The Art of Multiple Imputation: Overcoming Limitations Born from Missing Data among Covariates

Kenneth W. Steve & Li Leung

U.S. DOT Research and Innovative Technology Administration
Bureau of Transportation Statistics
1200 New Jersey Avenue SE Washington, DC 20590
Kenneth.Steve@dot.gov, Li.Leung@dot.gov

Abstract

The Bureau of Transportation Statistics (BTS) has conducted the National Census of Ferry Operators (NCFO) biennially since 2006. Data are collected from approximately 260 ferry operators currently operating in the United States. This data are used to maintain the national ferry database containing information regarding routes, vessels, passengers and vehicles carried, funding sources, etc. As with many surveys or censuses of businesses, ferry operators have shown a reluctance to provide information. More specifically, some operators consider passenger boarding data to be business sensitive information. While a significant number of operators simply don't provide the information, others may ask that it not be made public. This presents BTS with challenges in regards to producing accurate population parameters for ferry passenger boardings.

In an effort to generate a more useful picture of the true number of passenger boardings for the 2006 NCFO, a SAS macro for multiple imputation (MI) was employed (Giesbrecht, 2008). The current paper discusses some of the difficulties in trying to reproduce that effort within the 2008 NCFO dataset. An initial discussion of the preliminary analyses and resulting data edits is followed by a description of various MI models fit in an attempt to overcome problems associated with relatively large amounts of missing data. The missing data patterns for both years were non-monotone; therefore a Markov chain Monte Carlo method was used to estimate missing data. Initial models produced many error messages as a result of multicolinearity among regressors, implausibly imputed values as a result of lack of specification in the model itself and finally empty imputation cells as a result of missing data among regressors. In the end, a clean MI model was developed that provided properly imputed estimates of passenger boardings for all cases.

Background

Although ferries have a long history of moving passengers and freight in America, less is known about this mode of transportation than any of the other modes. Regularly surveyed, routine statistics like the number of ferry operators and the number of passengers carried were undocumented prior to the establishment of the National Census of Ferry Operators (NCFO). Part of this knowledge gap was due to the industry's structure. State and local public transportation agencies operate some ferry systems, but others are privately owned and operated. Another complication is that many operators provide ferry services as well as dinner and sightseeing cruises, whale watching and other types of excursions. As such, it is often difficult to separate these activities. Finally, variability in the size of ferry operations gives rise to dramatic differences in how they run their operations and maintain their records. These issues, coupled with the fact that the total population of operators is quite small, creates many challenges with regards to collecting and reporting ferry data in the United States.

The Transportation Equity Act for the 21st Century (TEA–21) (P.L. 105-178), section 1207(c), directed the Secretary of Transportation to conduct a study of ferry transportation in the United States and its possessions. In 2000, the Federal Highway Administration (FHWA) Office of Intermodal and Statewide Planning conducted a survey of approximately 250 ferry operators to identify: (1) existing ferry operations including the location and routes served; (2) source and amount, if any, of funds derived from Federal, State, or local governments supporting ferry construction or

operations; (3) potential domestic ferry routes in the United States and its possessions and to develop information on those routes; and (4) potential for use of high speed ferry services and alternative-fueled ferry services. The Safe, Accountable, Flexible Efficient Transportation Equity Act—A Legacy for Users (SAFETEA-LU) Public Law 109-59, Section 1801(e) requires that the Secretary, acting through the Bureau of Transportation Statistics (BTS), shall establish and maintain (biennially) a national ferry database containing current information regarding routes, vessels, passengers and vehicles carried, funding sources and such other information as the Secretary considers useful.

While the original data collection in 2000 was conducted by the Volpe National Transportation Center, a branch of the U.S. Department of Transportation (DOT) on behalf of FHWA, subsequent data collections have been conducted by BTS. The geographic scope of the NCFO 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. Ferry operations are defined as 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 NCFO database contains ferry operation data for calendar years 1999, 2005, 2007 and 2009. Along with other sources of ferry data such as the U.S. Coast Guard and the Army Corps of Engineers, the database contains operator provided information 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. BTS has made revisions to the census questionnaire at each occurrence of the NCFO to improve the nature of the data collected and maximize the usefulness of the NCFO database. The NCFO database continues to be an important source of information for various industry agencies, and various federal and state funding agencies. Still, there is reluctance on the part of many ferry operators to provide complete and accurate information with regards to various aspects of their operation, most notably passenger and vehicle boardings. A multiple imputation model was used to impute the number of missing passenger boardings for the 2006 NCFO (Giesbrecht, 2008). The paper discusses the challenges in replicating that effort for the 2008 NCFO.

