)

The Item Response Rate (IRR)[12] is an item-specific data quality metric quantifying how many units respond to a specific item; thus, the IRR is calculated for each item of interest. The IRR is the proportion of the number of units with a valid response to item *x* to the number of units that require a response to item *x*. TIRR[13] intends to provide item-specific data quality from the standpoint of the entire census by multiplying the IRR by the URR. IRR and TIRR are calculated as follows:

---

[11] URR is defined in Section 2.1 of App D3-A Demographic Surveys and Decennial Censuses Response Rates in *U.S. Census Bureau Statistical Standards* [U.S. Census Bureau 2022] and its terms are defined in Sections 1.1-1.2. The rate is equivalent to APPOR Response Rate 2 (RR2) (AAPOR, 2016).

[12] The rate and its terms are defined in Section 3.1 of App D3-A in *U.S. Census Bureau Statistical Standards* [U.S. Census Bureau 2022].

[13] The rate and its terms are defined in Section 3.1 of App D3-A in *U.S. Census Bureau Statistical Standards* [U.S. Census Bureau 2022].

$$IRR_x = \left(\frac{ITEM_x}{IREQ_x}\right) \times 100\% \quad (2)$$

$$TIRR_x = (IRR_x \times URR) \div 100\% \quad (3)$$

Where:

- $ITEM_x$ = the number of units with a valid response to item *x*.
- $IREQ_x$ = the number of units whose response is required for item *x*.

A response was required for item *x* unless it is a valid skip item. For example, it was a valid skip when a ferry operator did not report vehicle capacity for a vessel that did not carry vehicles.

For NCFO, an IRR was calculated by table and thus "unit" meaning differs depending on the table that the IRR is calculated for due to the data structure of NCFO. Operator table has only one record (i.e., one row) for each ferry operator. However, other tables could have multiple records (rows) for the same ferry operator. For example, a ferry operator needs at least two records[14] in the Terminal table because a set of an origin and destination terminals is required at minimum (i.e., one record for the origin terminal and the other record for the destination terminal). A unit for calculating the IRR for an item in Operator table is a ferry operator while that for an item in Segment table is a segment, that in Terminal table is a terminal, and that in Vessel table is a vessel. Thus, the number of units with respect to calculating the IRR varies across the tables.

3.4.3. Modified Quantity Response Rate (MQRR)

The URR, IRR, and TIRR measure the data quality based on the count of responses. In an establishment survey, such as the NCFO, values of the responses are also important to be considered. For example, if only 10 percent of the ferry operators responded to the annual passenger boarding item, the IRR for the passenger boarding is 10 percent. Suppose these 10 percent are very large in terms of fleet size and the number of serving terminals, and thus their total passenger boarding covers 90 percent of the total passenger boarding of all the ferry operators. This coverage of 90 percent in terms of values of the responses is also important to be reported. The Modified Quantity Response Rate (MQRR) is intended to measure the quantity aspect of the data quality and is an item-level metric of the quality of quantity.

The MQRR is devised based on the Quantity Response Rate (QRR)[15] defined in *U.S. Census Bureau Statistical Standards* [U.S. Census Bureau 2022]. The QRR is mainly for an economic survey/census where a survey unit would be different from a tabulation unit and weighting and adjustments are involved. Thus, the QRR is not applicable to NCFO in its current form. A modified version of the QRR is devised reflecting characteristics of the NCFO. The MQRR is the proportion of the observed total of item *x* (expressed as a percentage) and is calculated as follows:

---

[14] There might be a case where only one record in Terminal table is found for a ferry operator. This is probably because the operator failed to report the pairing segment information.

[15] The definition and terms of QRR are found in Sections 1.4 and 2.1 of App D3-B Economic Surveys and Censuses Response Rates in *U.S. Census Bureau Statistical Standards* [U.S. Census Bureau 2022].

$$MQRR_x = \left(\frac{T_{x,Obs}}{T_{x,Est}}\right) \times 100\% \quad (4)$$

Where:

- $T_{x,Obs}$ = observed total of values on item x in the census.
- $T_{x,Est}$ = estimated total for item x.

$T_{x,Obs}$ is calculated by summing all values of item x in the final internal tables[16]. $T_{x,Obs}$ includes only values for ferry operators that responded to the census. Meanwhile, $T_{x,Est}$ adds imputed values for ferry operators who did not respond to the census or item x. Thus, $T_{x,Est} = T_{x,Obv} + T_{x,Imp}$, where $T_{x,Imp}$ = sum of imputed values of item x for nonresponding ferry operators eligible for the NCFO. For example, when item x is passenger boarding, $T_{x,Obs}$ is sum of all reported passenger boarding counts and thus is regarded as the *observed* total passenger boarding. Meanwhile, $T_{x,Est}$ adds total of imputed passenger boarding counts for nonresponding ferry operators. Imputation was performed based on the growth ratio of 2018 and 2020 NCFO and is discussed in the next section.

3.5. ESTIMATING NONRESPONSE BIAS

3.5.1. Definition of Bias

Bias of an estimator is defined as the difference between the population parameter value (i.e., true value) and the expected value of the estimator of the parameter and is expressed as follows:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta \quad (5)$$

Where:

- θ = true parameter value.
- $\hat{\theta}$ = estimator of the parameter.
- $E(\cdot)$ = expected value.

In a laboratory setting, a bias is the difference between the average of measurements made on the same object and its true value [National Institute of Standards and Technology 2012] and the average of measurements corresponds to the expected value of the estimator in Equation 5. Since the true value is rarely available in practice, it is often estimated under certain conditions.

Nonresponse bias is a specific type of bias in survey statistics due to nonresponse. This bias occurs when we fail to collect data for a subset of units selected for the survey and respondents are meaningfully different from nonrespondents. For a sample mean, the bias of the sample respondent mean is estimated as follows [OMB 2006]:

[16] The sum can be larger than the sum released in the public tables because some values in the public tables are suppressed due to confidentiality.

$$Bias(\bar{x}_R) = \bar{x}_R - \bar{x}_S \quad (6)$$

Where:

- $\bar{x}_S$ = mean based on all sample cases.
- $\bar{x}_R$ = mean based only on respondent cases.

In a survey program, $\bar{x}_R$ is calculated with the current sampling weight and $\bar{x}_S$ is calculated with the sampling weights adjusted for nonresponse. Thus, the nonresponse bias is the difference between the mean based on values of only respondents and the mean based on values of the full sample (i.e., respondents and nonrespondents). Since the nonrespondents do not provide their values, they must be estimated by weight adjustments for nonresponse, called nonresponse adjustment weights.

This study was different in two aspects from a study where the above equation can be applied directly. First, this study was primarily interested in the bias in the *total* of respondents' values, not the mean. Second, the NCFO is a *census*, not a survey, meaning nonresponse adjustment weights were not developed for the NCFO. Considering these two distinctions, this study modified Equation 6 as follows to fit the study purpose:

$$Bias(T_{x,R}) = T_{x,R} - T_{x,C} = T_{x,NR} \quad (7)$$

Where:

- $T_{x,R}$ = total of values in x from respondents.
- $T_{x,C}$ = total of values in x of the census (including respondents and nonrespondents).
- $T_{x,NR}$ = total of values in x of nonrespondents.

The total of the census, $T_{x,C}$, is calculated as a sum of $T_{x,R}$ and $T_{x,NR}$. Because values of *x* for nonrespondents are missing, they must be imputed to calculate $T_{x,NR}$. It should be noted that not all missing values could be imputed, especially for perennial nonrespondents who had not participated in the several past censuses[17].

3.5.2. Measure of Size (MOS) Grouping

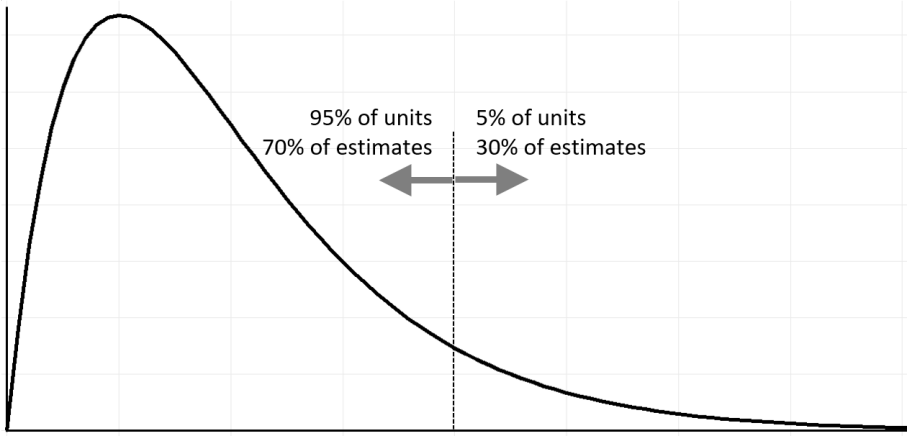
There were two variables of primary interest in the NCFO nonresponse bias estimation, passenger boarding and vehicle boarding. Imputing the boarding counts could be done at either the operator-segment-terminal-vessel level or operator level. Imputation at the operator-segment-terminal-vessel level requires more detailed data and more intensive efforts in preparing the data than that at the operator level. Imputation at the operator level implies that the total boarding count for a ferry operator was imputed. Meanwhile, when imputation at the operator-segment-terminal-vessel level was performed, a boarding count for each of the segments under the same operator should be imputed first and the boarding counts then summed across the segments to calculate the total boarding count for the ferry operator. This study performed imputation at the operator level since the efforts for preparing the data for imputation were deemed suitable for this study, and a rough estimate of the bias would be

[17] For these operators, only their contact information is known.

adequate for the study; it should be noted that a separate study focusing on imputing boarding counts at the operator-segment-terminal-vessel level is currently being performed.

According to Lineback and Thompson [2010], business/establishment surveys were often distinctively different from demographic surveys because their data, such as the amounts of sales and products, could be severely right skewed (Figure 3). A right skewed distribution implies that a small portion of units (e.g., businesses or establishments) dominates the data values (e.g., only 5 percent of units comprises 30 percent of the total value in Figure 3) and this was true for the NCFO; some of the largest operators were governments. In addition, there was a concern about imputation quality that could vary considerably by the size of an operator in terms of ridership. Addressing these issues, the study proposed grouping ferry operators based on the boarding counts being measure of size (MOS) and imputing missing boarding counts by group. For example, ferry operators could be grouped into small or large groups based on passenger boarding. This method certainly does not produce results with a high accuracy. However, it was a practically viable method considering the limited time for the study. Imputation can be performed separately for the small and large groups.

Figure 3. Right-Skewed Distribution of Business Statistics



Source: Lineback and Thompson [2010].
Note: This figure was recreated based on Figure 1 in Lineback and Thompson [2010].

Grouping was determined heuristically based on a combination of natural breaks in the distribution, proportion of the total MOS value of a group, and clean numbers. For example, of passenger boarding, if a natural break occurred at 135,258, either 100,000 or 150,000 would be a cleaner break. Suppose a threshold of 100,000 for the small group results in 2 percent of all the operators being included while that of 150,000 results in 10 percent being included. A threshold of 150,000 would be a better choice than 100,000 because 2 percent might be viewed as too small to form a separate group. Due to the relatively rough nature of the estimation, three to five groups were considered appropriate.

3.5.3. Imputation

Imputation was performed using a growth ratio of 2018 and 2020 NCFOs by each MOS group. This growth ratio method is based on a strong assumption that a nonresponding ferry operator in the 2020 NCFO experienced the average growth of the responding operators in the same MOS group. To calculate the growth ratio, boarding counts in both census years should be known. The growth ratio and rate were calculated as follows:

$$Growth\ Ratio_g(x) = \frac{T_{x,R,g,2020}}{T_{x,R,g,2018}} \tag{8}$$

$$Growth\ Rate_g(x) = \frac{T_{x,R,g,2020} - T_{x,R,g,2018}}{T_{x,R,g,2018}} \times 100\% \tag{9}$$

Where:

- *R* = respondents only.
- *g* = MOS group (e.g., small, medium, and large).
- $T_{x,R,g,2020}$ = total of values in *x* of respondents in group *g* in census year 2020.
- $T_{x,R,g,2018}$ = total of values in *x* of respondents in group *g* in census year 2018.

Growth Ratio = 1 means there was no change in the total between the two censuses and *Growth Ratio* > 1.0 means the total in the 2020 NCFO was greater than that in the 2018 NCFO.

Using the growth ratios calculated by Equation 8, values missing in either the 2018 or 2020 NCFO were imputed with the following equations:

$$\hat{x}_{i,2020} = Growth\ Ratio_g(x) \times\ x_{i,2018} \tag{10}$$

$$\hat{x}_{i,2018} = \frac{1}{Growth\ Ratio_g(x)} \times x_{i,2020} \tag{11}$$

Where:

- $\hat{x}_{i,year}$ = imputed value of *x* of a ferry operator *i*.
- $x_{i,year}$ = observed value of *x* of a ferry operator *i*.
- *g* = MOS group of the operator *i*.

To apply these equations, a ferry operator must have provided data on *x* in either the 2018 or 2020 NCFO. Accordingly, ferry operators that had not responded to *x* in *both* census years were not eligible for imputation and this limited the number of included ferry operators. Once missing values were imputed, the total of imputed values in *x* of nonrespondents, $T_{x,NR}$, was calculated by summing all imputed values. This was the estimated nonresponse bias in the census total for item *x* according to Equation 7.

3.6. IDENTIFYING INFLUENTIAL VARIABLES

3.6.1. Influential Variables

Variables associated with nonresponse were identified so that subgroups based on the identified variables can be examined to find ways to increase responses in future censuses and to identify potential issues for data users in analyzing the 2020 NCFO data. Although there were many variables available in the 2020 NCFO data, not all variables were considered appropriate for analysis because values of the variables should be available for both respondents and nonrespondents. These variables whose values were known for all the units in the frame were typically “design variables” because they were used to design a sampling plan for the survey, such as stratified sampling and cluster sampling.

