U.S. Department of Transportation
Research and Innovative Technology Administration

TR-004

Bureau of Transportation Statistics
Technical Report

August 2008

Multiple Imputation of Missing Passenger Boarding Data in the National Census of Ferry Operators

by Lee H. Giesbrecht

The Bureau of Transportation Statistics (BTS), a component of the Research and Innovative Technology Administration (RITA) of the U. S. Department of Transportation (DOT), conducted the National Census of Ferry Operators in 2006. This data collection updated information collected by the Federal Highway Administration in 2000. The resulting database contains ferry operation data for calendar year 2005 along with other sources of ferry data such as the U.S. Coast Guard and the Army Corps of Engineers. Ferry operators were asked about their season of operation, vessel fleet, modes of access to their terminals, and information about the route segments that they serve between terminals such as the route segment length, average trip time, and the number of passengers served.

Ferry operations included are those providing itinerant, fixed route, common carrier passenger and/or vehicle ferry service. Ferry operations that are exclusively nonitinerant (e.g., excursion services—whale watches, casino boats, day cruises, dinner cruises, etc.), passenger-only water-taxi services not operating on a fixed route, LoLo (Lift-on/Lift-off) freight/auto carrier services, or long-distance passenger-only cruise ship services are not included within the scope of this census. The geographic scope includes ferries operating within the United States and its possessions, encompassing the 50 states, Puerto Rico, the U.S. Virgin Islands, and the Commonwealth of the Northern Mariana Islands. In addition to ferry operators providing domestic service within the United States and its possessions, operators providing services to or from at least one U.S. terminal are also included.

BTS identified 230 ferry operators that were in business in 2005 that fall within the scope outlined above. Of those, approximately 92 percent responded to the census questionnaire. Data are missing because not all ferry operators responded to the census. However, some data variables for nonresponding ferry operators were completed based on information from other sources (e.g., vessel characteristics). In particular, passenger and vehicle boarding data are blank in the database for ferry operators that did not respond to the census, did not have access to these numbers, refused to report them, or required BTS to keep them confidential.

Need for Imputation

About 15 percent of the ferry route segments (part of the ferry route between two terminals) have missing values for passenger boarding data in the 2006 National Census of Ferry Operators. The sum of passengers for all nonmissing values (including those for which the operator required confidentiality) is about 89 million. This incomplete count is arguably less useful than an estimate of all passengers, which would include the 15 percent of route segments with missing values. Estimates of passenger-miles traveled and other passenger-related statistics will also not be as useful unless they are based either on complete or accurate estimates of passenger boarding data.

Multiple imputation techniques allow values to be imputed for missing data along with a measure of variability for estimates computed from the imputed values. As a federal statistical agency, the Bureau of Transportation Statistics strives to fully inform users about the quality of its data. Providing users with a measure of the variability added to the data due to imputation helps to satisfy this goal.

The Imputation Model

The basic idea of multiple imputations is to impute plausible values for the missing data (in this case, missing passenger data) from a distribution of values multiple times. This way, one can estimate distributions from the multiple replicates of the data. The method chosen for the missing ferry passenger boarding data fits a linear regression model that uses auxiliary information about the source of the missing data (ferry operator and route segment variables) along with prior data (from the 2000 data collection) to construct probability distributions of plausible values from which to impute the number of passengers. This method has the advantage of using a model to compute values of the missing data based on information known about the source (which results in better imputed values), while also providing a measure of uncertainty around estimates that make use of the missing values. The technique was first posited by Rubin (1987),[1] who is credited as its developer. Multiple

[1] Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

imputation is widely regarded, today, as the method of choice due to its appropriate treatment of imputation variance.