Methods

BTS identified a total of 240 valid ferry operations to be included in the 2007 NCFO. A paper questionnaire was sent to each of these ferry operators to request information about their operation. Those who did not respond to the paper questionnaire were called on the telephone to encourage their participation, and potentially take their information over the phone. In the end, approximately 89 percent of the valid operations responded to the census questionnaire. Among the completed questionnaires 355 individual operator segments (i.e., a ferry route between two terminals serviced by a unique operator) were identified as active being serviced in 2007. Approximately 20 percent of these active ferry route segments had missing passenger boarding data in the 2007 NCFO.

The sum of passengers for all nonmissing values was about 87 million in the 2007 database.[1] The 2005 estimate for the number of annual passenger boardings was 108 million (Giesbrecht, 2008). The goal of this effort was to produce a national estimate of the total annual passenger boardings for US ferry vessels that was, first of all accurate, and secondly, comparable in some way to the previous estimate. To the extent possible, the same methods were used for deriving the 2007 estimates that were used for the 2005 estimates.

The previous effort to overcome the absence of passenger boarding data for all operator segments utilized multiple imputation approach. A prior covariance matrix was derived from the 2005 NCFO data and covariates were imputed based on logical decisions prior to fitting the MI models for 2007. For the current set of analyses, no prior covariance matrix was identified due to the time frame allowed to complete the project. In addition, covariate imputation was only conducted on one variable at the very last step to overcome missing imputations. In other words, attempts were made to fit models with the least amount of alteration to the existing database as possible.

[1] For each NCFO (i.e., 2000, 2006. 2008, etc …), the data are collected within that calendar year, but the data are collected for the previous calendar year.

Preliminary Analyses

Prior to fitting any multiple imputation (MI) models, a series of summary statistics and scatter plots were produced using SPSS version 12.0.1. The first step was to compute the bivariate correlations of the covariates to the dependent variable to be imputed (i.e., passenger boardings). As you can see in table 1, very few of the covariates were significantly correlated to the dependent variable. This seemed counter intuitive as one might expect the passenger capacity of the vessel and or other vessel characteristics to be related to the number of passenger boardings. A ferry operator is not going to invest money into a vessel that cannot be recouped by passenger fares. On the other hand, an operator will invest enough money into a vessel to be sure that passenger fares are not being lost at the dock. The only covariates that appear to be correlated with passenger boardings based on the initial correlations were the annual total number of vehicle boardings (Vehbrdgs; $r = .76$, $p < .001$), and the average number of vehicle boardings during peak traffic periods (Pk1veh; $r = .43$, $p = .002$, Pk2veh; $r = .34$, $p = .035$).

Table 1: Covariate Correlations with Passenger Boardings

Covariate	n	r	p-value	Covariate	n	r	p-value
Avtriptime	282	-.072	.225	Daysawk	215	.064	.353
Paxseas	285	.089	.133	Pk1pax	102	-.077	.441
Vehbrdgs	130	.763	<.001 **	Pk1veh	49	.425	.002 **
Vehseas	134	.120	.167	Pk2pax	82	-.071	.527
Segleng	285	-.045	.447	Pk2veh	38	.344	.035 *
Typspd	279	-.008	.889	Auto	285	.003	.960
Paxcap	274	.001	.983	Parking	285	.009	.875
Lanefeet	134	-.046	.595	Transbus	285	.006	.918
Horsepower	240	-.015	.820	Interbus	285	.030	.608
Selfprop	284	.028	.637	Litheavrail	285	.036	.547
Breadth	275	.042	.483	Amtrak	285	.034	.567
Length	274	.053	.380	Truck	285	.029	.624
Captons	152	-.009	.909	Frghtrail	285	-.024	.684
Nettons	220	-.034	.617	Metro	285	-.035	.552
Caryveh	285	-.038	.524	Ratereg	285	.053	.368
Caryfrt	52	.037	.794	Pbpown	285	.096	.104
Tripsaday	210	-.046	.508	Pbprop	285	.122	.039 *

* Significant at $\alpha = .05$; ** Significant at $\alpha = .01$.

Given the unexpected pattern of correlations among the covariates, scatter plots were produced for each covariate against passenger boardings. The goal of reviewing the scatter plots was to evaluate the data distributions for normality (when appropriate), and identification of outliers. In figure 1, the scatter plot for vehicle and passenger boardings, five operator segments were flagged as potential outliers. Great care must be taken in deciding whether a case is an outlier. The fact that a data point does not group closely with the rest is not sufficient evidence to conclude that it should be removed from the analysis. In this case however, there was a consistent pattern across covariate scatter plots that indicated that these operator segments may be fundamentally different from the rest. Further investigation into the individual cases revealed that these were all very large volume, state owned operator segments (i.e., Staten Island, Alaska Marine Highways, etc.).

