Environmental Science Processes & Impacts

Accepted Manuscript

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Environmental Science
Processes & Impacts
Formerly Journal of Environmental Monitoring
www.rsc.org/jem
ROYAL SOCIETY OF CHEMISTRY

Evaluation of Imputation Methods for Microbial Surface Water Quality Studies

(Prepared for: Environmental Science: Processes & Impacts)

Chiping Nieh[†,*]

Samuel Dorevitch[†]

Li C. Liu[§]

Rachael M. Jones[†]

[†]Division of Environmental and Occupational Health Sciences, School of Public Health,
University of Illinois at Chicago, 2121 W. Taylor Street, Chicago, IL 60612-7260

[§]Division of Epidemiology and Biostatistics, School of Public Health, University of
Illinois at Chicago, 1603 W Taylor Street, Chicago, IL 60612-4394

ABSTRACT

Longitudinal studies of microbial water quality are subject to missing observations. This study evaluates multiple imputation (MI) against data deletion, mean or median imputation for replacing missing microbial water quality data. The specific context is data collected in Chicago Area Waterway System (2007 – 2009), where 45% of *Escherichia coli* and 53% of enterococci densities were missing owing to sample analysis deficiencies. Imputation methods were compared performing a simulation study using complete observations with introduced missing values and subsequently compared with the original data with missing observations. Coefficients for *E. coli* densities in linear regression models predicting somatic coliphages density show that MI introduces the least bias among other methods while controlling Type I error. Further exploration of utilizing different MI implementations is recommended to address the influence of missing percentage on MI performance and to explore sensitivity to the degree of violation of the missing completely at random assumption.

KEYWORDS Missing data; Multiple imputation; microbial water quality data; *Escherichia coli*

INTRODUCTION

Long-term studies of surface water quality can provide insight into environmental and ecosystem dynamics, and the determinants of water quality.[1,2] Sample collection, however, may periodically be interrupted in such studies owing to equipment and analysis failures, etc. These events result in missing values. Though a variety of statistical techniques are available for the analysis of data with missing values, there are circumstances, such as performing statistical modeling of health outcomes using water quality measures as a predictor, in which a complete dataset, with no missing values, is required. Additionally, a complete data set containing the initially planned sample size ensures the required statistical power.

Common known methods to deal with missing data include data omission (DO), arithmetic mean imputation (AMI), median imputation (MedI), regression imputation (RI), and multiple imputation (MI). The DO method excludes observations with any missing component. As a result, the sample size the statistical power of the study is reduced. DO is the most common method for dealing with missing data and is the default for many statistical software programs. The AMI and MedI methods replace all missing values with the same value, the arithmetic mean or median, respectively, of the observed data. These two methods prevent sample size reduction and have the advantage of not changing the sample mean or median of the variable, but variance is reduced by the imputation.[3] This misleading decrease in variance can (erroneously) improve the statistical significance of comparisons of means or other data analyses and lead to false conclusions. The DO, AMI and MedI methods do not consider relationships between variables in the imputation, which may be appropriate for univariate analyses, but could lose information in a multivariate context. RI is a technique in which missing values are estimated by a regression model developed by predicting the observed values from other variables in the data set. While RI utilizes relationships in multivariate data, the use of fitted values for imputation over identifies the relationships between variables. MI is a simulation-based method that creates *m* data sets with imputed values, that has been enabled by improvements in computer technology. Unlike the other methods described, MI maintains the variance of the original data set,[4] and considers both sampling variability and uncertainty introduced by missigness in the imputation of missing values. By creating *m* data sets, it enables the variation introduced by imputation to be compared across the *m* imputed data sets.[5]

In the context of microbial surface water quality, investigators have utilized a variety of methods to handle missing values. For example, in two multi-year studies, Whitman et al.[6] and Bezuidenhout et al.[7] omitted observations with missing values; excluding the season and month of missing *E. coli* density values, respectively, from reported microbial water quality trends. In contrast, Nevers et al.[8] replaced an unspecified number of *E. coli* density values missing throughout the study period with a value equal to the average of the three previous and three subsequent values. In other settings, missing water chemistry and hydrology values have been replaced by multiple imputation,[9,10] observed arithmetic mean, or median.[2,11]