As previously mentioned, the passenger data are at the ferry route segment level. This requires that *covariates* be standardized at this level of analysis. The covariates in the imputation model are variables with nonmissing values that will be used as predictors of the missing passenger values. These variables are not all descriptive of the ferry route segment. Some are descriptive of the ferry operator, the ferry vessels, or the ferry terminals. As large a number as possible of covariates (as long as they are logically related to the number of passengers) is desired to improve the predictability of the imputation model. The process of fitting the model began with a version that included 10 variables and no geographic information. This model resulted in a very wide range of imputed values and, therefore, a large imputation variance. It was felt that geographic information would greatly improve the model's predictive power. Subsequent changes to the model included a metro/nonmetro variable for both terminals, terminal access variables, and an indicator of whether either terminal served a national park. Census division for either terminal was added. Each addition reduced the variance of the estimate of total passengers due to imputation. Finally, the Census division variables were replaced with variables for each state. This resulted in a model that would not converge. It may be that the state variables resulted in too many unique records. The final model included the following variables listed in box A.

SAS Proc MI (version 8.02)[2] was used to run 10 multiple imputations for missing 2005 passenger boarding values. The Markov Chain Monte Carlo (MCMC)[3] option was used with an informative prior distribution based on the 2000 data that contained all the same variables as the main model from 2000. One exception is the variable that indicates the operator requested that his passenger data be kept confidential. This was not asked in 2000, so the values for this variable in the prior distribution were all zero.

Determining a Maximum Imputation Value

A very conservative approach was used to determine a range of plausible values from which to impute. No minimum value was specified, and the maximum value was based on the assumption that vessel capacity would be fully utilized at each route segment.

[2] Documentation for the SAS MI Procedure can be found as of the date of this publication at http://support.sas.com.
[3] More about the MCMC option in SAS Proc MI is contained in the SAS documentation. For further reading about Markov Chain Monte Carlo techniques, see http://www.stat.columbia.edu/~liam/teaching/neurostat-spr07/papers/mcmc/mcmc-gibbs-intro.pdf as of May 2008.

Box A: Names and Descriptions of Variables Used in Final Model

Variable name	Variable description
OPID	Operator ID number
SEGID	Segment ID number
MACROSYS	Segment length from contractor computations
SEGLEN	Segment length reported by ferry operator
AVGTIME	Average one-way travel time reported by operator
MONTHS	Number of months ferry operates per year
PASSENGERS	Number of passengers per year
CONF	Operator requested passenger data be kept confidential
BOATS	Total number of vessels operator runs (not per segment)
AVGCAP	Average capacity of vessels
METRO	At least one of the two segment terminals is in a core based statistical area (CBSA)
NPS	At least one of the two segment terminals serves a national park
AUTO	At least one of the two segment terminals has auto access
PARKING	At least one of the two segment terminals has parking
TRANSIT	At least one of the two segment terminals has access to public transit bus
INTERCITY	At least one of the two segment terminals has access to intercity bus
LHRAIL	At least one of the two segment terminals has access to light or heavy transit rail
AMTRAK	At least one of the two segment terminals has access to Amtrak rail
NEWENGLAND	One of the two segment terminals is in New England Census Division
MIDATLANTIC	One of the two segment terminals is in Mid Atlantic Census Division
ENCENTRAL	One of the two segment terminals is in East North Central Census Division
WNCENTRAL	One of the two segment terminals is in West North Central Census Division
WSCENTRAL	One of the two segment terminals is in West South Central Census Division
ESCENTRAL	One of the two segment terminals is in East South Central Census Division
SATLANTIC	One of the two segment terminals is in South Atlantic Census Division
MOUNTAIN	One of the two segment terminals is in Mountain Census Division
PACIFIC	One of the two segment terminals is in Pacific Census Division

SOURCE: Variables used in this model were selected from the U.S. Department of Transportation, Research and Innovative Technology Administration, Bureau of Transportation Statistics, National Census of Ferry Operators, 2006, or else derived from variables in this file.

SAS Proc MI allows the analyst to control the range of valid values for the imputed variable. A reasonable upper limit for the imputation of missing passengers was determined based on information about vessels and route segments. The upper limit for the missing data was set as the total annual passengers computed using the following logic.