While vehicle boardings during peak periods was initially shown to be a significant predictor of annual passenger boardings, a closer look at the scatter plots indicated that their predictive value may be dramatically diminished once outliers were removed (see Figure 2). The same was true for many other potential covariates (e.g., publicly operated ferries – see Figure 3). After reviewing the scatter plots and investigating specific cases, a decision was made to remove the five operator segments identified as outliers. Once removed, the correlations were rerun to evaluate the impact their removal had on the individual covariates relationship to passenger boardings.

Figure 1: Scatter Plot of Passenger and Vehicle Boardings

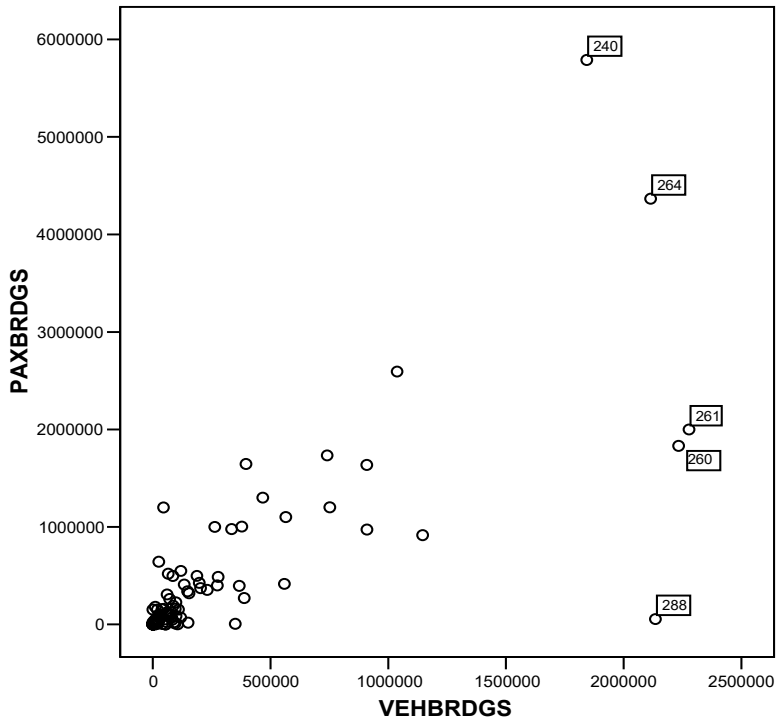


Figure 2: Scatter Plot of Passenger and Peak Vehicle Boardings

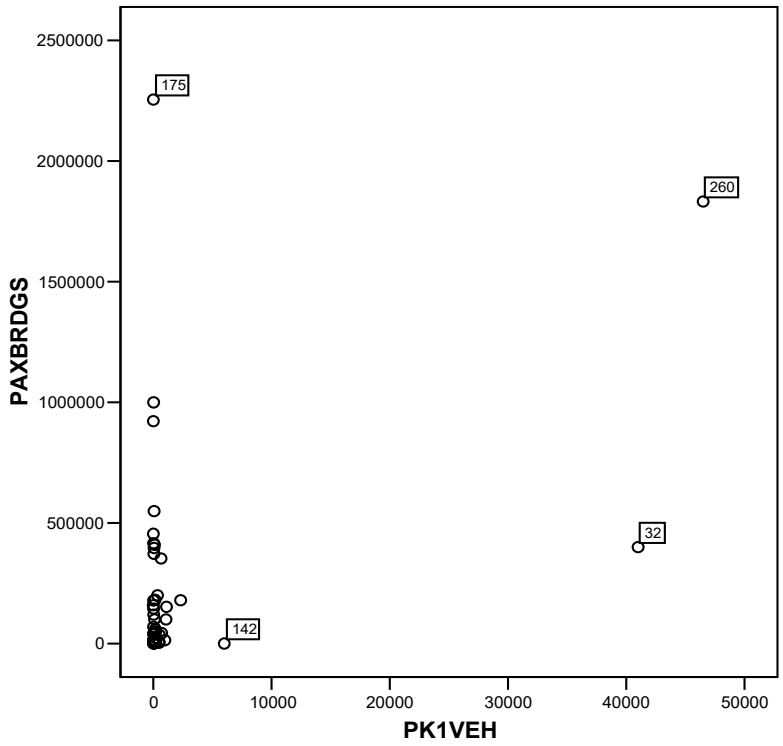
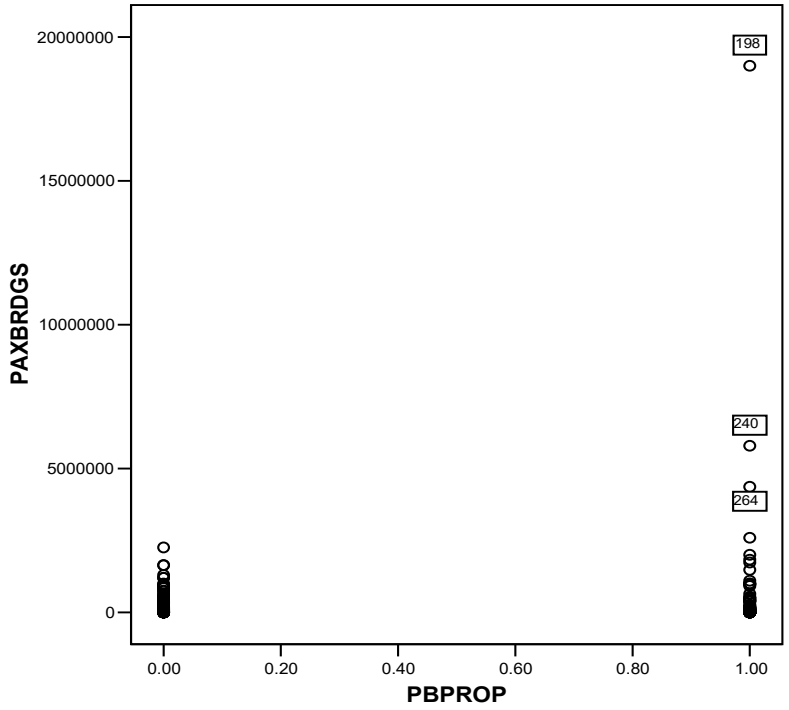