The rationale investigators use to select a strategy for the management of missing environmental monitoring data is rarely reported, such that guidance is limited for new analysis problems. Studies outside of the field of environmental monitoring that compare multiple missing data management strategies[3,12-14] can be difficult to evaluate for inference to microbial water quality owing to their use of synthetic, non-multicollinear, continuous, or normally distributed data for method comparison. The objective of this study is to compare four strategies for the management of missing microbial surface water data – data omission, arithmetic mean, or median replacement, and multiple imputation – using environmental measurements of microbial density in freshwater. We hypothesize that multiple imputation preserves the data structure and produces less biased and more precise statistical inferences than the imputation and omission methods tested.

The data used were collected through the Chicago Health, Environmental Exposure, and Recreation Study (CHEERS). CHEERS was an observational cohort epidemiological study that characterized the risk of acute gastrointestinal illness among recreators performing secondary-contact water recreation. Water recreation took place on either the Chicago Area Waterways System (CAWS), an engineered system that receives 70-90% of its flow from wastewater treatment plants, or general use waters. The frequency of gastrointestinal illness was similar among recreators on the CAWS and on general use waters, but higher than that among persons who did not participate in water recreation.[15] During participant recruitment microbial water quality was measured with the intent of evaluating health risk as a function of exposure to water-borne microbes. However, between 08/01/2008 and 05/08/2009 deficiencies in analyses performed at a commercial laboratory led the research team to discard 963 of 2,155 (45%) *E. coli* and 1,121 of 2,155 (52%) enterococci density results from all locations. Absent these data, the exposure of many study participants cannot be determined for the purpose of exposure-response analyses. Imputation of these data would enable the assignment of exposure to all recreators.

Previous applications of MI to environmental data have found the method to effectively recover missing information,[9,10,16] which suggests that the method may recover missing information in the context of CHEERS. To our knowledge, however, the MI method has not been applied to microbial surface water quality. This application could pose a challenge for MI because the data exhibit high temporal and spatial variability, and the data in CHEERS have high rates of missingness. We used a simulation approach to test our hypothesis. Specifically, we defined a subset of complete data for which all variables were observed, introduced missing values using simulation, and applied the imputation and omission methods. We evaluated each method by comparing *i)* the distributions of microbe densities, and *ii)* linear regression model coefficients fitted after treatment of missing values. Subsequently, the methods were applied to the original data, to impute *E. coli* and enterococci values missing due to laboratory deficiencies.

MATERIALS AND METHODS

Data. We considered water quality measurements in the North Branch System and Cal-Sag Channel of the CAWS, which were collected seasonally (n= 1,206): 8/2007-10/2007, 4/2008-10/2008, and 4/2009-7/2009. Study locations included: (North to South) Bridge Street, Skokie Rowing Center, Lincoln Avenue, River Park, Clark Park, and North Avenue in the North Branch System (Figure 1), and (East to West) Beaubien Woods, Riverdale Marina, Alsip, and Worth in the Cal-Sag Channel (Figure 2).

Data for this study was limited to these locations in the CAWS for a primary reason that indicator microbes were present well above the method detection limits and protozoan pathogens were detected frequently, relative to study locations in Lake Michigan. Samples were collected at these ten locations throughout the three-year study period, and so capture temporal and spatial variability of microbial water quality data.

Microbial water quality was described by the densities of: *E. coli* (colony forming units [CFU]/100mL), enterococci (CFU/100mL), F+ coliphages (plaque forming units [PFU]/100ML), somatic coliphages (PFU/100mL), *Giardia* cysts (#/10L) and *Cryptosporidium* oocysts (#/10L). Sample collection and analytical techniques were described elsewhere.[15] Briefly, the four indicator microbes were measured every 2 hours during participant recruitment (1-4 times per day), while the protozoan pathogens were measured every 6 hours (1-2 times per day). Chemical and physical measures of water quality were measured when the indicator microbes were measured, and included: dissolved oxygen (DO, mg/L), pH, conductivity (mmho/cm), water temperature (°C), and turbidity (NTU). Rainfall was described by the magnitude, duration, intensity and time since the last rainfall event, where rainfall events were distinguished by at least 6 hours without rainfall.[17] Combined sewer overflow (CSO) events were described by the magnitude, duration, intensity and time since the last event anywhere in the North Branch or Cal-Sag Channel, where events were distinguished by at least 1 hour without CSO.[17]