However, the frame data file for the 2020 NCFO contained only contact information of the ferry operators, and thus only the locational geographic information of the operators, such as state, are known for all the 246 operators in the frame. Meanwhile, there were a limited number of variables that the study was able to find values for nonrespondents using historical NCFO data. Some of these variables had values that were not likely to change over years. For example, whether an operator served the National Park Service (NPS) or not was unlikely to change over time. The missing values of these variables for nonrespondents were imputed with a high confidence using the past 2 NCFOs (i.e., 2016 and 2018 NCFOs) since they were not expected to vary over the past 3 census years (i.e., 2016, 2018, and 2020). The following six variables were identified and included in the analysis: (1) Accept Public Funding, (2) Report On Behalf of Government, (3) Operator State, (4) Serve NPS, (5) Segment Type, and (6) Population. Operator State came from the NCFO frame data file. Population was not included in NCFO but was publicly available in ACS for the calendar year of 2019 (corresponding to the 2020 NCFO); total population of terminal locations was used. These six variables served as predictors in the analysis to identify variables that were associated with nonresponse, called influential variables in this report.

3.6.2. Conditional Inference Tree

To identify variables that influenced nonresponse, a conditional inference tree method was employed. A conditional inference tree is a nonparametric class of decision trees and shares commonality with typical decision trees. The principal difference between the conditional inference tree and a typical decision tree is that the conditional inference tree model selects influential variables based on the statistical association with the response variable, while a typical tree model, such as classification and regression trees (CART), does so based on the information measure (e.g., Gini coefficient) of node impurity. In the conditional inference tree, influential variables are included only when they are statistically significant. Meanwhile, influential variables not statistically significant can still be included in CART trees. The conditional inference tree handles continuous variables better than other decision trees because statistical tests devised for continuous variables are used. Moreover, the conditional inference tree can deal with complex interactions among influential variables more effectively than other trees because statistical tests can detect interactions and adjust for them properly.

The conditional inference tree model considers interactions among all included influential variables without explicitly specifying interactions. Phipps and Toth [2012] found that the interactions among the predictors are important in analyzing establishment nonresponse. They applied a regression tree model to describe the association between establishment

characteristics and its response propensity on the Occupational Employment Statistics (OES) survey and concluded a nonresponse bias would be nonignorable without proper adjustments. Interactions among the influential variables could affect analysis results and a conditional inference tree method effectively controls for potential interactions.

The permutation test framework developed by Strasser and Weber [1999] was used to select the influential variable and also determine the best split in the conditional inference tree. An influential variable corresponding to the minimum of Bonferroni-adjusted P-values is selected for each node. When the variable was selected, the optimal binary split was determined based on a two-sample statistic measuring the discrepancy between the split samples. This two-step procedure recurses at each child node and the conditional inference tree is also known as unbiased recursive partitioning. Variables included in the final tree are statistically related to the response variable (indicator for nonresponse in this study).

4. Results and Discussion

4.1. DATA QUALITY METRICS

Four rates measuring different aspects of data quality were computed for the entire census. A unit response rate was computed also by subgroup so that subgroups showing differentiating data quality could be identified. The four calculated rates are presented in separate subsections.

4.1.1. Unit Response Rate (URR)

Table 3 shows URRs for the entire 2020 census and by subgroup. Out of the 246 operators contacted, 164 responded to at least one answer, but 160 provided acceptable data to be labeled unit responses. (Refer to Section 3.2. Determining Unit Response for the definition of a unit response). This yielded an URR of 65 percent for the 2020 NCFO (the first row of the table). Operators reporting on behalf of government or accepting public funding were more likely to respond to the census than their counterparts, which was expected. These relationships were confirmed by the chi-square test of independence at 0.05 significance level. Among the operators that received public funds, 91 percent responded to the census, while 76 percent of the operators that did not receive public funds responded.

Table 3. Unit Response Rate (URR) of 2020 NCFO by Subgroup

Subgroup	Category	Nonresponse	Response	Total	Percent URR
All	None	86	160	246	65
Report on behalf of government (*p*-value < 0.001)	No	42	77	119	65
	Yes	13	83	96	86
Accept public fund (*p*-value < 0.01)	No	25	79	104	76
	Yes	8	81	89	91
Ticket revenue (*p*-value = 0.08)	<50%	9	51	60	85
	≧50%	39	102	141	72
Number of Segments (*p*-value = 0.27)	2	26	89	115	77
	3-6	17	44	61	72
	>6	4	27	31	87
Number of Vessels (*p*-value = 0.13)	1-2	34	77	111	69
	3-6	13	55	68	81
	>6	6	28	34	82

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, 2024.
Note: Reported *p*-values from the Chi-square tests for independence.

Operators whose ticket revenue was less than 50 percent of their total revenue tended to respond to the census more often, but this relationship was not statistically significant at the 0.05 level; it was significant at the 0.1 level. Subgroup comparison with respect to the numbers of segments and vessels found no clear difference across the categories. Although operators with more than 6 segments or with 3 or more vessels showed higher URRs than the other categories, it was not statistically valid at the 0.1 significance level. The chi-square test was performed after the upper two levels (i.e., 3-6 and >6) were combined, and the results did not change the conclusions. Thus, the numbers of segments and vessels managed by the operators were not associated with response status of the operators.

4.1.1.1. Responses of Past NCFOs

Table 4 shows URRs of the past 4 NCFOs: 2014, 2016, 2018, and 2020 NCFOs. The URR increased from 46.5 percent in 2014 NCFO, to 61.9 percent in 2016 NCFO, to 75.6 percent in 2018 NCFO, and dropped to 65 percent in 2020 NCFO. For the 2020 NCFO, the data year was 2019 and the data collection year was 2020. Although the data year of 2019 was pre-pandemic, the data collection was performed during the pandemic. Thus, the URR being lower than the previous census is believed to be attributable to the pandemic. Despite all efforts to locate all ferry operators eligible to be part of the frame, new operators may have come into existence without BTS awareness. It was noteworthy that several operators were not included in the frame file but ended up participating in the census and were included in the number of the invited operators. For example, the frame file used to invite ferry operators to the 2020 NCFO included 245 operators, but two operators that were not found in the frame file eventually participated in the census and one operator in the frame was later found to be duplicate, resulting in a total of 246 invited operators for the 2020 NCFO.

Table 4. Unit Response Rate (URR) of the Past 4 NCFOs

URR	Census Year			
	2014	2016	2018	2020
Number of responding operators	120	161	180	160
Number of operators invited to census	258	260	238	246
Percent URR	46.5	61.9	75.6	65.0

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset, 2018 NCFO Dataset, 2016 NCFO Dataset, and 2014 NCFO Dataset,* available at https://www.bts.gov/NCFO as of April 2024.

Table 5 shows participation of ferry operators over the past four NCFOs based on the operator ID variable. Participation means that a ferry operator submitted information to the NCFO but the submitted information may not necessarily satisfy the definition of a unit response. Thus, the number of responses is equal to or less than the number of participations. A total of 227 operators were found to have participated in at least one of the four censuses; about 10 to 30 operators have never participated in the four censuses. A total of 86 ferry operators (38 percent) participated in all four NCFOs, 53 in three NCFOs, 45 in two NCFOs and 43 in one NCFO. It is possible that operators having not participated in a census were in fact out of business in a corresponding data year meaning they were ineligible to be part of the NCFO. However, it was speculated that such operators, if existed, would be rare and thus would not affect potential conclusions drawn from Table 5.

Table 5. Participation of Operators over the Past 4 NCFOs Based on Operator ID

NCFO						
2014	2016	2018	2020	Number of Censuses	Number of Operators	
O	O	O	O	4	86	86
X	O	O	O	3	53	36
O	X	O	O			8
O	O	X	O			2
O	O	O	X			7
X	X	O	O	2	45	16
X	O	X	O			3
X	O	O	X			14
O	O	X	X			10
O	X	O	X			2
O	X	X	O			0
X	X	X	O	1	43	13
X	X	O	X			12
X	O	X	X			5
O	X	X	X			13

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset, 2018 NCFO Dataset, 2016 NCFO Dataset, and 2014 NCFO Dataset,* available at https://www.bts.gov/NCFO as of April 2024.

4.1.2. Item Response Rate (IRR) and Total Item Response Rate (TIRR)

IRR and TIRR were calculated for each of the 19 key response variables using Equation 2 and Equation 3. Calculation of IRR was done for each table. This implies that $IREQ_x$, the denominator of IRR (i.e., the number of responses required for item x), changes depending on the table to which the item x belongs. For example, *IREQ* for Accept Public Funding is 164, the number of the ferry operators having participated in the 2020 census, since the item is in Operator table. Meanwhile, *IREQ* for Segment Name item is the number of segments that the 164 operators *should* have reported on the 2020 census. Since the true number of segments was unknown, due to the fact that some operators have not reported all their segments, 934 (the number of segments included in the released data) was used for *IREQ* for Segment Name.

It is noteworthy that 934 segments included not only segments reported by the 164 operators but also segments BTS added during the data edit process. Through the data edit process (e.g., automated data edit and analyst's manual review), BTS made efforts to verify the reported segments using operator websites and, if segments were found missing, BTS added missing segments if possible. These efforts led to adding several segments that the operators failed to report in the 2020 census. For example, when an operator reported two terminals and reported only one segment, BTS added the pairing segment for a returning trip since a ferry boat most likely operated two ways between the terminals[18]. However, not all segments that the operators failed to report were discovered and added by BTS's data edit process.

Table 6 presents IRRs and TIRRs of the 19 key variables and TIRRs were calculated by multiplying IRRs by the URR, 65 percent, following Equation 3; IRRs for an extended list of items (47 items) are presented in Appendix A. For IRR calculation, $IREQ_x$ was set to equal to the number of rows in a corresponding table, meaning it was assumed that there was no valid

[18] One segment consisted of a ferry boat running from Terminal A to Terminal B and the pairing segment consisted of a ferry boat running from Terminal B to Terminal A. When there are more than two terminals, there might be no pairing segments due to a possible one-way looping route (e.g., a route where Terminal A→Terminal B→Terminal C →Terminal A).

skip. Although this assumption was not necessarily true for some variables, identifying a valid skip in a variable often required information stored in a table different from that of the variable of interest, which involved manual review of all relevant variables across different tables. For this reason, TIRR was calculated based on the assumption of no valid skip.

Table 6. Item Response Rate (IRR) and Total Item Response Rate (TIRR) of 19 Key Response Variables in 2020 NCFO

19 Key Response Variable (x)	Table	Percent IRR_x	Percent $TIRR_x$
Accept public funding	Operator	100	65
Operator city		100	65
Operator state		100	65
Operator name		100	65
Average trip time	Operator-segment	81	53
Most used vessel		81	53
Passenger boarding		81	53
Segment length		81	53
Season end date		79	51
Season start date		79	51
Vehicle boarding		81	53
Segment name	Segment	81	53
Terminal 1 city	Terminal	99	64
Terminal 2 city		89	58
Terminal 1 state		99	64
Terminal 2 state		89	58
Terminal 1 name		99	64
Terminal 2 name		89	58
Vessel name (of most used vessel)	Vessel	81	53

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

Items in Operator table show the highest TIRR, while those associated with segment questions in the Operator-Segment and Segment tables show the lowest TIRR. This is anticipated since the segment questions required detailed data that may not be readily available to the respondent as the respondent may have needed to retrieve data from different sources/systems or requested assistance from other departments. Continual efforts are being made to improve the data quality of a future censuses by making the census questionnaire and instrument more intuitive, easier to respond, and more forgiving to correct mistakes, especially for the questionnaire sections regarding segments.

4.1.3. Modified Quantity Response Rate (MQRR)

MQRR could be calculated for each variable where imputation is possible using Equation 4 and was calculated for two boarding count variables, Passenger Boarding and Vehicle Boarding. These two boarding variables are important since it is BTS's mission to collect and publish statistics on intermodal and multimodal passenger movement,[19] and the ferry boarding counts are such statistics. Moreover, the boarding counts determine the amount of federal fund distributed to a ferry operator according to Fixing America's Surface Transportation (FAST) Act

[19] Title 49 U.S.C. § 6302(b)(3)(B)(vi)(X).

of 2015[20]. Calculating MQRR requires imputation and missing values in the two boarding counts were imputed. MQRR was calculated for passenger boarding and vehicle boarding, separately.

4.1.3.1. MQRR for Passenger Boarding

MQRR for Passenger Boarding in the 2020 NCFO is calculated as follows:

$$\begin{aligned} MQRR_{Passengers} &= \left(\frac{Total\ Passengers_{\ Obs}}{Total\ Passengers_{Est}}\right) \times 100\% \\ &= \left(\frac{113{,}990{,}893}{153{,}852{,}554}\right) \times 100\% \\ &\approx 74\% \end{aligned} \tag{12}$$

In comparison, TIRR = 41 percent (Table 6) for Passenger Boarding. This means although the proportion of the ferry operators having provided passenger boarding data in the 2020 NCFO was 41 percent, the proportion of the total of the values they provided to the total that would be if all the operators provided passenger boarding data is considerably higher, 74 percent.

There are some caveats involved in MQRR calculation. First, the estimated passenger total, the denominator, should have included all 246 ferry operators invited to the 2020 NCFO. However, some of the operators did not participate in not only the 2020 NCFO but also past NCFOs, implying no information was available for imputing their boarding counts, thus imputation was impossible. Second, the imputation was performed based on growth ratio in passenger boarding counts over two consecutive censuses, 2018 and 2020 NCFOs. This means ferry operators that provided valid data in passenger boarding in at least one of the two NCFOs were included in the MQRR calculation, and this limits the number of included operators to 177. These imply the calculated MQRR is not comparable to the TIRR since the TIRR reflect all 246 ferry operators.

4.1.3.2. MQRR for Vehicle Boarding

MQRR for Vehicle Boarding in the 2020 NCFO was calculated as follows:

$$\begin{aligned} MQRR_{Vehicles} &= \left(\frac{Total\ Vehicles_{\ Obs}}{Total\ Vehicles_{Est}}\right) \times 100\% \\ &= \left(\frac{26{,}607{,}017}{28{,}621{,}575}\right) \times 100\% \\ &\approx 93\% \end{aligned} \tag{13}$$

MQRR is much higher than TIRR = 41 percent (Table 6); as noted earlier, TIRR is identical among the two boarding count variables (Passenger Boarding and Vehicle Boarding). This means although 41 percent of the ferry operators provided vehicle boarding data in the 2020 NCFO, the proportion of the total of the values they provided to the total that would be if all the operators provided valid vehicle boarding data is considerably higher, 93 percent. The caveats involved with calculating the MQRR for passenger boarding discussed in the previous section also apply here.