Figure 3: Scatter Plot of Passenger Boardings and Public/Private Operation



With the outliers removed, a more predictable pattern of correlations emerged among the covariates (see Table 2). The average trip time for a segment was negatively related to the total number of passenger boardings (Avtriptime; r = -151, p = .012). This makes sense as the longer it takes to make a trip between two ports, the fewer number of trips you can make. Another expected pattern that emerges is that the length of the season for ferrying passengers is positively related to the total number of passenger boardings (Paxseas; r = .186, p = .002). While the relationship between vehicle boardings and passenger totals is slightly decreased after removing the outliers (Vehbrdgs; r = -.703, p = <.001), the length of the season for carrying vessels is now a significant predictor (Vehseas; r = .198, p = .023).

Table 2: Covariate Correlations with Passenger Boardings (Outliers Removed)

Covariate	n	r	p-value	Covariate	n	r	p-value
Avtriptime	278	-.151	.012 *	Daysawk	211	.062	.372
Paxseas	281	.186	.002 **	Pk1pax	101	-.071	.480
Vehbrdgs	128	.703	<.001 **	Pk1veh	48	.604	<.001 **
Vehseas	132	.198	.023 *	Pk2pax	81	-.062	.582
Segleng	281	-.041	.495	Pk2veh	37	.506	.001 **
Typspd	276	.046	.499	Auto	281	-.005	.939
Paxcap	270	.130	.033 *	Parking	281	.001	.983
Lanefeet	131	-.028	.747	Transbus	281	.024	.691
Horsepower	237	.005	.938	Interbus	281	.056	.350
Selfprop	280	.049	.417	Litheavrail	281	.038	.528
Breadth	272	.008	.897	Amtrak	281	.030	.618
Length	271	.038	.538	Truck	281	.064	.289
Captons	151	.006	.943	Frghtrail	281	-.029	.627
Nettons	217	-.031	.652	Metro	281	-.046	.441

Caryveh	281	.062	.297	Ratereg	281	.015	.807
Caryfrt	51	.025	.863	Pbpown	281	.069	.475
Tripsaday	207	-.054	.443	Pbprop	281	.040	.247

* Significant at α = .05; ** Significant at α = .01.

Considered individually, peak vehicle boardings still appear to be correlated with total passenger boardings. Individual bivariate correlations however, may be deceptive when not taken in context with covariates of which they may be a derivative. In an attempt to get a better feel for the true structure of the relationships between the covariates and the outcome, a series of linear regression models were fit so that the relationships between the covariates and outcome could be considered simultaneously. In the initial model, passenger boardings were regressed on the covariates shown to be significant in the previous set of bivariate correlations. In this model (see Table 3), only the length of the passenger season (Paxseas; beta = .292, p = .045), and the total number of vehicle boardings (Vehbrdgs; beta = .581, p < .001) appear to be significant predictors of total passenger boardings. In this case, beta is the standardized slope coefficient, showing the relative impact of each covariate.

Table 3: Linear Regression Models Predicting Passenger Boardings

Model 1			R-Squared = .540		R-Squared Adj. = .530
Covariate	β	SE	Beta	t-value	p-value
Avtriptime	-142.038	85.963	-.067	-1.652	.099
Paxseas	10,825.660	5482.507	.292	1.975	.045 *
Vehbrdgs	.797	.065	.581	12.275	<.001 **
Vehseas	-1292.027	5896.484	-.034	-.219	.827
Paxcap	-3.802	40.373	-.004	-.094	.925
Pk1veh	9.094	15.685	.102	.580	.562
Pk2veh	-7.036	12.899	-.100	-.545	.586

* Significant at α = .05; ** Significant at α = .01.