Missing Data. During CHEERS study, external quality control (QC) was performed using blinded spiked samples. Spiking involved the subdivision of a water sample into two samples. A known concentration of microbes was added into the first sample and the second sample was not manipulated. Recovery was then calculated by dividing the microbe concentration detected in the spiked sample by the sum of the expected concentration added to the spiked sample and the microbe concentration detected in the non-spiked sample. During the period 08/01/2008-05/08/2009, laboratory performance for *E. coli* and enterococci density was poor, as indicated by (1) unusually high variability in the recovery of the two microbes, and (2) failures to detect these bacteria in waters samples collected downstream of water treatment plants where the microbes were typically numerous. Laboratory internal quality control measures did not indicate a problem with sample handling and analysis, but blinded spiked matrix samples frequently yielded zero percent recovery for indicator microbes. *E. coli* and enterococci data quality returned to acceptable levels after a different laboratory began analyzing samples in May, 2009. During this period, data quality for coliphages and protozoan pathogen analyses, which were conducted at a different laboratory, remained excellent. The insufficient laboratory performance resulted in discarding all *E. coli* and enterococci data analyzed during the time period. Of the 1,206 *E. coli* and 1,206 enterococci measurements in the CAWS (North Branch and Cal-Sag Channel), 45% and 53% were excluded, respectively.

The mechanisms by which data were missing have been classified by Rubin[18] as: *i*) missing completely at random (MCAR), in which the probability of a value being missing is not related to both the observed and unobserved data, *ii*) missing at random (MAR), in which the probability of a value being missing is related to the value of observed data, but not to its own value, and *iii*) missing not at random (MNAR), in which the probability of a value being missing is related to its unobserved value. The event of MCAR means that the missing data were a random subset of the original data, such that the true multivariate distribution was preserved in the non-missing values.[19] The mechanism of missingness influenced the selection of omission and imputation strategies.

The data studied herein were missing owing to laboratory error. The problematic samples were collected at multiple locations and days, and are expected to span the range of water quality and weather conditions observed during the entire study. The laboratory was blinded to the location of sample collection and anticipated water quality. Thus, there was no reason to suspect the probability of poor laboratory performance (e.g., the event of a missing value) to be associated with microbe density in the sample or with other observed values, suggesting that these values are MCAR. In addition, consistent with a MCAR pattern, the distributions of the $\log_{10}$ densities of the other microorganisms, along with chemical and physical measures of water quality, collected on days when valid *E. coli* results were reported by the laboratory to be qualitatively similar to the distributions measured during the period of unacceptable data quality, even though two-sample Kolmogorov-Smirnov test indicated that majority of them do not have the same distributions (Table 1). Enterococci results were not presented in Table 1. As described in the following paragraph, the quality of imputation methods were compared by evaluating inferences drawn from the imputed data regarding somatic coliphages density. Because $\log_{10}$ *E. coli* densities were associated with $\log_{10}$ somatic coliphages density, meaning a significant parameter estimate of *E. coli* in a multivariate regression model predicting somatic coliphages density, while $\log_{10}$ enterococci densities were not, analyses were limited to *E. coli* imputation.

Simulation Study. To enable evaluation of imputation methods against real values, a *complete* data set was created in which no *E. coli* density values were missing (n = 622). The approach was to introduce a MCAR pattern into the complete data by random deletion, and impute the deleted values using each of the four methods. Simulation included 1,000 replications of the following steps: *i*) randomly delete 45% of *E. coli* density values, equal to the percentage of missing data in the original data set, *ii*) fill in missing values using one of the imputation methods of interest, *iii*) fit a linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \sum_{j=2}^{p} \beta_j x_{ji} + \varepsilon_i \qquad \text{Equation 1}$$

where $\varepsilon_i \sim i.d.d.N(0, \sigma^2)$, y_i is the $\log_{10}$ somatic coliphages density, x_{1i} is the $\log_{10}$ *E. coli* density, and x_{ji} are other dependent variables; and iv) retrieve parameter estimates of $\log_{10}$ *E. coli*, β_1. The retrieved parameter estimates were used to compare imputation methods.