---

[20] Pub. L. No. 114-94, § 1112, 129 Stat. 1312 (2015).

4.2. NONRESPONSE BIAS

Nonresponse bias in two variables is of interest in this study: Passenger Boarding and Vehicle Boarding. To estimate nonresponse bias, imputing missing values in the two variables was necessary, and imputation at operator level was performed using the measure of size group-based growth ratios.

4.2.1. Measure of Size Group

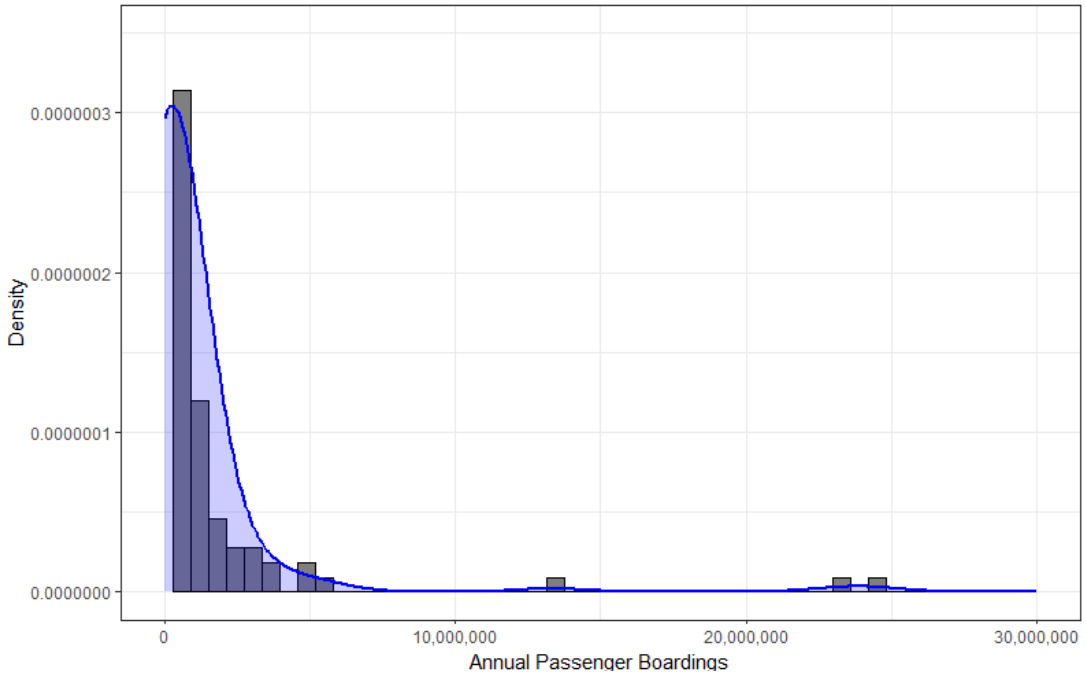
MOS of a ferry operator for imputation is the corresponding historical boarding count: passenger boarding for imputing missing passenger count and vehicle boarding for imputing missing vehicle count. MOS groups were determined considering natural breaks in distribution, clean values of break points, and proportion of each group in the total.

4.2.1.1. Passenger Boardings

Distribution of passenger boardings was examined and Figure 4 shows distribution of nonzero passenger boarding counts of 2018 and 2020 NCFO. The distribution[21] is right skewed, hinting at the presence of a few notably large values. This kind of a right-skewed distribution is a unique characteristics of business/establishment surveys discussed by Lineback and Thompson [2010] (Figure 3).

[21] The histogram was created with 50 bins and the density plot was created with the kernel density bandwidth determined by the unbiased cross validation method.

Figure 4. Distribution of Nonzero Passenger Boardings in the 2018 and 2020 NCFO



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: The histogram was created using 50 bins and the density plot is created with the bandwidth determined by the unbiased cross validation method.

To determine the MOS groups, passenger boarding in the 2018 and 2020 NCFO were analyzed. A total of 177 ferry operators that had a nonzero passenger boarding count in either or both of the census years 2018 and 2020 were included in the analysis. When passenger boarding in one year was available, that was used to represent the operator's MOS. When passenger boarding in both years were available, the average of the two counts was used. Among the 177 operators, 105 have valid passenger boarding counts in both census years.

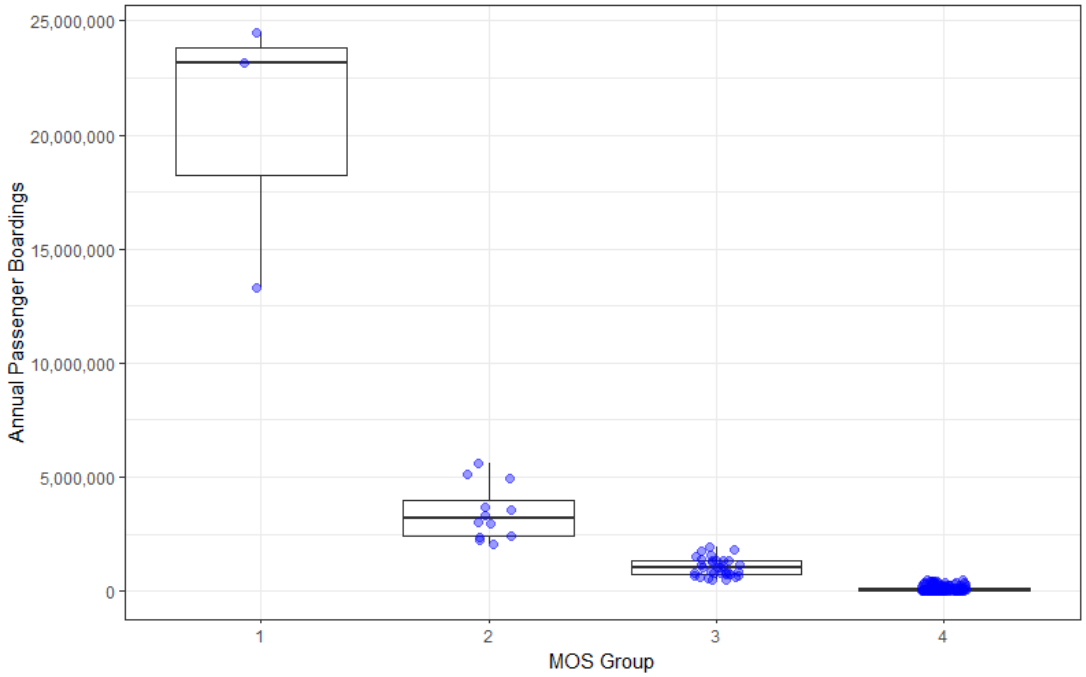
Determining the cutoffs in passenger boardings for grouping operators was done in an iterative manner. As seen in Figure 4 and Figure 5, there are three operators found on the right tail and they are distant from the remaining operators. These three operators comprised 40 percent of the total boarding count of all 177 operators and were formed to be one group (Figure 5). Another group that consisted of operators with small passenger boardings was formed to comprise 10 percent of the total boardings and included 129 operators. Two more groups were formed to comprise similar percentages, 27and 23 percent. As noted earlier, grouping was determined considering natural breaks in distribution, clean values of break points, and proportion of each group in the total. As a result, four groups were formed and are summarized in Table 7 and visualized in Figure 5.

Table 7. Measure of Size Group Based on Passenger Boardings

Group	Description	Definition	Number of Operators[a]	Total Boarding Counts[b]	Percent of All Groups
1	Extra Large (XL)	>= 10,000,000	3	60,942,777	40
2	Large (L)	10,000,000 > & >=2,000,000	12	41,403,641	27
3	Medium (M)	2,000,000 > & >= 500,000	33	34,583,713	23
4	Small (S)	500,000 > & > 0	129	14,845,905	10

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
[a] 177 ferry operators were included in determining MOS groups and they had nonzero values in the passenger boardings in either or both of the census years, 2018 and 2020.
[b] Based on the 2018 and 2020 NCFO data. When values were available in both years, an average was used.

Figure 5. Boxplots of Annual Passenger Boardings by MOS Group

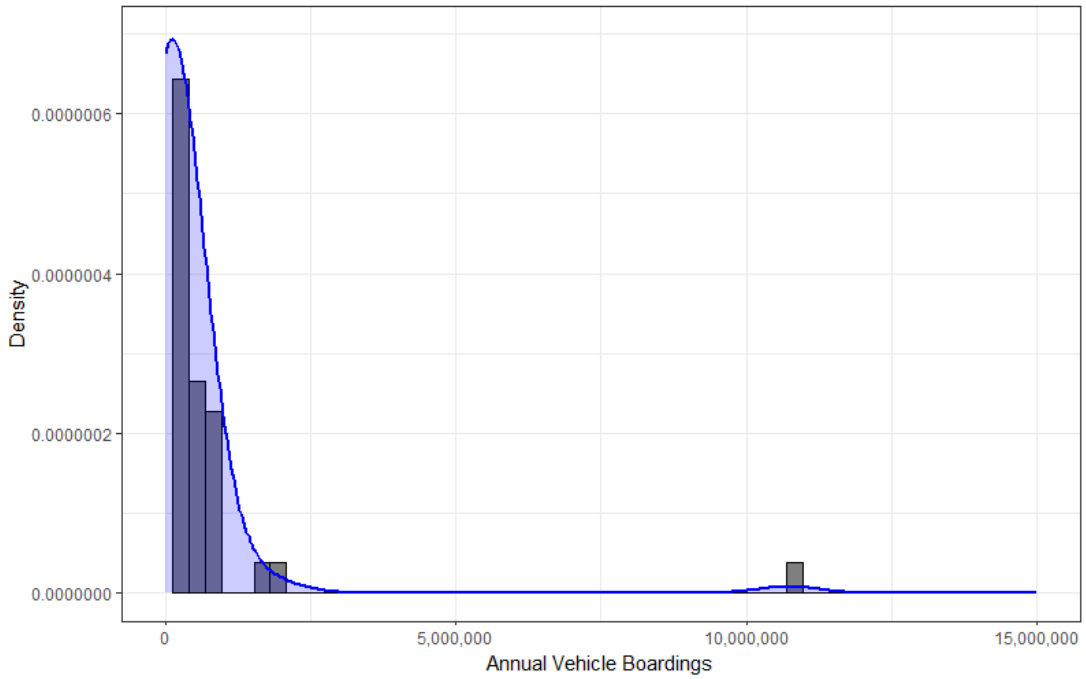


Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: The boxplot was based on 177 ferry operators with nonzero passenger boardings in either or both of the census years, 2018 and 2020.

4.2.1.2. Vehicle Boardings

Figure 6 shows the distribution of nonzero vehicle boarding counts of the 2018 and 2020 NCFO. It is right skewed due to the presence of a few notably large values.

Figure 6. Distribution of Nonzero Vehicle Boarding Counts in the 2018 and 2020 NCFO



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: The histogram was created with 55 bins and the density plot was created with the bandwidth determined by the unbiased cross validation method.

To determine the MOS groups, vehicle boardings in the 2018 and 2020 NCFO were analyzed. A total of 95 ferry operators were included in the analysis, and they had nonzero vehicle boarding count in either of the two census years, 2018 and 2020. When vehicle boarding in both years were available, average of the two counts was used. Among the 95 operators, 65 have valid vehicle boarding in both census years. The number of operators included in the analysis is notably smaller than that for passenger boarding analysis because some operators and associated vessels did not carry vehicles.

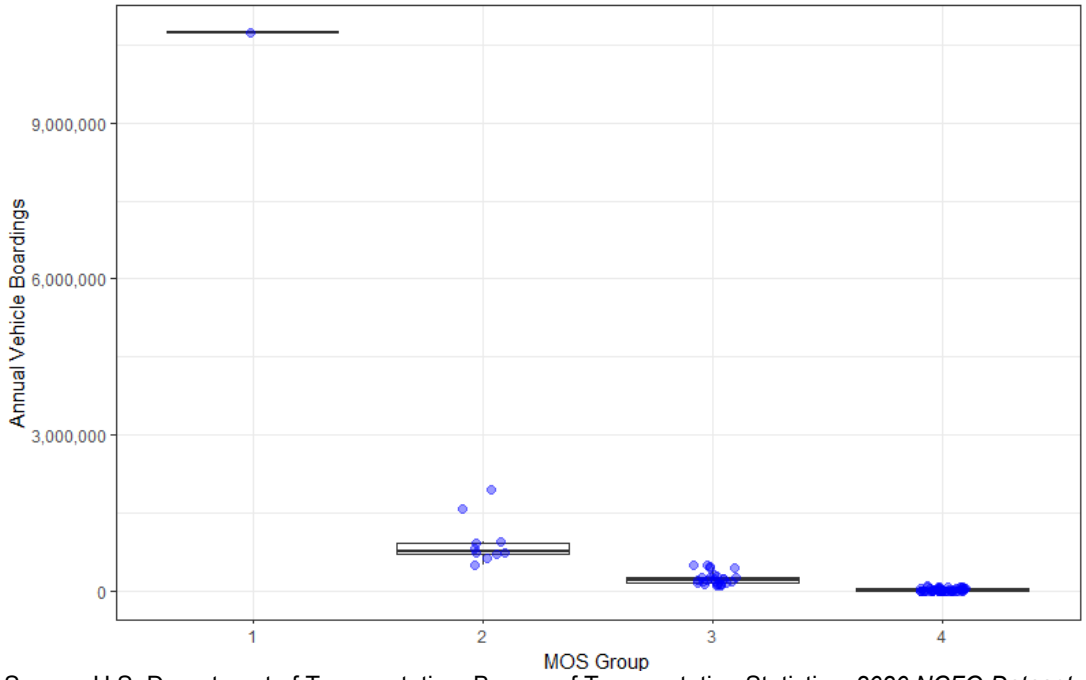
Determining the cutoffs in vehicle boardings for grouping operators was done in an iterative manner. As seen in Figure 6 and Figure 7, there is one operator far away from the rest of the operators and the operator alone comprised 41 percent of the total vehicle boarding of the 95 ferry operators. That operator alone formed a standalone group. A group of operators with small vehicle boardings was formed to comprise 5 percent of the total boarding and includes 57 operators. Two middle groups were formed to comprise 31 and 23 percent. The resulting four groups are provided in Table 8 and visualized in Figure 7.