In an attempt to further tease out the unique effects of the various covariates in predicting the number of annual passenger boardings for each operator segment, forward step-wise regression was used with the same covariates to see if a different solution would be produced (see Table 4). In this model, the length of the passenger season does not remain a significant predictor. The length of the vehicle season enters the model based on a p <.1 but does not appear to be significant.

Table 4: Forward Step-wise Regression Models Predicting Passenger Boardings

Model 2			R-Squared = .835		R-Squared Adj. = .817
Covariate	β	SE	Beta	t-value	p-value
Vehbrdgs	.790	.100	.823	7.906	<.001 **
Vehseas	9042.960	5026.378	.187	1.799	.089

* Significant at α = .05; ** Significant at α = .01. Probability to enter = .1.

After running the stepwise regression model, additional scatter plots were produced among vehicle boardings, vehicle season, passenger capacity and passenger season against passenger boardings as an additional inspection for outliers. This led to the identification of 6 more operator segments that were removed from the data file. From the original 355 operator segments, a total of 11 were removed leaving 344 operator segments to be used for the multiple imputation procedures. All of the operator segments removed was from large scale, state owned ferry operations. One critical thing to keep in mind here is that only cases were passenger boarding data exists are being removed from the data set. We cannot remove cases where this data are missing because we would not get an imputed value from the multiple imputation procedure if they weren't present in the final data file being used at that step. Fortunately, the fact that cases with missing passenger data are not represented in the scatter plots insulates us from making the mistake of removing

these cases based on the preliminary analysis.

Once these cases were removed, the stepwise regression model was rerun resulting in two significant predictors of passenger boardings (see Table 5). While the overall amount of explained variability appears to be reduced in this model (Adj. R-squared = .744 vs. .835), this reduction may due to the reduction in sample size and the fact that there is less variability to be predicted (i.e., removing extreme values truncates the variability of the dependent variable). The more important factor here is that we have a cleaner model for predicting the outcomes to be imputed.

Table 5: Linear Regression Models Predicting Passenger Boardings

Model 3		R-Squared = .749			R-Squared Adj. = .744
Covariate	β	SE	Beta	t-value	p-value
Vehbrdgs	1.032	.068	.764	15.091	<.001 **
Vehseas	9840.588	2393.562	.208	4.111	<.001 **

* Significant at α = .05; ** Significant at α = .01. Probability to enter = .1.

A final look at the individual correlations based on the reduced sample reveals four significantly correlated covariates (see Table 6). As one might expect, the passenger capacity of the vessel most often used and the length of the passenger season for the operator segment also appear to be significantly related to the number of annual passenger boardings. It is important to keep in mind however that none of the covariates evaluated thus far are free of missing data. While some covariates may have less missing data than others, the multiple imputation procedure requires at least one of the covariates (preferably all) to have no missing data. This issue will be addressed further in the next section.

Table 6: Covariate Correlations with Passenger Boardings (Final)

Covariate	**n**	**r**	**p-value**
Vehbrdgs	123	.811	<.001 **
Vehseas	127	.189	.033 *
Paxcap	264	.281	.001 **
Paxseas	274	.246	<.001 **

* Significant at α = .05; ** Significant at α = .01.

Multiple Imputation Models

Once the preliminary analyses were complete, the data file (n = 344) was imported into SAS (version 9.3) to run a series of multiple imputation models. In order to establish a rough baseline for comparison; the initial model included all covariates without specification (i.e., min or max values and rounding) for the covariates or dependent variable. Given the changes to the census questionnaire from previous years, and the short window of time for imputing the missing passenger data, no prior covariance matrix was used. As such, all models used the expectation-maximization (EM) method for estimating the prior covariance matrix. The Markov Chain Monte Carlo (MCMC) method was used for estimating missing values in all models due to fact that the pattern of missingness was non-monotonic. All other settings were also left to SAS defaults. Ten imputations were made for every missing operator segment.

The results of the initial model were less than stellar (see Table 7). The model failed to converge on an acceptable solution for either the covariance matrix or the imputed data set after the default of 200 iterations. Furthermore, the covariance matrices for both were also singular. A covariance matrix is singular when there is zero variability within a covariate or perfect correlation exists between two or more covariates. A second model was run with iterations increased to 500 and the criterion for resolution was relaxed to p = .05. The algorithm for the prior covariance matrix converged after 109 iterations. It was still singular and the covariance matrix for the imputed data failed to converge and was singular. The lack of specification for the dependent variable also resulted in negative passenger imputation values (see Min, Table 7). Unless scores of ferry passengers fall overboard on a regular basis, this result is implausible.

Table 7: Model 1 – All Covariates No Specification.