Imputation Methods. We considered four imputation methods: *i*) data omission, DO, *ii*) arithmetic mean imputation, AMI, *iii*) median imputation, MedI, and *iv*) multiple imputation, MI. DO was implemented by excluding all observations associated with each missing *E. coli* density. AMI was implemented by replacing all missing values of *E. coli* density by the arithmetic mean value of *E. coli* densities remaining after deletion from the *complete* data set. MedI was implemented by replacing all missing values of *E. coli* by the median value of *E. coli* densities. MI was implemented utilizing the Markov Chain Monte Carlo (MCMC) imputation mechanism, which accommodates an arbitrary missing data pattern. The Proc MI statement in SAS was used to generate *m* = 5 imputed data sets. The Proc MI statement has two major imputation algorithms, *i*) propensity score with the approximate Bayesian bootstrapping technique,[20] and *ii*) regression model with data augmentation (DA) technique.[21] Due to the presence of a non-monotone missingness, DA algorithm was utilized.[22,23] The DA algorithm involves repetition of an imputation step (I-step) and a posterior step (P-step). In the I-step, a covariance matrix is generated from the observed data and specified regression model, and missing values are imputed with the addition of random noise. A new covariance matrix is generated using the imputed data, and the P-step is initiated. In the P-step, new elements of the covariance matrix are randomly selected from a posterior distribution based on the imputed data in I-step. The I-step is initiated, and the cycle repeats until the covariance matrices converge. The algorithm is implemented *m* times to generate *m* sets of imputed data.

Collins et al.[24] addressed the question of what variables should be included in the imputation model by comparing parameter estimates obtained using various numbers of variables in imputation and found that the more variables in the model (e.g., a richer model), the better imputation results. Therefore, we added as many variables as possible in the imputation model, including: date, location (dummy variable), enterococci density, *Giardia* cyst density, *Cryptosporidium* oocyst density, somatic coliphages density, F+ coliphages density, sampling hour, pH, dissolved oxygen, conductivity, turbidity, water temperature, solar radiation, time since last rain, and magnitude of last CSO.

We considered, but did not implement a time-series averaging approach, which is a type of arithmetic mean imputation in which a missing value is imputed with the mean of temporally adjacent observed values, such as was employed Nevers et al.[8] Unlike the dataset used by Nevers et al.,[8] in which daily measurements were made at a fixed set of locations, in CHEERS, locations were typically sampled on weekends, and the frequency of sampling a given location was based on patterns of recreational use of surface waters. As a result locations were rarely sampled more than two consecutive days. For many locations, the sampling frequency was less than weekly. Thus the six temporally adjacent data point approach used by Nevers et al.[8] would likely span dates that were weeks, and potentially months, apart. This approach was judged inappropriate for the context of the present study.

Method Comparison. The methods were first compared based on the distribution of $\log_{10}$ *E. coli* after imputation or omission relative to the real data. Since 1,000 replications were simulated, the distribution characteristics (e.g., mean and variance) were calculated for each replicate and averaged for comparison to the real data. Previous work suggests that all methods should preserve the central tendency, while AMI and MedI are expected to reduce variance in the distribution. Replicating this result provides a general verification of the integrity of the analyses.

The primary evaluation of the methods, however, is based on statistical inferences, specifically the regression coefficient for the variable $\log_{10}$ *E. coli* density, denoted β_1. The regression model is specified in Equation 1. Initially, the goal was to use microbial indicator to predict pathogen densities, because pathogens cause adverse health outcomes among water users. However, in initial analyses, the magnitude of correlations among pathogens and *E. coli* or enterococci densities were weak. Therefore, we used $\log_{10}$ somatic coliphages density as the dependent variable.

The specific independent variables included in the regression model were selected by backwards-step variable selection ($\alpha = 0.05$): sample date, location (dummy variable), $\log_{10}$ *E. coli* density, F+ coliphages $\log_{10}$ density, dissolved oxygen, and turbidity. During model selection, multicollinearity was evaluated by the variance inflation factor (VIF), and found to be acceptable, with VIF < 10.[25] For multiply imputed data, the regression coefficients, β_1, fitted to the $m = 5$ imputations were pooled using Rubin's rule.[5] This pooling adjusts for the within-imputation and between-imputation variances.[26] The evaluation of MI is based on statistical inferences, like the pooled estimate for β_1, instead of the individual filled-in values in each imputed data set because the method introduces uncertainty and variability in each estimate for each missing value.