The following route segment and vessel information were considered:

- **Segment length** – Missing segment lengths were imputed with values computed using geographic information system software. The software computed segment length using precise coordinates for the two ferry terminals and presumed waterway paths between the terminals from the U.S. Army Corps of Engineers Navigable Waterway Network GIS database.
- **Average travel time per segment** – Missing values for average time were imputed with the average time of 14.1 minutes per mile for all route segments with missing passenger data but with average time reported.
- **Number of months segment operated per year** – Route segments with missing values for the number of operating months were assumed to be operated year-round and 12 months was imputed.
- **Number of vessels available per segment** – The number of vessels available was computed by dividing the total number of vessels per operator by the number of route segments per operator.
- **Average vessel capacity per operator** – This was computed by summing the vessel capacity fields (if operator-reported capacity was missing, data from the U.S. Coast Guard were used) and dividing by the number of vessels for each operator.
- **Average capacity per route segment** – The number of vessels per segment was then multiplied by the average vessel capacity per operator (because multiple vessels may be used on the same route segment) to get the capacity available for each route segment.

The number of runs per year was based on an assumption of an 8-hour work day (no data on work day length were available from the ferry survey). It was assumed that each round trip took the average time multiplied by two with zero time to load and unload passengers. This was multiplied by the number of days operating per year (based on the number of operating months per year). This estimated value of the total number of trips per year was then multiplied by the available passenger capacity per trip, thereby resulting in the maximum possible number of passengers for that ferry route segment. The highest passenger count possible based on these criteria for the highest capacity missing route segment was 7,358,400. The lowest upper limit that could be used for a missing route segment that still allowed the imputation model to converge was about 400,000. The imputation model was run for each missing route segment using the upper limit computed as described above and a lower limit of zero for each missing route segment.

Imputation Results

Table 1 shows the estimated total annual passengers and passenger miles for all states, along with their associated 95 percent confidence intervals (CI) and coefficients of variation (CV)[4]. The CIs and CVs were computed based on the standard deviations across 10 imputation replicates.

It is likely that even the lower bound of the 95 percent confidence interval of 104 million passengers is much closer to the actual number of passengers for 2005 than the total computed without imputation of about 89 million because it is the most conservative estimate that accounts for the missing data. Some other estimates based on the imputed passenger data include state totals for California and Washington of 9,350,649 and 14,695,039, respectively. These estimates also have imputation error associated with them of plus or minus 592,402 and 380,069, respectively. The total number of passengers for Alaska, 711,809, has no imputation error because there were no route segments with missing data. Note that several other states have no imputation error as well. Estimates for some states, such as Massachusetts, may still be useful despite having a 4.9 percent CV for passengers, but many other states have imputation errors too large for accurate reporting. It should also be noted that state totals cannot be revealed for four states, New York, Connecticut, South Carolina, and Wisconsin (see first row of table), due to confidentiality restrictions. None of these states had any missing passenger data, but reporting a state total for any of these states would reveal the confidential data for some ferry operators.

Next Steps

The estimated number of passengers for 2005 may now be reported by BTS, along with its estimated variance due to imputation. Other estimates based on the passenger boarding data may also be computed and reported, such as passenger miles traveled. Care will be taken to ensure confidentiality for operators who requested their data be kept confidential. The confidential data were used in the production of the imputation replicates.

It may be possible to impute data for missing passenger reports from the 2000 ferry database and compare results. The methodology for imputing the 2000 data, however, must be different because there is no source for an informative prior census, which will likely result in a larger variance due to imputation. It is not clear whether or not a statistically significant difference could be detected using this methodology. It may also be possible to further reduce the range of plausible values for each operator with missing values on an individual basis. These ideas may be explored in future research.

In the next round (2008) of the census, additional data collection should be considered to better inform the imputation process for this variable. Information such as the schedule/number of trips for each route, the usual vessel for each route, and whether or not the route includes vehicles or is a passenger-only route would be helpful in this effort.

[4] The CV is computed by dividing the standard error by the estimate.