SAS Code

```
PROC MI data = work.ncfo SEED = 37851 NIMPUTE = 10 OUT = ncfomipass;
  MCMC CHAIN = multiple DIPLAYINIT INITIAL = em;
  VAR paxbrdgs amtrak auto avtriptime breadth captons caryfrt caryveh daysawk frghtrail horsepower
  interbus laneft length litheavrail metro nettons parking paxcap paxseas pbprop pbprown pk1pax pk1ve
  Pk2pax pk2veh ratereg segleng selfprop transbus tripsaday truck typspd vehbrdgs vehseas;
RUN;
```

Log

```
WARNING: The EM algorithm (MLE) fails to converge after 200 iterations.
WARNING: A covariance matrix computed in the EM process is singular.
WARNING: The EM algorithm (posterior mode) fails to converge after 200 iterations
WARNING: The initial covariance matrix for MCMC is singular.
WARNING: The posterior covariance matrix is singular.
```

Multiple Imputations of Passenger Boardings

Imputation	N	Mean	SE	Min	Max	Sum
1	344	185,356.10	18,135.65	-533791	2,000,000	63,762,497
2	344	186,752.59	19,323.98	-1204754	2,000,000	64,242,892
3	344	181,470.55	18,925.46	-1021384	2,000,000	62,425,869
4	344	173,706.94	19,028.62	-849604	2,000,000	59,755,187
5	344	186,524.70	19,227.07	-1049541	2,000,000	64,164,497
6	344	194,723.65	18,126.15	-678041	2,000,000	66,984,934
7	344	184,224.60	18,721.17	-728645	2,000,000	63,373,264
8	344	178,793.03	18,063.15	-728539	2,000,000	61,504,803
9	344	192,083.91	18,608.46	-583127	2,000,000	66,076,864
10	344	187,283.80	18,762.99	-822624	2,000,000	64,425,627
Total	3,440	185,091.99	5,905.69	-1,204,754	2,000,000	636,716,434

In an attempt to reduce the complexity of the model and thus reduce the problems associated with multicolinearity among covariates (and hence avoid singular covariance matrices), A third model was fit using only the regressors shown to be significant predictors under simultaneous inference in the preliminary analyses (i.e., annual vehicle boardings and vehicle boarding season length). This simplified model's prior covariance matrix converged after 20 iterations, and the imputed data matrix converged after 11 iterations (see Table 8). While there was little difference in the standard errors associated with the average of the imputations, but we still observed negative imputed passenger values. Given that we are ultimately interested in estimating the total number of passenger boardings for the calendar year 2007, imputing negative boarding values is again implausible and will have a major impact on these estimates.

Table 8: Model 3 – Two Covariates (no specification).

SAS Code

```
PROC MI data = work.ncfo SEED = 37851 NIMPUTE = 10 OUT = ncfomipass;
  MCMC CHAIN = multiple DISPLAYINIT INITIAL = em;
  VAR paxbrdgs vehbrdgs vehseas;
RUN;
```

Log

```
NOTE: The EM algorithm (MLE) converges in 20 iterations.
NOTE: The EM algorithm (posterior mode) converges in 11 iterations.
```

Multiple Imputations of Passenger Boardings

Imputation	N	Mean	SE	Min	Max	Sum
1	303	179,375.71	18,982.08	-434,514.99	2,000,000	54,350,839.25
2	303	187,891.91	19,211.47	-412,683.83	2,000,000	56,931,248.34

3	303	189,517.50	19,343.72	-417,116.35	2,000,000	57,423,803.23
4	303	186,929.41	19,315.83	-544,696.47	2,000,000	56,639,611.41
5	303	179,309.55	18,946.19	-301,384.48	2,000,000	54,330,792.19
6	303	183,571.12	19,238.67	-390,095.94	2,000,000	55,622,047.99
7	303	186,445.08	19,117.84	-447,116.73	2,000,000	56,492,858.27
8	303	185,150.75	19,048.60	-337,908.23	2,000,000	56,100,677.32
9	303	188,086.29	19,455.47	-694,845.60	2,000,000	56,990,144.70
10	303	185,071.72	19,183.31	-286,927.88	2,000,000	56,076,730.74
Total	3,030	185,134.90	6,058.09	-694,845.60	2,000,000	560,958,753.44

To further refine the imputation model, we added specifications for all variables included in the estimation. For passenger and vehicle boarding, the minimum value was set to 0, while there was no limit set to the maximum value. For the length of the vehicle boarding season the minimum was set to 1 month while the maximum was set to 12. With these changes to the model, we see that the matrices again converged after 20 and 11 iterations respectively with no errors (see Table 9). We also see that the minimum values among the imputed data appear within range, and that the mean and total passenger boardings for each set of imputations appear to be raised. Even though we have removed a number of the operator segments with the largest volume of passenger boardings, these totals still appear to be too low (see Giesbrecht, 2008).