Table 8. MOS Group Based on Vehicle Boardings

Group	Description	Definition	Number of Operators[a]	Total Boarding Count[b]	Percent of All Groups
1	Extra large (XL)	>= 10,000,000	1	10,805,029	41
2	Large (L)	10,000,000 > & >=500,000	10	8,365,547	31
3	Medium (M)	500,000 > & >= 100,000	27	6,132,876	23
4	Small (S)	100,000 > & > 0	57	1,303,565	5

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
[a] 95 ferry operators were included in determining MOS groups that had nonzero values in the vehicle boardings in either or both of the census years, 2018 and 2020.
[b] Based on the 2018 and 2020 NCFO data. When values were available for both years, an average was used.

Figure 7. Boxplots of Annual Vehicle Boardings by MOS Group



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: The boxplot was based on 95 ferry operators with nonzero vehicle boardings in either or both of the census years, 2018 and 2020.

4.2.2. Growth Rate and Ratio

Growth rates and ratios were calculated by for each MOS group using Equation 8 and Equation 9. For these calculations, values in both the 2018 and 2020 NCFOs were needed.

4.2.2.1. Passenger Boarding

While examining the 2018 and 2020 NCFO data for calculating the growth, several operators were noted for abnormal changes in passenger boardings (e.g., growth ratios being too small or too large) leading to extensive review of their data in passenger and vehicle boarding in the 2014 through 2020 NCFO. The study determined that six operators suffered from entry errors in passenger boarding. For example, one operator's passenger boarding count in 2018 NCFO was 10 times that of the 2020 NCFO, while the passenger boarding in the 2014, 2016, and 2020

NCFOs were similar for that operator, and the vehicle boarding in the 2018 NCFO was similar to that the 2020 NCFO. This led to conclusion that the passenger boarding in 2018 NCFO suffered from entry error and needed to be corrected. Some operators had abnormal counts in their passenger boarding but could not be determined to be erroneous. The passenger boarding counts of the six operators were corrected using historical passenger boardings and other relevant data of these operators[22].

The 2018–2020 growth rates and ratios in passenger boarding were calculated, excluding the six operators[23]. Table 9 presents calculated rates and ratios by MOS group along with the number of operators included in the calculation. A total of 105 operators that reported nonzero passenger boardings in both the 2018 and 2020 NCFO were included in the calculation. The overall growth rate and ratio of these operators indicates a 1.91 percent increase from the 2018 to 2020 census year. However, they vary across the four MOS groups, ranging from a 1.95 percent decrease in Group 2 (i.e., large operators) to a 10.5 percent increase in Group 3 (i.e., medium operators).

Table 9. Growth Rates and Ratios of Passenger Boardings in 2018 and 2020 NCFO

Group	Number of Operators	Percent Growth Rate	Growth Ratio
1 (Extra-Large Operators)	2	-0.83	0.992
2 (Large Operators)	7	-1.95	0.981
3 (Medium Operators)	23	10.5	1.105
4 (Small Operators)	73	5.33	1.053
All	105	1.91	1.019

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

As noted in Section 3.5.3. Imputation, imputing missing boardings based on the growth ratios is based on the strong assumption that a nonresponding ferry operator experiences the average growth of the responding operators in the same MOS group. However, considering the imputation was to provide a rough estimate of a bias in total boardings when only observed boarding counts were used to calculate the total boarding, applying the growth ratios for imputation is deemed acceptable. Imputation of boarding counts based on a more sophisticated approach is currently being conducted and would provide a more reliable estimate of the total boarding. The growth ratio was equivalent to the slope of the linear model describing passenger boarding. The visualized regression lines by group were presented in Appendix B.

4.2.2.2. Vehicle Boarding

While examining the 2018 and 2020 NCFO data, several operators were noted in their abnormal changes in vehicle boarding. After extensive review of their data in boarding counts in the 2014 through 2020 NCFOs, the study determined that four operators suffered from entry errors in vehicle boarding. The study corrected the erroneous vehicle boarding counts of the four operators using historical boarding counts and other relevant data of from these operators.

Growth rates from the 2018-2020 vehicle boardings were calculated excluding these four operators. Table 10 presents calculated rates and ratios by MOS group along with the number

[22] It should be noted that these corrections were made only for this study, not in the released data sets.
[23] Including the six operators results in similar the rates and ratios.

of operators included in the calculation. A total of 65 operators[24] that reported nonzero vehicle boardings in both the 2018 and 2020 NCFOs were included in the calculation. The overall growth rate and ratio of these operators indicate a 0.3 percent decrease from the 2018 to 2020 census years. However, the rate and ratio vary across the four groups, ranging from an 8.5 percent decrease in Group 4 (i.e., small operators) to a 1.54 percent increase in Group 1 (i.e., extra-large operator).

Table 10. Growth Rates and Ratios of Vehicle Boardings in 2018 and 2020 NCFO

Group	Number of Operators	Percent Growth Rate	Growth Ratio
1 (Extra-Large Operators)	1	1.54	1.015
2 (Large Operators)	7	-1.58	0.984
3 (Medium Operators)	20	-0.30	0.997
4 (Small Operators)	37	-8.50	0.915
All	65	-0.30	0.997

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

The growth ratios in Table 10 were used to impute missing passenger boardings in the 2018 or and 2020 NCFOs. The growth ratio was equivalent to the slope of the linear regression describing vehicle boarding. The visualized regression lines by group are presented in Appendix B.

4.2.3. Nonresponse Bias Estimate

Nonresponse biases were estimated on passenger and vehicle boarding counts in the 2018 and 2020 NCFOs and are presented separately.

4.2.3.1. Passenger Boarding

Biases were calculated for overall and by MOS group in the 2018 and 2020 NCFOs using Equation 7 and are presented in Table 11. Based on the 177 operators included in the bias estimation, the total bias due to nonresponse in the 2020 NCFO was estimated to be about 40 million passengers undercounted, 35 percent of the observed total. About 13 million out of the total bias came from Group 1 and about 11 million came from Group 2. The number of observations (second column of the table) indicates the number of the operators having provided passenger boardings in both the 2018 and 2020 NCFOs and those having provided in one of the two censuses.

[24] The number of the ferry operators having reported nonzero vehicle boardings (i.e., 65) is smaller than that of the operators having reported nonzero passenger boardings (i.e., 105) because many of the 105 operators have carried only passengers.

Table 11. Estimated Bias in Passenger Boardings by MOS Group in 2018 and 2020 NCFOs

Group	Number of Operators	Passenger Boardings	Type	2018 NCFO	2020 NCFO
All	177	Count	Observed[a]	143,428,652	113,990,893
			Estimated[b]	151,036,807	153,852,554
		Bias	Number[c]	-7,608,155	-39,861,661
			Percentage[d]	-5%	-35%
Group 1	3	Count	Observed[a]	61,141,307	47,418,283
			Estimated[b]	61,141,307	60,633,641
		Bias	Number[c]	0	-13,215,358
			Percentage[d]	0%	-28%
Group 2	12	Count	Observed[a]	35,806,139	30,077,815
			Estimated[b]	41,760,128	40,946,737
		Bias	Number[c]	-5,953,989	-10,868,922
			Percentage[d]	-17%	-36%
Group 3	33	Count	Observed[a]	32,539,630	26,070,244
			Estimated[b]	33,304,055	36,803,843
		Bias	Number[c]	-764,425	-10,733,599
			Percentage[d]	-2%	-41%
Group 4	129	Count	Observed[a]	13,941,576	10,424,551
			Estimated[b]	14,831,317	15,468,333
		Bias	Number[c]	-889,741	-5,043,782
			Percentage[d]	-6%	-48%

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

[a] Observed total count was the sum of the reported passenger boarding counts: $T_{x,R}$ in Equation 7.

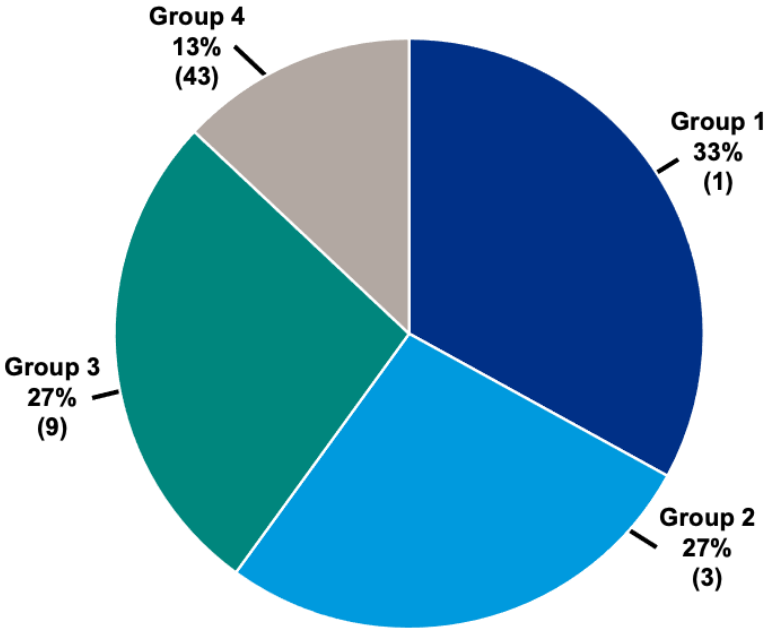
[b] Estimated total count was the sum of the reported and imputed passenger boarding counts: $T_{x,C}$ in Equation 7.

[c] Bias was calculated using Equation 7: $T_{x,R} - T_{x,C}$.

[d] *Percentage* = (*Bias* ÷ *Observed Total Count*) × 100%.

Figure 8 shows the percentage contribution of each group to the total bias in the 2020 NCFO, displayed in percentage, and the number in the parenthesis is the number of nonresponding operators. The largest percentage of the bias came from Group 1, followed by Group 2, then Group 3 is slightly behind Group 2. Of the total bias, 33 percent is attributable to one extra-large ferry operator (Group 1) having not responded to the passenger boarding item in the NCFO 2020 and 27 percent is attributable to three large nonresponding operators (Group 2). The top two groups, Groups 1 and 2, accounted for 60 percent of the bias. Four operators missing their passenger boarding counts in the 2020 NCFO are responsible for 60 percent of the total bias showing the importance of obtaining passenger boarding data from big operators in estimating an accurate national passenger boarding count.

Figure 8. Contribution to the Total Bias in Passenger Boarding in the 2020 NCFO by Group



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: The number in parenthesis is the number of ferry operators with missing passenger boarding.

4.2.3.2. Bias Estimate for Entire Ferry Passenger Population

As noted earlier, the total bias should be estimated ideally for all 246 ferry operators invited to the 2020 NCFO. However, imputing missing passenger boarding based on the 2018–2020 growth ratio limited the number of operators to 177 that reported nonzero passenger boardings in at least one of the two census years. The biases shown in Table 11 are underestimated, and the degree of the underestimation cannot confidently be assessed without more information about nonresponding operators.

With a set of assumptions, however, the potential range of underestimation could be roughly estimated. A total of 69 operators[25] were excluded from the bias estimation in Table 11. The degree of underestimation depends on which MOS groups these operators belong, which is unknown. The first assumption was made on the group membership of the 69 operators. Two likely scenarios were devised corresponding to the upper and lower ends of the probable range of the underestimation. The upper end of underestimation assumes that the 69 operators belong to one of the three groups excluding Group 1 (extra-large operators)[26] and further assumes that

[25] 246 (number of operators invited to 2020 NCFO)–177 (number of operators included in the bias estimation analysis) = 69. These 70 operators had not provided data on passenger boardings in both 2018 and 2020 NCFOs including those having not participated in both censuses.

[26] Examining 2014, 2016, 2018, and 2020 NCFO data identified, there were only three ferry operators having reported passenger boardings greater than 10 million in any of the four censuses. Thus, it was unlikely that any of the 69 nonresponding operators could have carried at least 10 million passengers, suggesting all the 69 operators fell in one of the three smaller groups, Groups 2, 3, and 4.

the distribution across the three groups is the same as in Figure 8 in terms of the number of operators: 5.5 percent for Group 2, 16.4 percent for Group 3 and 78.2 percent for Group 4.[27] The lower end of underestimation assumes that all the 69 operators fall into Group 4, small operators. Although it was not likely that all the 69 operators belong to Group 4, it was likely that a great majority of them are small (i.e., Group 4).

Another assumption made was that the missing passenger boarding of each of the 69 operators was equal to the average of the imputed passenger boarding for their corresponding group. For example, an operator belonging to Group 2 was assumed to have its missing passenger boarding equal the average of the imputed passenger boardings of Group 2; the average for Group 2 is 3,622,974 and was calculated as follows:

$$\frac{Total\ Bias\ in\ Group\ 2}{Number\ of\ Nonresponding\ Operators\ in\ Group\ 2} = \frac{10{,}868{,}922}{3} = 3{,}622{,}974 \quad (14)$$

Where:

- *Total Bias in Group 2* is found in Table 11.
- *Number of Nonresponding Operators in Group 2* is found in Figure 8 (the number in the parenthesis).

As noted in Equation 7, the bias is essentially the total of imputed values.

With the above two assumptions, the lower and upper ends of the probable range of the underestimation could be calculated as follows:

Lower End:

$$(69 \times 1.00) \times \frac{5{,}043{,}782}{43} = 8{,}093{,}511 \quad (15)$$

Upper End:

$$(69 \times 0.055) \times \frac{10{,}868{,}922}{3} + (69 \times 0.164) \times \frac{10{,}733{,}599}{9} + (69 \times 0.782) \times \frac{5{,}043{,}782}{43} = 33{,}574{,}023 \quad (16)$$

Thus, the amount of the underestimation (i.e., the bias) could range from about 8.1 million to about 34 million, and the amount of the estimated total bias in passenger boardings for the entire ferry population during calendar year 2019 (i.e., 2020 NCFO) ranges from 48 million[28] (42 percent of the observed total) to 73 million[29] (64 percent of the observed total). Increasing the unit response rate and item response rate on passenger boarding will be critical in estimating accurate national total passenger boarding count. Based on the estimated range of the potential total bias, the national total passenger boarding count would have ranged between

---

[27] The number of nonresponding operators in Figure 8 = 3:9:43, and the corresponding percentages = 5.5%:16.4%:78.2%.