Table 1: Total Estimated Annual Passenger Miles, All States

State	Passengers	CI (+/-)	CV	Passenger Miles	CI (+/-)	CV
NY/CT/SC/WI	30,773,467	—	—	159,166,232	—	—
AK	711,809	—	—	27,765,052	—	—
AL	26,373	—	—	67,793	—	—
AR	224,123	218,306	52.3%	168,732	173,243	52.4%
AZ	476,958	731,373	76.0%	679,645	1,098,549	82.5%
CA	9,350,649	592,402	3.4%	66,559,250	74,536	0.1%
DE	100,710	—	—	9,710	—	—
FL	1,368,531	173,608	6.8%	14,117,073	85,202	0.3%
GA	2,510,672	2,679,801	57.2%	1,464,799	900,781	31.4%
HI	158,947	—	—	2,525,953	—	—
IL	1,409,578	1,235,336	43.0%	782,142	831,105	54.2%
KY	1,638,053	961,879	26.4%	1,081,744	1,204,952	56.8%
LA	6,853,093	2,713,414	16.9%	6,124,207	7,582,914	63.2%
MA	5,339,118	530,858	4.9%	42,295,505	4,836,189	5.8%
MD	1,118,809	961,358	43.2%	201,424	161,386	40.9%
ME	1,773,987	317,259	9.1%	5,436,695	660,725	6.2%
MI	3,380,338	972,310	15.3%	51,380,536	380,047	0.4%
MN	6,000	—	—	243,421	—	—
MO	310,999	—	—	2,686,118	—	—
MS	259,091	479,145	99.0%	704,229	52,148	3.8%
MT	2,220	—	—	231	—	—
NC	2,880,974	213,729	3.4%	14,449,028	1,221,618	4.3%
NH	180,455	218,194	59.9%	1,017,703	1,297,101	65.0%
NJ	9,108,890	—	—	46,283,785	—	—
OH	1,067,814	391,850	17.4%	6,196,437	1,352,301	11.1%
OR	609,227	383,932	30.7%	65,277	40,746	31.8%
PA	115,108	—	—	17,250	—	—
RI	202,678	228,704	60.7%	9,963,079	12,366,913	63.3%
TN	669,962	1,077,361	85.8%	127,550	216,207	86.5%
TX	7,498,892	1,225,272	8.6%	15,858,229	225,118	0.7%
UT	23,451	—	—	79,392	—	—
VA	1,439,977	1,152,501	29.2%	12,161,605	10,752,105	45.1%
VT	2,031,342	315,081	8.3%	3,602,117	124,127	1.8%
WA	14,695,039	380,069	1.3%	124,460,491	10,232,599	4.2%
WV	46,059	—	—	8,659	—	—
Total	**108,363,392**	**4,673,546**	**2.3%**	**617,751,094**	**16,141,448**	**1.3%**

KEY: CI = confidence interval; CV = coefficients of variation.
— = values were not imputed; therefore, there is no imputation error.

SOURCE: U.S. Department of Transportation, Research and Innovative Technology Administration, Bureau of Transportation Statistics, National Census of Ferry Operators, 2006, augmented with imputed values for passengers and passenger miles.

About this Report

This report was prepared by Lee H. Giesbrecht, survey statistician and project manager for the 2006 National Census of Ferry Operators.

This report presents findings from the 2006 National Census of Ferry Operators (NCFO) augmented with imputed values for passengers and passenger miles. Due to the imputation procedures used to calculate missing data, totals in Table 1 may not correspond to calculations obtained from using only the data in the NCFO. The 2006 NCFO data were collected from 230 ferry operators by the Bureau of Transportation Statistics, a component of the Research and Innovative Technology Administration in the U.S. Department of Transportation. The data were supplemented by other sources of ferry data, such as the U.S. Coast Guard and the Army Corps of Engineers. The database contains information on ferry systems, including operators, routes, vessels, and passenger and vehicle boarding. The ferry database is available online – www.bts.gov.

For questions about this or other BTS reports, call 1-800-853-1351, email answers@bts.gov or visit www.bts.gov.