Upon further inspection, it becomes clear that not all 344 operator segments are included in the imputed datasets. Each set only includes 303 observations. As previously mentioned, the multiple imputation procedure assumes at least one of the covariates has no missing data. With the MCMC method, the model attempts to estimate all missing values simultaneously. When all variables in the model are missing data for a given observation, there are no knowns by which to estimate the other missing values for that observation. In this instance there were 41 cases where all variables were missing data. This will result in reduced estimates of passenger boardings and increased standard errors due to the reduced sample size.

Table 9: Model 4 – Two Covariates (with specification).

SAS Code

```
PROC MI data = work.ncfo SEED = 37851 NIMPUTE = 10 OUT = ncfomipass
  MINIMUM = 0 0 1
  MAXIMUM = . . 12
  ROUND = 1;
  MCMC CHAIN = multiple DISPLAYINIT INITIAL = em;
  VAR paxbrdgs vehbrdgs vehseas;
RUN;
```

Log

NOTE: The EM algorithm (MLE) converges in 20 iterations.
NOTE: The EM algorithm (posterior mode) converges in 11 iterations.

Multiple Imputations of Passenger Boardings

Imputation	N	Mean	SE	Min	Max	Sum
1	303	191,889.51	18,763.09	2	2,000,000	58,142,521
2	303	197,284.32	18,908.17	2	2,000,000	59,777,148
3	303	191,827.54	18,750.88	2	2,000,000	58,123,745
4	303	194,358.26	18,914.09	2	2,000,000	58,890,553
5	303	199,018.72	18,930.59	2	2,000,000	60,302,673
6	303	192,588.82	18,764.24	2	2,000,000	58,354,413
7	303	196,819.93	18,795.78	2	2,000,000	59,636,439
8	303	195,099.49	18,869.86	2	2,000,000	59,115,147
9	303	195,373.09	18,937.17	2	2,000,000	59,198,045

10	303	196,235.86	18,972.01	2	2,000,000	59,459,465
Total	3,030	195,049.55	5,955.57	2	2,000,000	591,000,149

In a final effort to refine the model and overcome the problems associated with missing data among the covariates, the passenger capacity of the vessel was included into the imputation model. While passenger capacity was not shown to be a significant predictor of passenger boardings in the final linear model in the preliminary analyses, it did have a significant correlation with the number of annual passenger boardings within the data set currently being analyzed. It also had far less missing data than either of the two covariates currently included in the imputation model.

Before it could be included to resolve the issue of missing data among covariates, efforts were taken to generate data for missing values within the passenger capacity variable. When vessel characteristics where known, the length and breadth of the vessel were compared to other vessels of the same size to impute the passenger capacity of the vessel for a given operator segment. When the vessel characteristics were not known, the average passenger capacity of all vessels among operator segments with missing passenger boarding data was imputed. The final model included imputed missing values for all 344 observations (see Table 10), with increased passenger boarding estimates and reduced standard errors.

Table 10: Model 5 - Three Covariates (with specification).

SAS Code

```
PROC MI data = work.ncfo SEED = 37851 NIMPUTE = 10 OUT = ncfomipass
  MINIMUM = 0 0 1 0
  MAXIMUM = . . 12 .
  ROUND = 1;
  MCMC CHAIN = multiple DISPLAYINIT INITIAL = em;
  VAR paxbrdgs vehbrdgs vehseas paxcap;
RUN;
```

Log

```
NOTE: The EM algorithm (MLE) converges in 23 iterations.
NOTE: The EM algorithm (posterior mode) converges in 13 iterations.
NOTE: The data set WORK.NCFOMIPASS has 3440 observations and 37 variables.
```

Multiple Imputations of Passenger Boardings

Imputation	N	Mean	SE	Min	Max	Sum
1	344	212,565.42	17,324.04	2	2,000,000	73,122,503
2	344	216,702.85	17,621.49	2	2,000,000	74,545,780
3	344	217,445.39	17,437.51	2	2,000,000	7,4801,214
4	344	211,235.08	17,312.78	2	2,000,000	72,664,869
5	344	219,078.48	17,608.24	2	2,000,000	75,362,996
6	344	218,727.08	17,542.07	2	2,000,000	75,242,117
7	344	223,729.66	17,627.63	2	2,000,000	76,963,004
8	344	227,624.69	17,988.91	2	2,000,000	78,302,892
9	344	212,053.76	17,234.49	2	2,000,000	72,946,492
10	344	215,787.36	17,571.97	2	2,000,000	74,230,852
Total	3,440	217,494.98	5,536.26	2	2,000,000	748,182,719

As a test of the idea that blindly increasing the number of covariates in a multiple imputation model improves the fit of the model, one last attempt was made to fit a model with all covariates. In this model, the range and scale of each variable was specified with increased iterations and relaxed criteria for convergence (see Table 11). Again it appears that we have surpassed the methods ability to overcome issues associated with missing values and multicolinearity. Not only are the covariance matrices singular and the model fails to converge, but the procedure is halted because not all of the imputed values are within the specified range.