[28] 39,861,661 (estimated bias for 177 operators responding to at least one of the two censuses) + 8,093,511 (lower end of estimated bias for 69 operators nonresponding to both censuses) = 47,955,172

[29] 39,861,661 (estimated bias for 177 operators responding to at least one of the two censuses) + 33,574,023 (upper end of estimated bias for 69 operators nonresponding to both censuses) = 73,435,684

162 million and 187 million passengers in 2019. Table 12 shows the estimates for the ferry passenger population in 2019.

Table 12. Estimated Total Passenger Boarding Count and Bias for Entire Ferry Population in 2019

2019 Ferry Passengers	Observed Total Count	Estimated Total Count	Estimated Total Bias
Lower End	113,990,893	161,946,065	47,955,172
Upper End		187,426,577	73,435,684

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

4.2.3.3. Vehicle Boarding

Biases were calculated overall and by MOS group in the 2018 and 2020 NCFOs using Equation 7 and are presented in Table 13 and Figure 9. Based on the 95 operators included in the bias estimation, the total bias due to nonresponse in the 2020 NCFO was estimated to be about 2 million vehicles undercounted (8 percent of the observed total). The 95 operators included is much smaller than the 177 operators included to estimate the bias of passenger boarding because many of the operators carried only passengers.

Table 13. Estimated Bias in Vehicle Boarding by MOS Group in 2018 and 2020 NCFOs

Group	Number of Operators	Vehicle Boardings	Type	2018 NCFO	2020 NCFO
All	95	Count	Observed[a]	26,416,295	26,607,017
			Estimated[b]	28,720,196	28,621,575
		Bias	Number[c]	-2,303,901	-2,014,558
			Percentage[d]	-9%	-8%
Group1	1	Count	Observed[a]	10,641,210	10,805,029
			Estimated[b]	10,641,210	10,805,029
		Bias	Number[c]	0	0
			Percentage[d]	0%	0%
Group2	10	Count	Observed[a]	7,679,712	8,365,547
			Estimated[b]	9,658,734	9,506,447
		Bias	Number[c]	-1,979,022	-1,140,900
			Percentage[d]	-26%	-14%
Group3	27	Count	Observed[a]	6,351,877	6,132,876
			Estimated[b]	6,627,465	6,609,519
		Bias	Number[c]	-275,588	-476,643
			Percentage[d]	-4%	-8%
Group4	57	Count	Observed[a]	1,743,496	1,303,565
			Estimated[b]	1,792,787	1,700,579
		Bias	Number[c]	-49,291	-397,014
			Percentage[d]	-3%	-30%

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

[a] Observed total count was the sum of the reported vehicle boarding counts: $T_{x,R}$ in Equation 7.

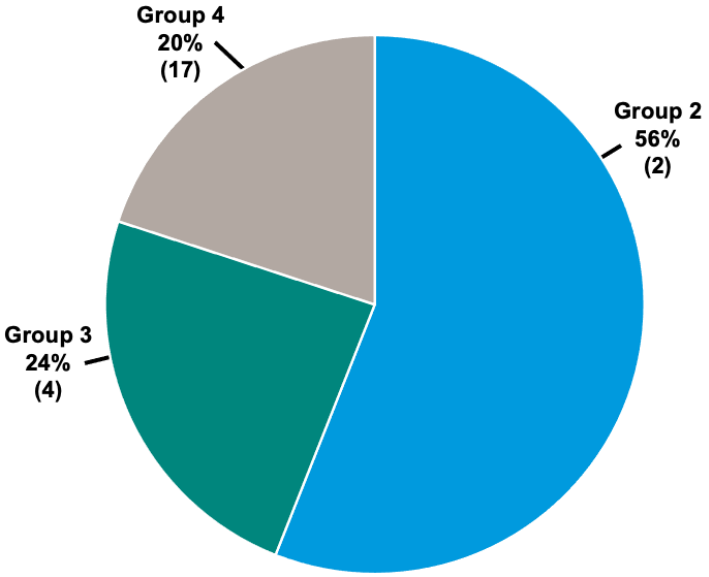
[b] Estimated total count was the sum of the reported and imputed vehicle boarding counts: $T_{x,C}$ in Equation 7.

[c] Bias is calculated using Equation 7: $T_{x,R} - T_{x,C}$.

Figure 9 shows a contribution of each group to the total bias in the 2020 NCFO, and the number in the parenthesis is the number of nonresponding operators. Group 1 is not shown in the figure because there was no nonresponding operator in the group, as only one operator fits into Group 1. The largest portion of the bias came from Group 2, followed by Group 3 and Group 4 closely behind Group 3. 56 percent of the total bias is attributable to the two large ferry operators (Group 2) having not responded to vehicle boarding item in NCFO 2020. This means

it is of critical importance to obtain vehicle boarding data from operators in Groups 1 and 2 to accurately estimate the national total vehicle boarding count.

Figure 9. Contribution to the Total Bias in Vehicle Boarding in the 2020 NCFO by Group



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: The number in parenthesis is the number of ferry operators with missing vehicle boarding.

4.2.3.4. Bias Estimate for Entire Ferry Vehicle Population

The estimated bias in Table 13 was limited to 95 operators having nonzero vehicle boarding in at least one of the two census years. Although it was a challenging task to estimate the bias for the entire ferry population with a high confidence, a rough estimate can be calculated with assumptions that are similar to those used for passenger boarding in the previous section. Assumptions on group membership and likely vehicle boarding count similar to those used in the passenger count were made when estimating the bias in total vehicle boarding count.

A total of 151 operators[30] were excluded in the bias estimation, and a portion of these operators carried only passengers and no vehicles. An assumption was needed on the proportion of the 151 nonresponding operators that have carried vehicles. Based on the ferry operators included in the bias estimation for passenger and vehicle boarding, 54 percent[31] of these operators carried vehicles meaning 46 percent of them carried only passengers. Thus, it was assumed that 54 percent of the 151 nonresponding operators carried vehicles. Accordingly, 82 operators were assumed to have carried vehicles and were used to estimate the of bias in vehicles carried

[30] 246 (number of operators invited to 2020 NCFO)–95 (number of operators included in the bias estimation analysis) = 151. These 151 operators have not provided data on vehicle boarding in both the 2018 and 2020 NCFOs including those having not participated in both censuses and those having not carried vehicles.

[31] In 2018 or 2020 NCFO, 177 operators carried passengers while 95 operators carried vehicles resulting in 54% (i.e., 100 × (95 ÷ 177)).

for the entire ferry population while the other 69 operators were assumed to have not carried vehicles and thus were excluded from the estimation.

As for the group membership of the 82 operators, two likely scenarios were devised corresponding to the upper and lower ends of the probable range were calculated. The upper end scenario assumes that the 82 operators fall into one of the three groups (Groups 2, 3, and 4) following the ratio in Figure 9: 8.7 percent for Group 2, 17.4 percent for Group 3 and 73.9 percent for Group 4.[32] The lower end scenario assumes that all the 82 operators fall into Group 4, small operators.

The vehicle boarding count of each nonresponding operator was assumed to be equal to the average of the imputed passenger boarding for their corresponding group. For example, an operator belonging to Group 2 is assumed to have its missing vehicle boarding count equal the average of the imputed vehicle boarding counts of Group 2, which is calculated as follows:

$$\frac{Total\ Bias\ in\ Group\ 2}{Number\ of\ Nonresponding\ Operators\ in\ Group\ 2} = \frac{1{,}140{,}900}{2} = 570{,}450 \quad (17)$$

Where:

- *Total Bias in Group* 2 is found in Table 13.
- *Number of Nonresponding Operators in Group* 2 is found in Figure 9 (the number in the parenthesis).

As noted in Equation 7, the bias is the total of imputed values.

With the above three assumptions, the lower and upper ends of the probable range of the underestimation can be calculated as follows:

Lower End:

$$(82 \times 1.00) \times \frac{397{,}014}{17} = 1{,}915{,}009 \quad (18)$$

Upper End:

$$(82 \times 0.087) \times \frac{1{,}140{,}900}{2} + (82 \times 0.174) \times \frac{476{,}643}{4} + (82 \times 0.739) \times \frac{397{,}014}{17} = 7{,}182{,}336 \quad (19)$$

[32] The number of nonresponding operators in Figure 6 = 2:4:17 and the corresponding percentages = 8.7%:17.4%:73.9%.

Thus, the amount of the underestimation ranges from about 1.9 million to about 7.2 million vehicles, and the amount of the estimated total bias in vehicle boardings for the entire ferry population in calendar year 2019 ranges from 3.9 million[33] (15 percent of the observed total) to 9.2 million[34] (35 percent of the observed total). Based on the estimated range of the potential total bias, the national total vehicle count would be between 31 million and 36 million vehicles in 2019. Table 14 shows the estimates for the ferry vehicle population in 2019.

Table 14. Estimated Total Vehicle Boarding Count and Bias for Entire Ferry Population in 2019

2019 Ferry Vehicles	Observed Total Count	Estimated Total Count	Estimated Total Bias
Lower end	26,607,017	30,536,584	3,929,567
Upper end		35,803,911	9,196,894

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

4.3. INFLUENTIAL VARIABLES

The operator-segment-terminal-vessel level data with 1,005 rows[35] was used in conditional inference tree analysis. Four analyses were performed: (1) unit response, (2) response to passenger boardings, (3) response to vehicle boardings, and (4) response to segment length. Table 15 shows the numbers of response and nonresponse cases at the operator-segment-terminal-vessel level. Of the 1,005 cases, 76 cases correspond to unit nonresponse[36]. In the boarding counts and segment length variables, 75 cases were missing values, and they were the same cases, meaning all three variables are missing in those 75 cases. It should be noted that the same operator may be found in multiple rows when that ferry operator reported multiple segments.

Table 15. Number of Nonresponses in Operator-Segment-Terminal-Vessel Data

Variables	Response	Nonresponse	Number of Cases
Operator name, vessel name, route origin, and route destination	929	76	1,005
Passenger boardings	897[a]	108	
Vehicle boardings	897[b]	108	
Segment length	897[c]	108	

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

[a] Includes 33 cases with zero passenger boardings.

[b] Includes 478 cases with zero vehicle boardings. Some of these cases were legitimate and no vehicles were carried on the corresponding segments (e.g., all the vessels serving the segments carried only passengers).

[c] There was no case where the segment length was zero.

---

33 2,014,558 (estimated bias for 95 operators responding to at least one of the two censuses) + 1,915,009 (lower end of estimated bias for 82 operators nonresponding to both censuses) = 3,929,567.

34 2,014,558 (estimated bias for 95 operators responding to at least one of the two censuses) + 7,182,336 (upper end of estimated bias for 82 operators nonresponding to both censuses) = 9,196,894.

35 The operator-segment-terminal-vessel level is a data set means each row of the data corresponded to a segment where an operator used a vessel to ferry between two terminals. Among these rows, 42 rows are missing at least the operator ID, segment ID, terminal ID, or vessel ID. For example, 34 rows do not have segment IDs.

36 Unit response was determined when valid values exist in (1) Operator Name, (2) at least two sets of Route Origin and Rout Destination, and (3) at least one Vessel Name. Please refer to Section 3.2. Determining Unit Response for the definition. All other cases resulted in unit *nonresponse*.

A conditional tree analysis was performed for each of the variable settings in Table 15 with and without state variables. Since the information of the state in which a ferry operator was located might be too specific or detailed, the analysis was performed with and without the state variable. The number of cases by state varies substantially, ranging from 2 to 126, and some states only had one operator serving two ferry segments. Thus, when such a state was identified to influence nonresponse, this likely indicates an issue with the operator in that state, not with the state itself. In these cases, it is impossible to differentiate between state effects and operator effects.

Four population variables were considered for the tree analysis: total population, population density (per square miles), indicator for metropolitan area, and 3-level urban/rural indicator. The best population variable was selected using the train-test split method based on the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) value. When a tree model with a population variable had the highest AUC value among the four models, it was determined to be the best model among the four models, and thus, the population variable included in the best model was deemed to be the best among the four variables. When the AUC values of the four models were identical, the total population variable was selected.