Performance of the imputation and omission methods across 1,000 replications were summarized using the metrics:[27] real parameter β_1 , estimated parameter $\bar{\beta}_1$, standardized bias (%), coverage rate (%), mean confidence interval width, and root-mean-square error (RMSE). The real parameter β_1 was the parameter estimate of *E. coli* yielded using the *complete* dataset to fit the regression in predicting the densities of somatic coliphages. Estimated parameter, $\bar{\beta}_1$ was the average of 1,000 parameter estimates, $\hat{\beta}_1^k$ where k ={1, 2, …, 1000}, of *E. coli* yielded through each simulation replication. Standardized bias was calculated

$$\frac{|\beta_1-\overline{\beta_1}|}{SD} * 100\% \qquad \text{Equation 2}$$

where *SD* was the standard deviation of the $\hat{\beta}_1^k$. The width of confidence interval was calculated for each replication and then the average width across 1,000 replications was reported. The coverage rate was the percent of simulations when β_1 fell within the 95% confidence interval of $\hat{\beta}_1^k$. According to Demirtas,[27] an approximate 95% coverage rate suggested that the rates of Type I error was well controlled. The root-mean-square error (RMSE) was calculated across replications as:

$$RMSE = \sqrt{\frac{1}{1000}\sum_{k=1}^{1000}\left(\beta_1-\hat{\beta}_1^k\right)^2}. \qquad \text{Equation 3}$$

RESULTS

Distribution Comparisons. As expected, AMI and MedI yielded smaller variance and MI yielded larger variance than the real data (Table 2). The distribution of the $\log_{10}$ *E. coli* density after data omission was, of the four methods, most similar to the real distribution, as indicated by the mean, median, 5th and 95th percentiles, and the standard deviation. Figure 3 shows scatter plots of the data for a randomly selected replication of the simulation relative to the complete data. The lack of realism introduced by the AMI and MedI methods are indicated by the high frequency of a single value (the mean or median). The oval-shaped cloud shows the magnitude of variability and uncertainty introduced by the MI method, which is expected to vary between the multiply imputed data sets.

Linear Model Inferences. The coefficient, β_1, for $\log_{10}$ *E. coli* density in predicting somatic colipages estimated by the different methods in the context of simulation are summarized in Table 3. The magnitude of bias in coefficients fitted with data treated by MI was smaller than observed with the other methods. Additionally, DO and MI had better coverage rates, 95.6% and 95.3% respectively, than AMI and MedI, 81.4% and 80.7%. A coverage rate of 95% indicates correct control of Type I error.[27] Overall, the better performance of MI in comparison to other methods was indicated by the higher coverage rate, and the smaller bias, mean CI width, and RMSE.

Original Dataset. When the four methods were applied to the *original* data set with *E. coli* values missing due to laboratory problems (Table 4), all methods yielded similar estimates of the mean $\log_{10}$ *E. coli* density, judging by the fact that all the standard errors overlap one another. As expected, the AMI and MedI methods yielded the smaller estimates of the standard errors than the DO and MI methods. Due to the introduction of random noise in MI, it is not surprising that microbe densities imputed using this technique had higher variance relative to the other methods tested, including DO.

The linear model coefficients for $\log_{10}$ *E. coli* density estimated with the *original* data after imputation or omission are summarized in Table 5. All coefficient estimates were statistically significantly different from zero. AMI and MedI resulted in equal estimates of β_1. The MI method gave the highest estimate for β_1 and smallest estimate of the standard error. Data treated by the DO method estimate for β_1 fell between the MI estimate and the AMI/MedI estimates, but had the largest standard error, which makes sense owing to the smaller sample remaining after data deletion.

DISCUSSION

Our objective was to evaluate the performance of three imputation methods – multiple imputation, arithmetic mean imputation and median imputation – and data omission for analysis of microbial surface water quality. Missing values occur frequently in long-term water quality studies.[2,6-11] To our knowledge, this is the first study to systematically compare methods for filling in missing microbial density values in surface water data. Our motivation for exploring methods to fill in missing values was specific to CHEERS, in which data missing due to poor laboratory performance prevented linkage between the environmental hazard of microbial surface water quality density and individuals conducting water recreation for the evaluation of health risk for epidemiologic analysis. However, the problem of missing microbe densities in surface water quality studies is ubiquitous.