Table 11: Model 6 – All Covariates (with specification).

SAS Code

```
PROC MI data = work.ncfo SEED = 37851 NIMPUTE = 10 OUT = ncfomipass
  MINIMUM = 0 0 0 1 0 0 0 0 1 0 0 0 0 10 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 . 0 0 1 0 1 0 1
  MAXIMUM = . 1 1 . . . 1 1 7 1 . 1 . . 1 1 . 1 . 12 1 1 . . . . 1 . 1 1 . 1 60 . 12
  ROUND = 1 1 1 . . . 1 1 1 1 . 1 . . 1 1 . 1 . 1 1 1 . . . . 1 . 1 1 1 1 . . 1;
  MCMC CHAIN = multiple DISPLAYINIT INITIAL = em;
  EM MAXITER = 500;
  EM CONVERGE = .05;
  VAR paxbrdgs amtrak auto avtriptime breadth captons caryfrt caryveh daysawk frghtrail horsepower
  interbus laneft length litheavrail metro nettons parking paxcap paxseas pbprop pbprown pk1pax pk1ve
  Pk2pax pk2veh ratereg segleng selfprop transbus tripsaday truck typspd vehbrdgs vehseas;
RUN;
```

Log

```
NOTE: The EM algorithm (MLE) converges in 109 iterations.
WARNING: A covariance matrix computed in the EM process is singular.
WARNING: The EM algorithm (posterior mode) fails to converge after 200 iterations.
WARNING: The initial covariance matrix for MCMC is singular.
ERROR: An imputed variable value is not in the specified range after 100 tries.
WARNING: The data set WORK.NCFOMIPASS may be incomplete.  0 observations and 37 variables.
WARNING: Data set WORK.NCFOMIPASS was not replaced because this step was stopped.
```

Conclusions

For any modeling effort, multiple imputation or otherwise, a thorough preliminary analysis is key to getting a better understanding of the dataset you are working with and gives clues as to how to overcome issues when trying to fit the model. Failing to become familiar with the dataset prior to conducting more complex analyses is like flying an airplane while blindfolded. If an analyst is not familiar with the characteristics of each variable in the dataset and their relationships to each other, he/she has no way of making informed decisions as to how to change a model that does not fit or does not prove useful in predicting outcomes.

That having been said, great care must be taken at each step in the preliminary analysis so as not to create an "alternate reality". Just because an observed data point does not cluster with the rest, does not mean it is bad data. Any decision to remove a data point from an analysis should only be made with careful consideration of the ultimate goal of the analyses to be conducted and the other characteristics associated with that data point. Ultimately the analyst must be able to defend every decision to include or omit an observation. As can be seen in the analyses and decisions made above, that may be more easier said than done as modeling data are as much art as it is science. Often time, the analysts familiarity with the population of interest (i.e., his/her gut), plays as big a role in decisions along the way as do hard, cold calculations.

Within this set of analyses several decisions were made that gave rise to other analyses and subsequent decisions. At any step along the way, arguments can be made that other decisions should have been made or different analyses conducted. In reality, if deadlines didn't have to be met, the authors would have continued to chase down loose ends in an attempt to better understand the data before producing a final product. Unfortunately, or maybe fortunately, the deadline for this project forced the authors to make hard decisions on how to handle the data. With more time, several other avenues would have been exhausted.

Although not every avenue explored within this project was presented in this paper, the authors understand that those explored were not entirely exhaustive. Efforts were made to transform the non-normal distribution of passenger capacity but the transformed variable was not extensively modeled within the preliminary analyses and multiple imputation procedures. There may have been other non-normal variables that, once transformed, proved to be more accurate and more reliable predictors of annual passenger boardings. We also did not fully vet the extent and nature of multicolinearity among covariates. Ideally, a more exhaustive investigation of correlations among covariates would be used to remove redundant regressors. Finally, we did not explore the use of informative prior covariance matrices as a

means to reducing the standard errors associated with imputations. Doing so would have required efforts similar those of which we've touched the surface of here on the 2005 database.

With the enough time and resources a complete set of preliminary analyses would be conducted on the 2005 data to develop an informative prior covariance matrix rather than using EM estimations. Ideally we would be able to evaluate the differences in estimations as a result of having an informative prior. Additionally, more care would be taken to investigate every loose strand underlying the decisions made along the way toward proclaiming the final number of annual passenger boardings among ferry operations in calendar year 2007. For example, it may be that the removal of individual observations may have been avoided by fitting more complex models during the preliminary analyses. Hindsight aside, it is important to note that any cases omitted from the imputation model would need to be reintroduced to each imputed dataset prior to performing further statistical analysis.

References (lit revue pending)

http://support.sas.com/rnd/app/papers/miv802.pdf

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