4.3.1. Unit Response

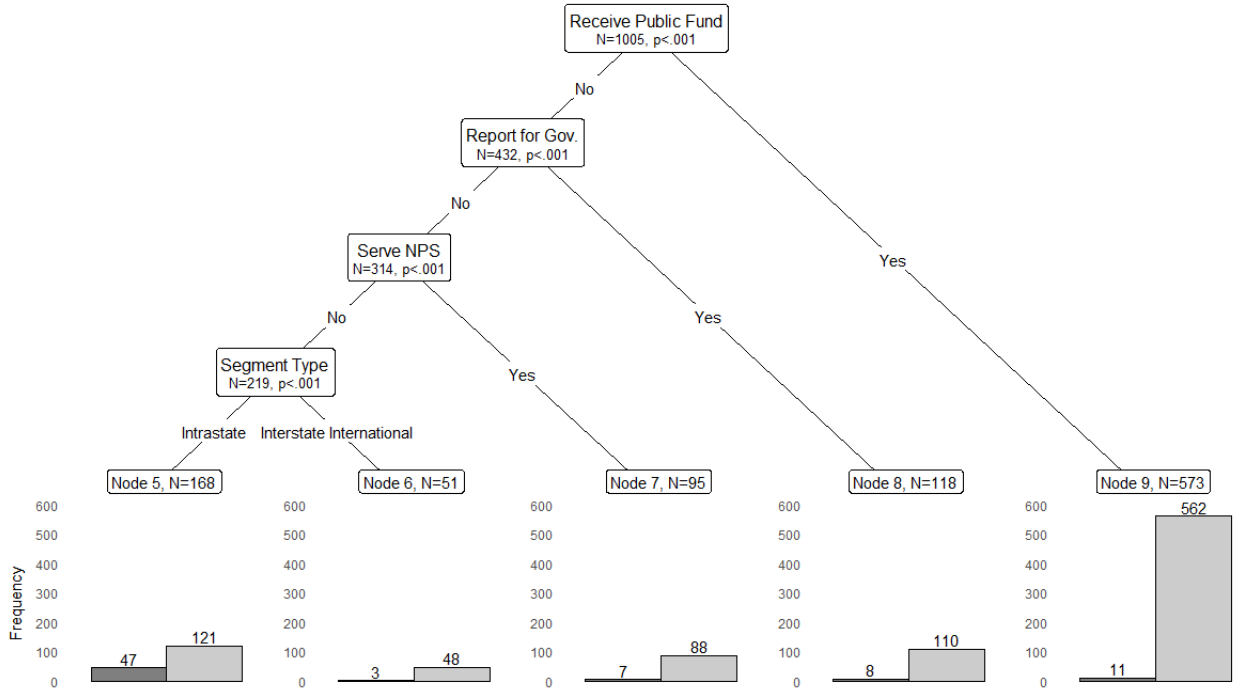
The target variable is a unit response. The target variable was recorded 1 (Yes) when an operator reported its name, at least two segments (two sets of origin and destination names), and at least one vessel name, and otherwise 0 (No). Tree models were developed with and without the state variables and the resulting trees were different.

To determine the best population variable for tree analysis, four tree models were developed, each with a different population variable. For the tree models without the state variables, all the four tree models resulted in the same AUC value (0.746). For the tree models with the state variables, the model with the total population resulted in the highest AUC value (0.840) while the other three models produced the same AUC value (0.838). Thus, the total population variable was included in the analysis to develop the final tree.

4.3.1.1. Without State Variables

Figure 10 shows the resulting tree without the state variable. Node 5 shows a higher nonresponse rate than the other terminal nodes. At Node 5 where 168 cases are gathered, 28 percent of them were mapped to operators determined to be nonresponsive. These cases correspond to operators not accepting public funds, not reporting on behalf of government, not serving national parks, and operating segments within a state.

Figure 10. Conditional Inference Tree Without State Variables for Unit Nonresponse

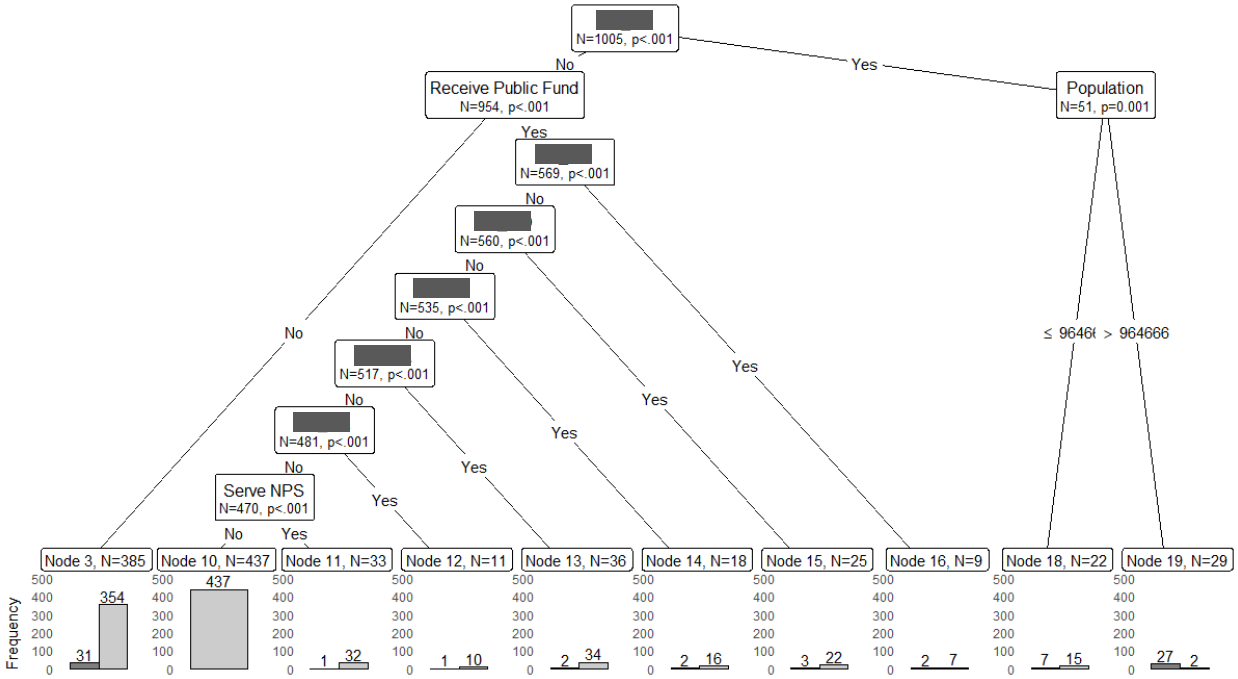


Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: *P*-value presented in an inner node was Bonferroni-adjusted *p*-value. Two bars of a bar plot at a terminal node correspond to nonresponse (left) and response (right).

4.3.1.2. With State Variables

Including the state variables changed the resulting tree (Figure 11). Several states were identified in the tree, and their names were redacted. Node 19 shows a high nonresponse rate and includes 29 cases. These cases correspond to operators located in a specific state, noted in the root node (i.e., Node 1), with their terminals found in high populous areas with more than 964,000 residents. To increase responses of these operators, it would be helpful to identify and work with stakeholders in that state who could advocate for the NCFO, such as industry associations for vessel operations or marine transportation.

Figure 11. Conditional Inference Tree with the State Variables for Unit Nonresponse



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: *P*-value presented in an inner node was Bonferroni-adjusted *p*-value. Two bars of a bar plot at a terminal node correspond to nonresponse (left) and response (right). State names were redacted.

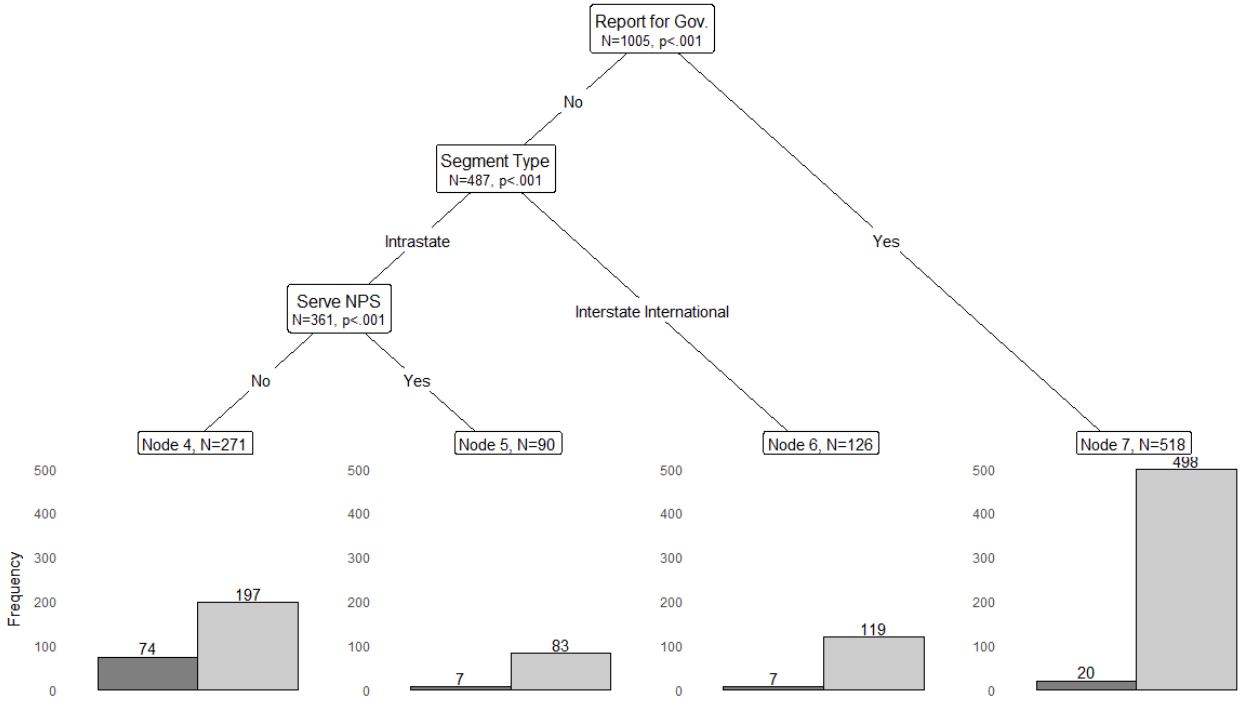
4.3.2. Response on Three Key Variables

A tree analysis was performed on each of the three key variables (i.e., passenger boardings, vehicle boardings, and segment length). As discussed earlier (Table 15), there were 108 missing cases for each variable, and the missing cases were identical across the three variables. Thus, the analysis results were found to be identical, and only the results of passenger boarding are presented.

4.3.2.1. Passenger Boarding without the State Variables

Figure 12 shows that the resulting tree without the state variable is similar to Figure 10, except the Receive Public Fund variable was excluded and Report for Government and Serve NPS positions were switched. Node 4 shows a higher nonresponse rate than the other nodes. Node 4 includes 271 rows with 74 rows with missing passenger boarding (i.e., 27 percent of nonresponse rate). These cases correspond to operators not reporting on behalf of government, operating within-state segments, and not serving national parks. The other three terminal nodes have nonresponse rates of 4 to 8 percent.

Figure 12. Conditional Inference Tree without State Variables for Nonresponse on Passenger Boarding



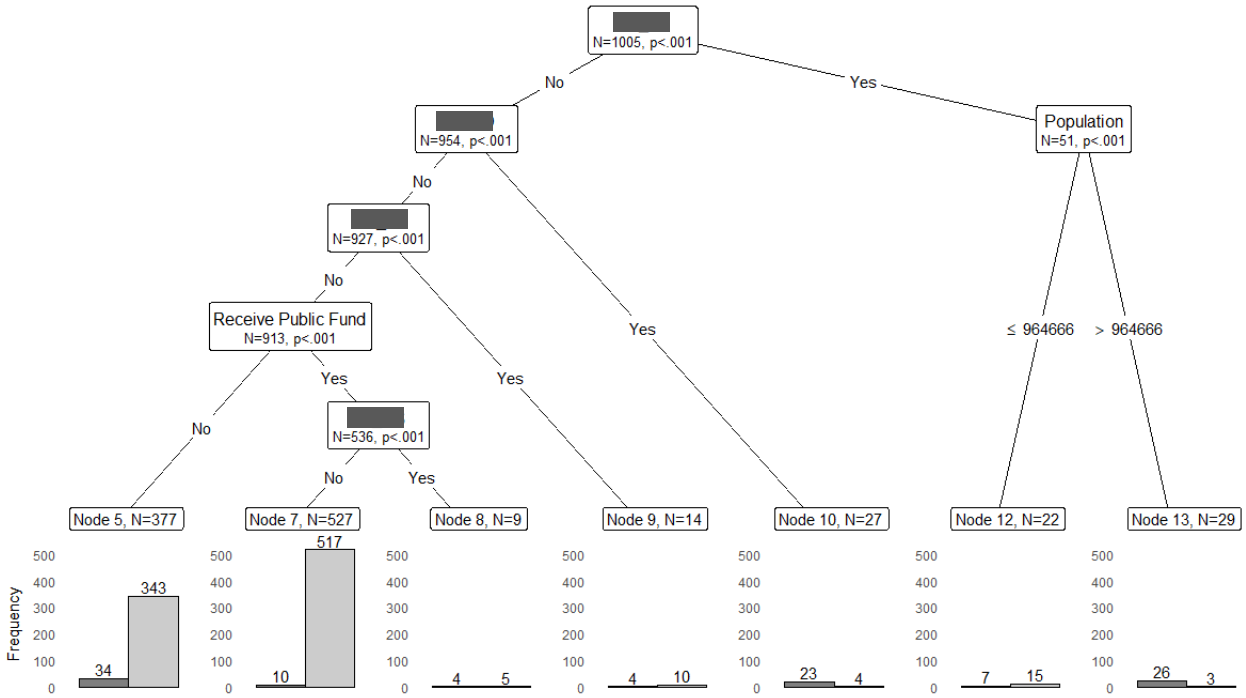
Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: *P*-value presented in an inner node was Bonferroni-adjusted *p*-value. Two bars of a bar plot at a terminal node correspond to nonresponse (left) and response (right).

4.3.2.2. Passenger Boarding with State Variables

Including the state variables changed the resulting tree (Figure 13). Several states were identified in the tree, and their names were redacted. Nodes 9, 10, 12, and 13 show a high nonresponse rate. However, Nodes 8 and 9 have 9 and 14 rows, respectively. Nodes 10, 12, and 13 correspond to two specific states. Node 13 corresponds to operators located in a specific state, noted in the root node (i.e., Node 1), and their terminals were found in high populous areas with more than 960,000 residents. The resulting tree is similar to that for the unit response (Figure 11), except Serve NPS was included, there were fewer state variables, and the location of Receive Public Fund variable was different in the tree.

To increase these operators' overall responses and item-specific responses on passenger boarding, vehicle boarding, and segment length, it would be helpful to identify and work with stakeholders in each state who could advocate for the NCFO, such as industry associations for vessel operations or marine transportation.

Figure 13. Conditional Inference Tree with State Variables for Nonresponse on Passenger Boarding



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.
Note: *P*-value presented in an inner node was Bonferroni-adjusted *p*-value. Two bars of a bar plot at a terminal node correspond to nonresponse (left) and response (right). State names were redacted.