Microbe densities in surface water exhibit high temporal and spatial variability, and it was not clear that MI could recovery missing information in this context given the high frequency of missing values, 45%. By using a simulation approach with a complete data set we were able to verify that MI can recover missing information to yield similar statistical inference to the complete data. A weakness of our simulation study from the public health perspective, however, was the evaluation of relationship between two indicator microbes – *E. coli* and somatic coliphages – rather than relationships between indicator microbes and protozoan pathogens, which can adversely impact human health. However, the weak relationship between the indicator bacteria and protozoan pathogens in these data may be unique, and the focus on two indicator microbes does not invalidate the MI evaluation.

We found that MI creates a relatively higher variance in the data after imputation (Table 2), but produces less biased regression coefficients relative to the other imputation and omission methods tested (Table 3). This finding concurs with observations in psychology and epidemiology.[3,12,13] An implication of our finding is that data omission[6,7] or imputation[5] methods used in previous studies of microbial surface water quality could have reported biased parameter estimates of *E. coli* densities. One expects that using an imputation approach in which some variety is introduced into the imputed value (e.g., imputing the mean of three previous and three subsequent values,[8] a season-specific mean, or in multi-location studies, a location-specific mean) could improve statistical inference relative to using an overall mean or median value, as was done in this study. However, Olinsky et al.[28] concluded that even though the degree of underestimation of variance using regression imputation is less than using mean imputation, MI still generated the less biased statistical inferences than RI.

A strength of this study was that the large number of samples collected in CHEERS enabled the creation of an artificial, complete dataset with which to test the imputation and omission strategies for the management of missing data relative to results from analysis of the real values. Another strength of this study was that the data were highly variable owing to collection at many locations, at different times of day, in three seasons, and over three years. To ensure an informative joint distribution was available for the MI method, environmental variables pertaining to solar radiation, microbial inputs into the CAWS (rain and combined sewer overflow), and chemical and physical measures of water quality were included in the imputation model. Our finding that MI effectively recovers missing in this context, particularly in light of the high rate of missing values (45% of *E. coli* values) provides important evidence that the MI method may be robust for environmental applications.

In our study, data were limited to those collected at CAWS. Other water bodies may have lower microbial densities, including a substantial proportion of samples below detection limits. Imputation of values below detection limits has been widely addressed in water quality literature, and can be performed within a multiple imputation framework. Additionally, the data used in this study are incomplete time series data, which is unique in comparing to any long-term time series microbial water quality data. Therefore, it is important for future studies to evaluate MI performance for time-series data, and specifically whether it is necessary to explicitly model the time-series in the multiple imputation model.[29]

Furthermore, future work should explore MI performance using different MI implementations and at other sites prior to recommending the method's wider application for microbial surface water quality. Analysis with these data can be extended to address the influence of the percentage of missing values on MI performance. In long-term microbial water quality studies, frequently, data are missing during thunderstorms or adverse weather conditions, and weather conditions often influence microbial densities: This event would result in a MNAR pattern. We hypothesize that inclusion of data about weather conditions is essential for the correct specification of the imputation model this context, and suggest that future analysis evaluate the sensitive of MI to MNAR patterns.[30] In the context of CHEERS, these results support the use of MI to fill-in missing values, thereby avoiding a substantial loss of human health data in analyses of water quality as a predictor of illness.

CONCLUSION

This study has demonstrated the use of MI can restore the preferred sample size and provide statistical inferences with less bias than other traditional imputation methods. Our findings suggest that MI is a useful tool to recover information that is lost due to unpredictable events. Given that our study considered data MCAR, this recommendation implies that such missing information also follows a MCAR pattern.

Table 1. Distributions of microbes (mean, standard deviation, SD, and median of log$_{10}$ densities) along with chemical and physical measures of water quality are similar when *E.coli* results were valid and invalid.