5. Conclusions

Results from analyzing nonresponse in the 2020 NCFO data led to the following conclusions:

- **Response rates varied across subgroups.** Unit response rate for the 2020 NCFO was 65 percent, yet the rate was found to vary across subgroups formed by reporting obligation and whether they accepted public funds. Operators that reported on behalf of government or operators that accepted public funds had notably higher response rates than their counterparts by 21 and 15 percentage points, respectively, which was statistically confirmed at 0.05 level. Subgroups by ticket revenue (i.e., <50% vs. ≥50%), number of segments (i.e., 2, 3-6, and >6), and number of vessels (i.e., 1-2, 3-6, and >6) also show variation in response rates, but they were not statistically significant at the 0.05 level; Ticket Revenue subgroup is significant at the 0.1 level. This implies data users focusing on a specific subgroup or comparing across subgroups should be cautious in interpreting these results.
- **A few large operators not responding to boarding counts were responsible for a large proportion of the bias in the total boarding counts of the respondents.** A ferry operator was considered large if they carried at least two million passengers or a half million vehicles annually. The estimated biases resulted in an undercount of about 40 million passengers and 2 million vehicles due to nonresponses, which correspond to 35 and 8 percent of the observed total boarding counts. It should be noted that these biases were for the limited numbers of ferry operators having provided essential data, not for the entire ferry population. As for the total passenger boarding count, four large operators failing to report their passenger boarding accounts for 60 percent of the total bias. This shows the importance of obtaining boarding count data from the large operators.
- **A probable range of the total bias for the entire ferry population was estimatable with assumptions.** Assumptions were required for nonparticipating ferry operators in their probable boarding counts, because only contact information is known for these operators. With the assumptions, the estimated nonresponse bias in the total passenger boarding count for the entire ferry population ranges from 48 million to 73 million passengers in comparison to the observed total of 114 million passengers. The estimated bias in total vehicle boarding ranged from 3.9 million to 9.2 million vehicles in comparison to the observed total of 26.6 million vehicles. Thus, the estimated national boarding counts in 2019 were between 162 million and 187 million passengers and between 31 million and 36 million vehicles.
- **Ferry operators in a specific state with certain characteristics were less likely to respond.** Based on conditional inference tree analysis, nonresponding operators were located in a specific state, especially with their terminals located in high populous areas. Targeting these operators for outreach and follow-up could increase participation and response in future census.

6. Recommendations

Based on the findings and conclusions, the following actionable items for improving data quality of a future NCFO are proposed:

- **Develop a process to keep track of responses for each of the key items.** The process will identify nonresponding ferry operators as a census rolls out so that BTS can quickly follow up with these operators. This recommendation would increase the unit response rate and assist in collecting quality data.
- **Develop a list of ferry operators grouped by historical boarding counts.** The list will help BTS identify which ferry operators would be critical in obtaining boarding count data so that BTS can prioritize follow-up contacts based on the list. This recommendation would increase the unit response rate and the item response rates on for the two boarding count items.
- **Develop a process to identify abnormal changes in three key items (passenger boarding, vehicle boarding, and segment length).** This process will help to identify responding ferry operators whose values could suffer from input errors, such as accidentally adding or excluding a digit in boarding counts. When an abnormal change is identified, BTS will follow up with the corresponding operator to verify the veracity of its input, and in the case an error is found, the operator could correct it in a timely fashion. This recommendation would improve the quality of data of the three key items.
- **Consider adding a question to the Segment Information section of the census questionnaire asking which cargo types (passengers, vehicles, and freight) are included at a segment level.** The proposed question in Segment Information section would improve data quality of the two boarding count items (passenger and vehicle boarding) and facilitate imputation of missing boarding counts by easily verifying zero vehicle boarding counts. A cargo type question was asked at the vessel level in the previous censuses. However, cargo type information at the vessel level is hard to use for verifying any restriction on a specific cargo type at a segment level. For example, when a segment carries only passengers, not vehicles, a ferry operator should enter zero to the vehicle boarding count. With the proposed question in in the survey, zero vehicle boarding would be easily verified based on the response to this question. Otherwise, cargo type information for all vessels serving that segment will need to be verified. This recommendation would improve the quality of data of the boarding counts.

7. References

American Association for Public Opinion Research (AAPOR). 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 5th edition. Lenexa, Kansas: American Association for Public Opinion Research. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf. Last accessed August 22, 2022.

Kweon, Young-Jun. 2021. *Quality Assurance Analysis Results of 2020 NCFO Pre-released Data* [internal document]. Washington, DC: U.S. Department of Transportation, Bureau of Transportation Statistics, Office of Data Development & Standards.

Lineback, Joanna Fane and Katherine J. Thompson. 2010. "Conducting Nonresponse Bias Analysis for Business Surveys, *Proceedings of the American Statistical Association, Section on Government Statistics*. Alexandria, VA: American Statistical Association.

National Institute of Standards and Technology. 2012. "2.1.1.3. Bias and Accuracy," *NIST/SEMATECH e-Handbook of Statistical Methods*. Washington, DC: U.S. Department of Commerce. https://www.itl.nist.gov/div898/handbook/mpc/section1/mpc113.htm. Last accessed November 10, 2023.

Nguyen, Aubrey. 2022. *Automate Data Edit Process for the National Census of Ferry Operators Using R Markdown* [slideshow]. Washington, DC: 2022 FCSM Research & Policy Conference.

Office of Management and Budget (OMB). 2006. *Standards and Guidelines for Statistical Surveys*. Washington, DC: Office of Management and Budget.

Phipps, Polly and Daniell Toth. 2012. "Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data," *The Annals of Applied Statistics*, 6, no. 2. Ann Arbor, MI: Institute of Mathematical Statistics. https://doi.org/10.1214/11-AOAS521. Last accessed May 7, 2024.

Strasser, Helmut and Christian Weber. 1999. *On the Asymptotic Theory of Permutation Statistics*. (January 1999 ed.) Report Series SFB Adaptive Information Systems and Modelling in Economics and Management Science, no. 27. Vienna, Austria: WU Vienna University of Economics and Business. https://doi.org/10.57938/ff565ba0-aa64-4fe0-a158-86fd331bee78. Last accessed May 7, 2024.

U.S. Census Bureau. 2022. *U.S. Census Bureau Statistical Quality Standards*. Washington, DC: U.S. Department of Commerce. https://www2.census.gov/about/policies/quality/quality-standards.pdf. Last accessed August 8, 2022.

Appendix A. Table-Based Item Response Rates

Calculation of IRR was done for each table and $IREQ_x$. The denominator (i.e., the number of responses required for item *x*) changes depending on the table to which the item *x* belongs, and thus $IREQ_{x,y}$ and $IRR_{x,y}$ are accurate. For example, *IREQ* for the Accept Public Funding item the number of ferry operators having participated in the 2020 census (i.e., 164 operators) since the item is in Operator table; please note that 164 ferry operators out of 246 invited to the 2020 NCFO had submitted their data (i.e., participated) and their data were released in the 5 tables.

Meanwhile, *IREQ* for Segment Name item is the number of segments that 164 operators should have reported on the 2020 census. The true number of segments was unknown since the operators may have not reported all their segments. Through the data edit process (e.g., automated data edit and analyst's manual review), BTS made efforts to verify the reported segments and, if segments were verified missing, add missing segments. These efforts led to adding several segments that the operators failed to report in the 2020 census. For example, when an operator reported only two terminals but reported one segment, BTS added the pairing reverse segment for a returning trip since a ferry most likely operated two ways between the terminals.[37] However, it was possible that not all segments that the operators failed to report were discovered by BTS's data edit process. *IREQ* for Segment Name item was 934 in the 2020 NCFO since there were 934 segments in Segment table where Segment Name item was found; the 164 ferry operators participated in the 2020 NCFO should have reported 934 segments including segments reported by the operators and those added by BTS.

Original variable names were used and they were matched with those in the published data tables and the data dictionary of the 2020 NCFO (Table 16).

Table 16. Table-Based Item Response Rate of 47 Variables in 2020 NCFO

Variable (*x*)	Table (*y*)	Number of Rows[a]	Percent $IRR_{x,y}$
Accepts_public_funding	Operator	164	100
Federal_state_local			100
Funding_revenue			77
Op_state			100
Op_strcity			100
Op_strzip			100
Operator_name			100
Trip_purpose			87
Average_trip_time	Operator-Segment	963	96
Avg_daily_brd_pax			80
Avg_daily_brd_veh			79
Most_used_vessel_id			96
Passengers			96
Route_rate_regulator			90
Route_rates_regulated			90
Segment_length			96
Segment_season_end			95
Segment_season_start			95

[37] One segment was where a ferry ran from Terminal A to Terminal B and the pairing reverse segment occurred when a ferry ran from Terminal B to Terminal A. When there were more than two terminals, there might be no pairing segments due to a possible looping route (e.g., a route of Terminal A → Terminal B → Terminal C → Terminal A).

Variable (*x*)	Table (*y*)	Number of Rows[a]	Percent $IRR_{x,y}$
Trips_per_year			96
Vehicles			96
Vessel_id1			90
Segment_name	Segment	934	96
Access_mode	Terminal	650	82
In_operation			99
Term_city			99
Term_state			99
Terminal_name			99
Terminal_operated_by			99
Terminal_operation			99
Terminal_owned_by			99
Terminal_ownership			99
Ada_accessible	Vessel	756	80
Cargo_type			95
Census_year_miles			95
Expected_lifespan			74
Fuel			91
Fuel_mileage			79
In_service			100
Passenger_capacity			95
Typical_speed			86
Vehicle_capacity			95
Vessel_name			100
Vessel_operated_by			95
Vessel_operation			95
Vessel_owned_by			95
Vessel_ownership			95
Year_built			95

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

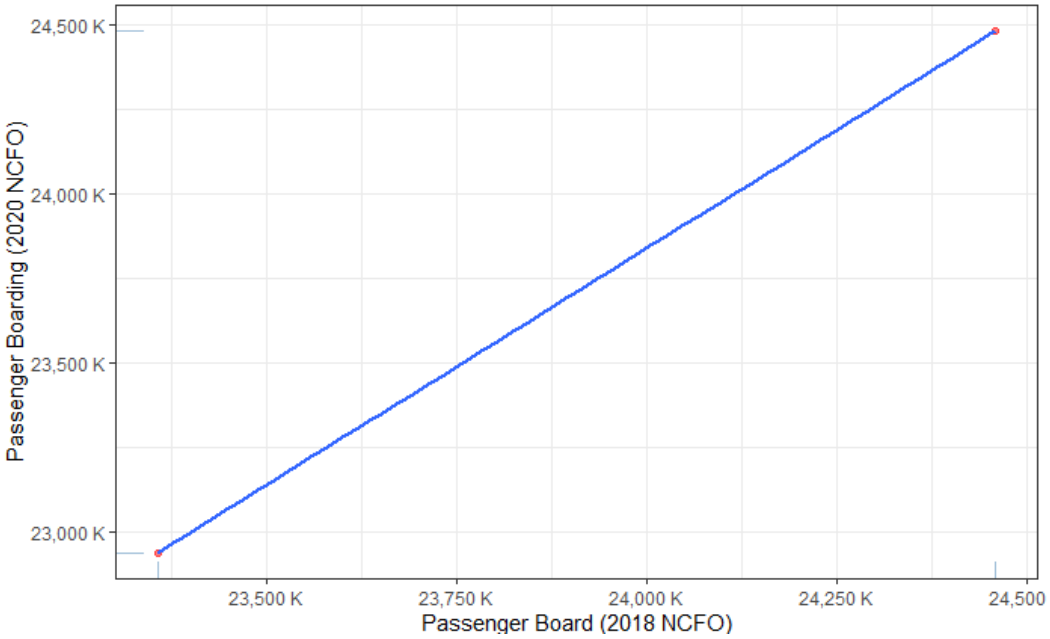
[a] Number of rows in table *y* = $IREQ_{x,y}$.

Appendix B. Visualization of Growth Ratios by MOS Group

Growth ratios shown in Table 9 and Table 10 are fundamentally identical to the slope coefficients of linear regressions without intercept between boarding counts in the two census years. The regression lines are visualized for passenger and vehicle boardings by MOS groups, separately. It should be noted that the regression for Group 1 (Figure 14) has only two data points, thus the residuals are zero.

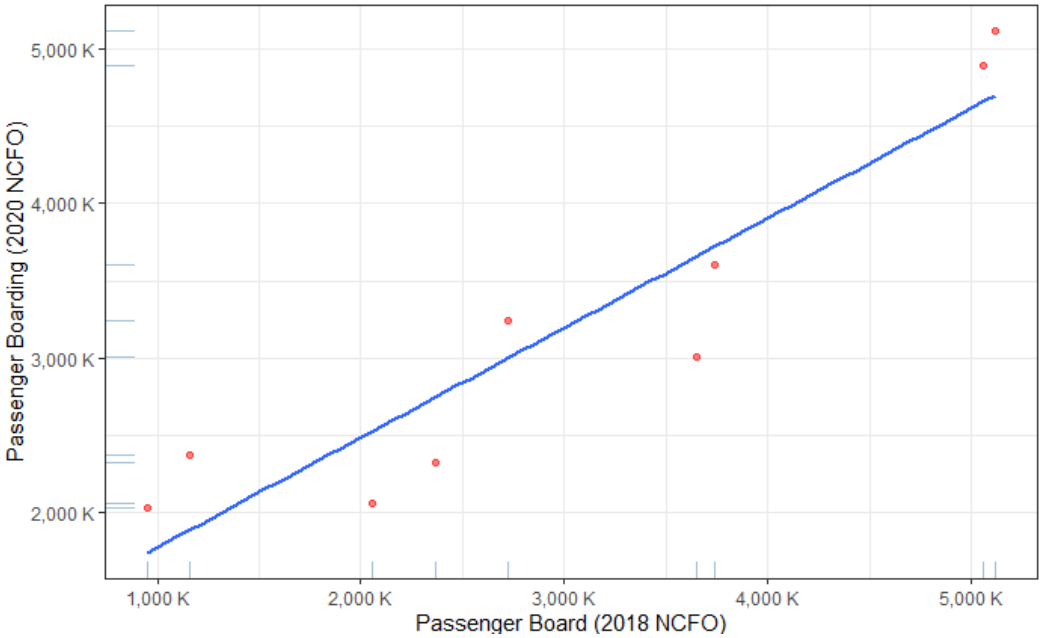
PASSENGER BOARDING

Figure 14. Growth Ratio of Passenger Boardings in 2018 and 2020 NCFO (Group 1)



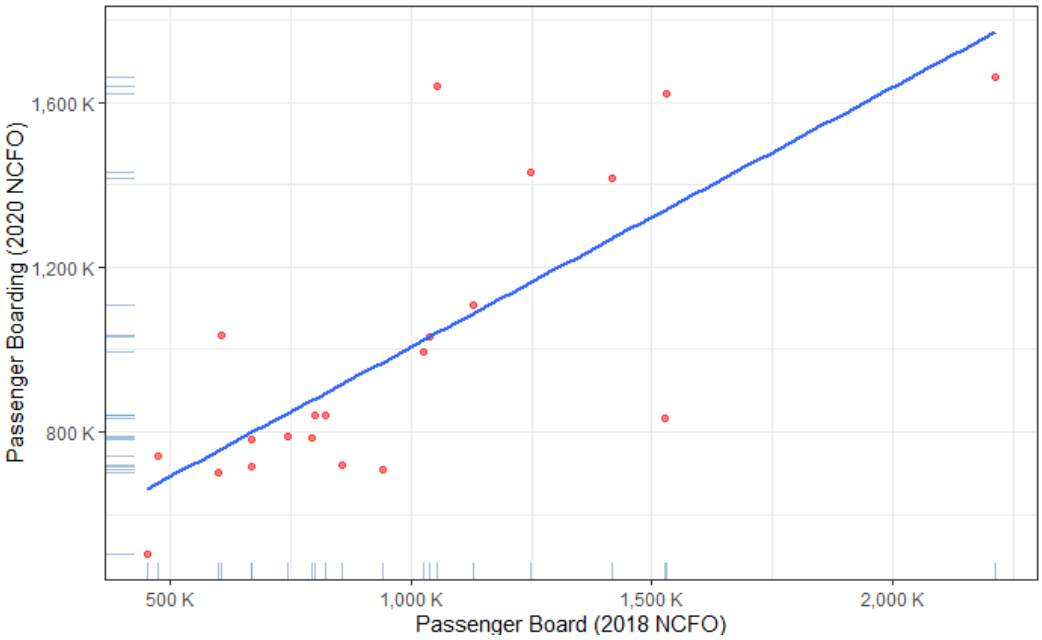
Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

Figure 15. Growth Ratio of Passenger Boardings in 2018 and 2020 NCFO (Group 2)



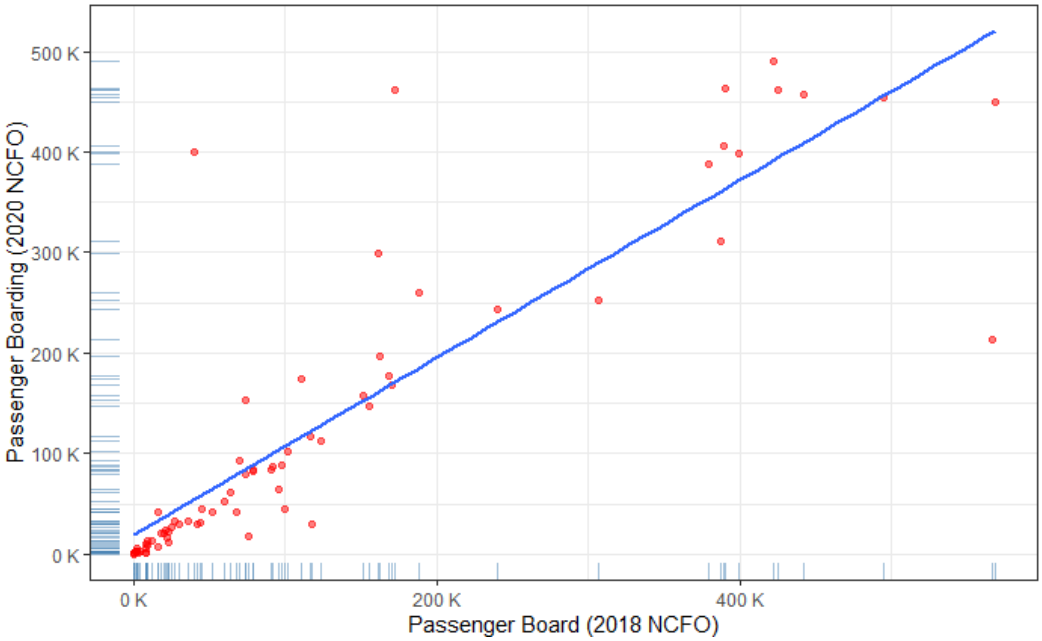
Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

Figure 16. Growth Ratio of Passenger Boardings in 2018 and 2020 NCFO (Group 3)



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

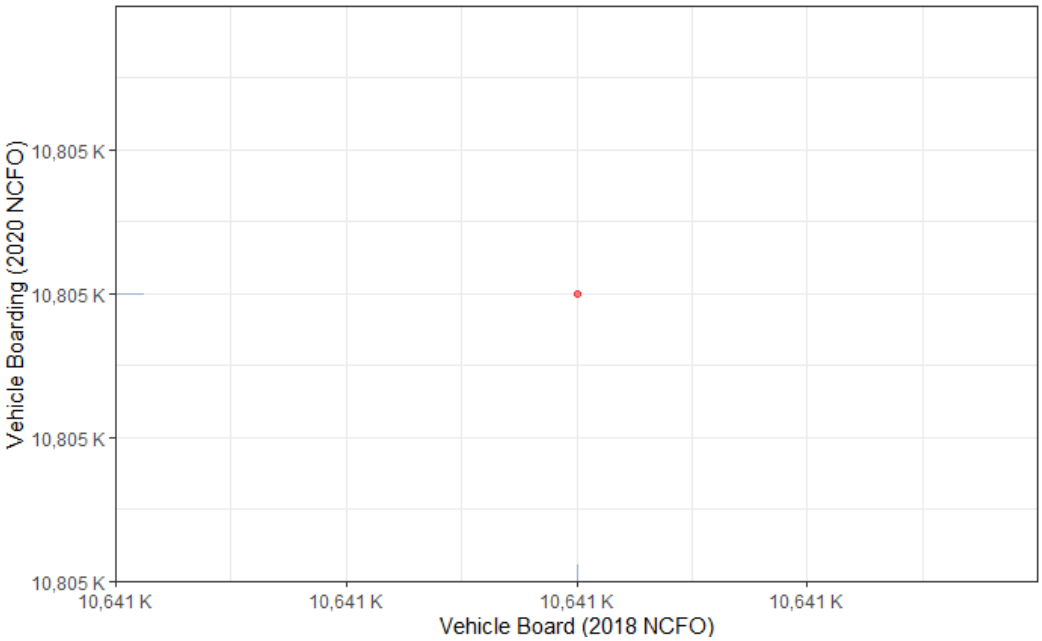
Figure 17. Growth Ratio of Passenger Boardings in 2018 and 2020 NCFO (Group 4)



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

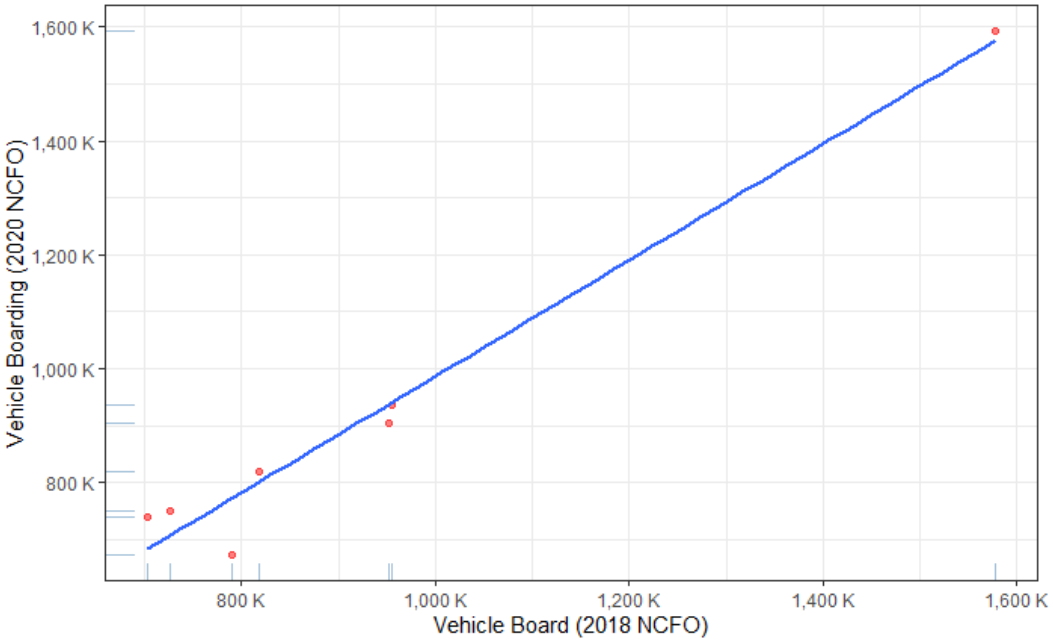
VEHICLE BOARDING

Figure 18. Growth Ratio of Vehicle Boardings in 2018 and 2020 NCFO (Group 1)



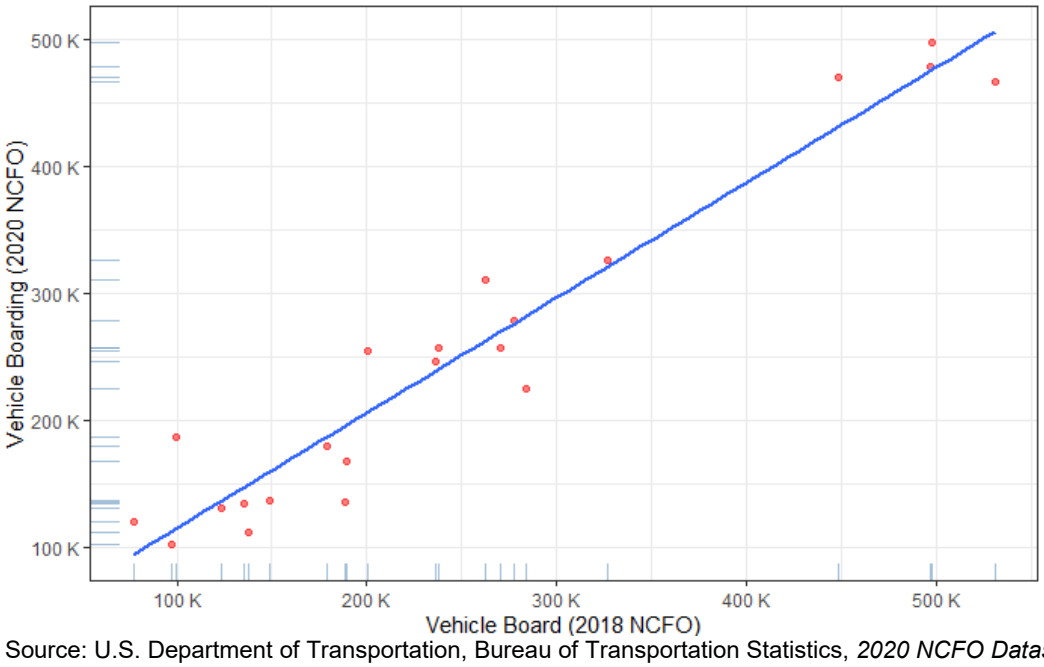
Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

Figure 19. Growth Ratio of Vehicle Boardings in 2018 and 2020 NCFO (Group 2)



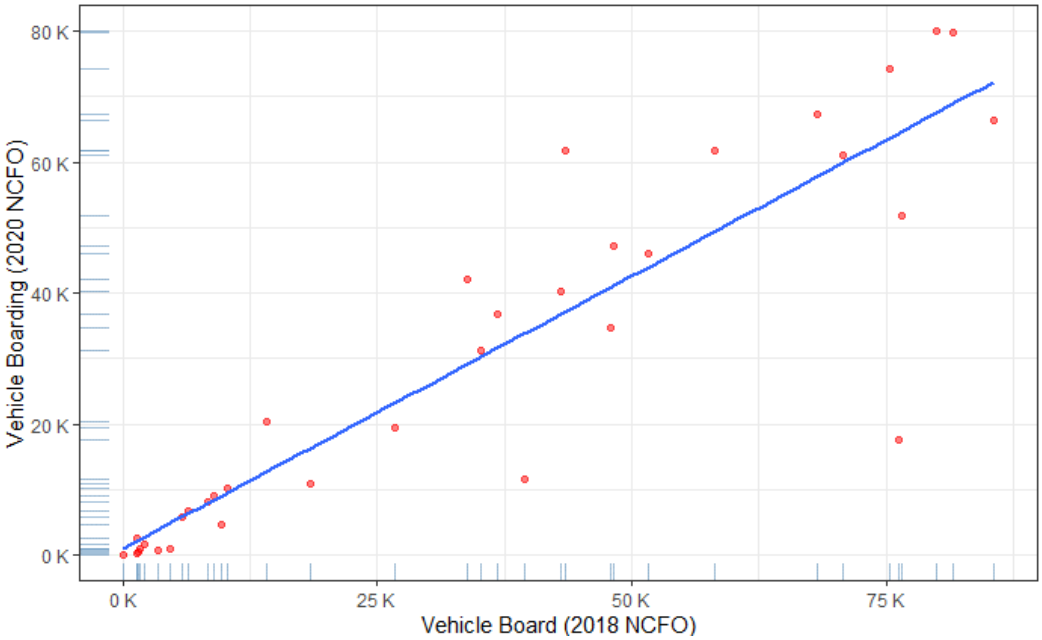
Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

Figure 20. Growth Ratio of Vehicle Boardings in 2018 and 2020 NCFO (Group 3)



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.

Figure 21. Growth Ratio of Vehicle Boardings in 2018 and 2020 NCFO (Group 4)



Source: U.S. Department of Transportation, Bureau of Transportation Statistics, *2020 NCFO Dataset* and *2018 NCFO Dataset*, available at https://www.bts.gov/NCFO as of April 2024.