	Subset with Valid *E. coli* Results				Subset with Invalid *E. coli* Results			
Measure (Unit)	No.	Mean	SD	Median	No.	Mean	SD	Median
Giardia (log$_{10}$ cysts/10L)	194	0.741	1.205	0.875	242	0.714	1.135	0.916
Cryptosporidium (log$_{10}$ oocysts/10L)	194	-0.85	0.967	-1.602	242	-0.34	1.169	-0.301*
Somatic coliphages (log$_{10}$ PFU/100mL)	642	2.264	0.93	2.415	415	2.433	0.928	1.279*
F+ coliphages (log$_{10}$ PFU/100mL)	642	0.934	0.845	0.903	415	1.199	0.909	2.602*
Dissolved Oxygen (mg/L)	414	7	2.015	6.62	256	7.483	2.027	7.395*
Conductivity (mmho/cm)	397	724.4	362.3	766	249	668.3	344.5	617*
Turbidity (NTU)	421	17.07	13.93	13.73	237	16.21	9.999	13.85
Solar radiation (W/m^2)	646	4.711	3.153	4.285	542	3.731	2.885	2.99*
Hours since last CSO (hour)	662	452.6	508	257.14	544	560.3	740.9	255.27*
Hours since last rain (hour)	662	61.19	66.94	39	544	59.43	66.21	36*

*Two-Sample Kolmogorov-Smirnov Test: $p < 0.05$.

Table 2. Distributions of $\log_{10}$ *E. coli* densities across simulation replications after data omission (DO), arithmetic mean imputation (AMI), median imputation (MedI), and multiple imputation (MI1-MI5, and the average of the five sets (Ave)), compared with the real, complete distributions.

				Multiple Imputation						
Statistic	DO	AMI	MedI	MI1	MI2	MI3	MI4	MI5	Ave	Real
N	364	662	662	662	662	662	662	662	662	662
Mean	2.730	2.730	2.751	2.733	2.725	2.732	2.731	2.733	2.731	2.730
Median	2.774	2.730	2.776	2.746	2.739	2.745	2.746	2.746	2.745	2.778
SD	1.006	0.746	0.745	1.022	1.026	1.023	1.027	1.020	1.023	1.006
5th percentile	1.173	1.488	1.490	1.147	1.130	1.144	1.135	1.146	1.140	1.204
95th percentile	4.226	3.961	3.963	4.300	4.300	4.303	4.304	4.300	4.302	4.230

Table 3. Simulation results for different imputation methods including data omission (DO), arithmetic mean imputation (AMI), median imputation (MedI), and multiple imputation (MI). Estimated β_1 represents the coefficient for $\log_{10}$ *E. coli* density in predicting somatic coliphages estimated by each imputation method.

Imputation Method	DO	AMI	MedI	MI
Real β_1	0.128	0.128	0.128	0.128
Estimated β_1	0.149	0.081	0.079	0.134
Standardized Bias (%)	40.9%	149.9%	162.2%	18.2%
Coverage Rate (%)	95.6%	81.4%	80.7%	95.3%
Mean CI Width	0.213	0.153	0.153	0.157
RMSE	0.057	0.057	0.058	0.035

Table 4. Distributions of *E. coli* $\log_{10}$ densities from the *original* data after imputation or omission. Imputation methods include data omission (DO), arithmetic mean imputation (AMI), median imputation (MedI), and multiple imputation (MI1-MI5, and the average (Ave)).

				Multiple Imputation					
Variable	DO	AMI	MedI	MI1	MI2	MI3	MI4	MI5	Ave
No.	662	1,206	1,206	1,206	1,206	1,206	1,206	1,206	1,206
Mean	2.730	2.730	2.751	2.794	2.866	2.784	2.845	2.843	2.826
SD	1.006	0.745	0.745	1.121	1.140	1.056	1.168	1.063	1.109
5th percentile	1.204	1.477	1.477	1.183	1.204	1.130	1.247	1.204	1.194
95th percentile	4.230	3.954	3.954	4.505	4.524	4.347	4.423	4.428	4.445

Table 5. Estimate for coefficient of $\log_{10}$ *E. coli* density ($\beta_1$in Equation 1) using imputed (n = 1,006) or omitted (n = 729) values from the original data. Imputation methods include multiple imputation (MI), arithmetic mean imputation (AMI), median imputation (MedI), and data omission (DO).

	MI	AMI	MedI	DO
Estimated β_1	0.165*	0.113*	0.113*	0.128*
Standard Error	0.018	0.026	0.026	0.028

* p < 0.05.

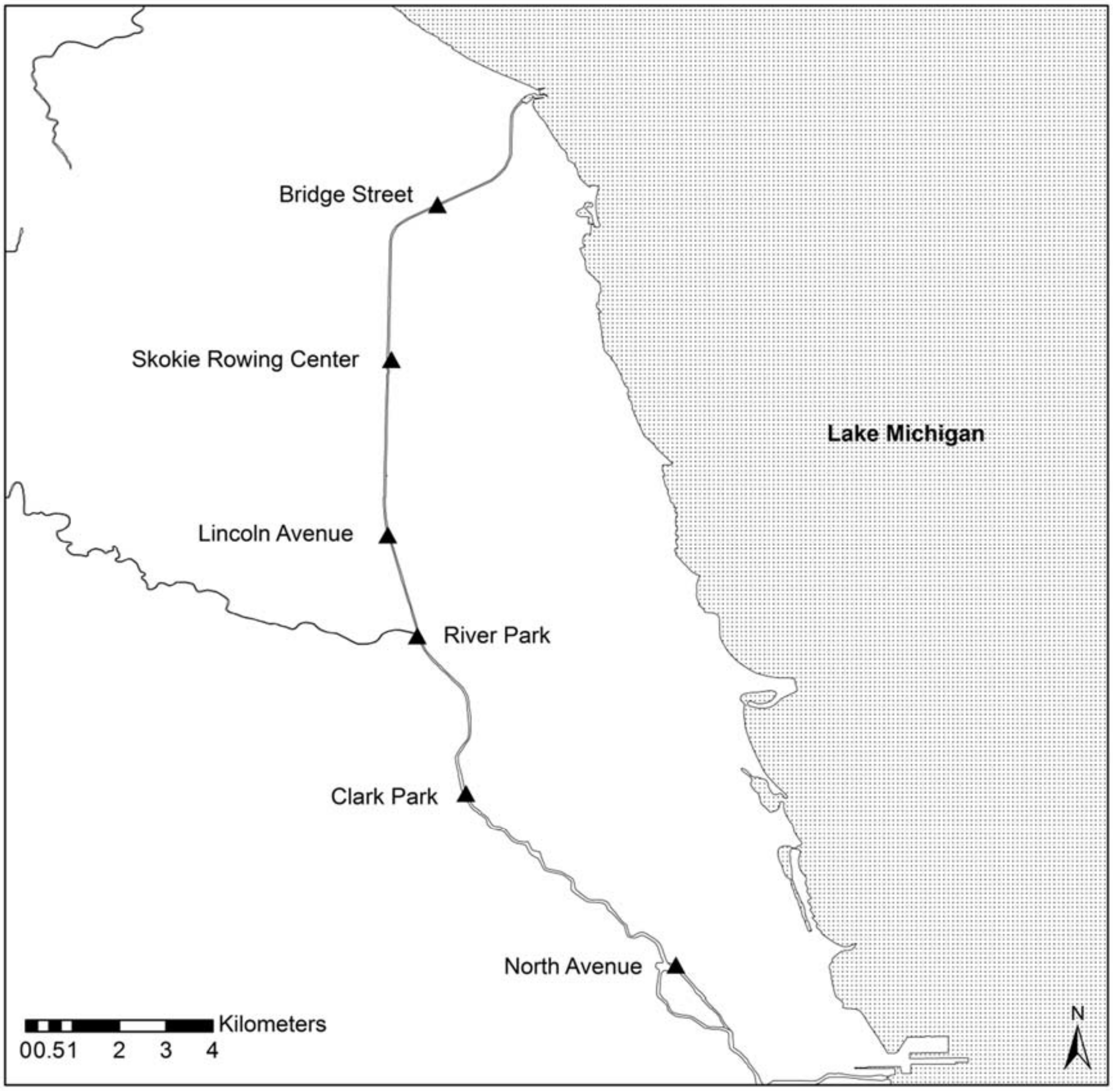

Figure 1. Map of study locations in the North Branch System. Map area includes downtown Chicago to Evanston, IL, the first suburb to the north of City of Chicago.

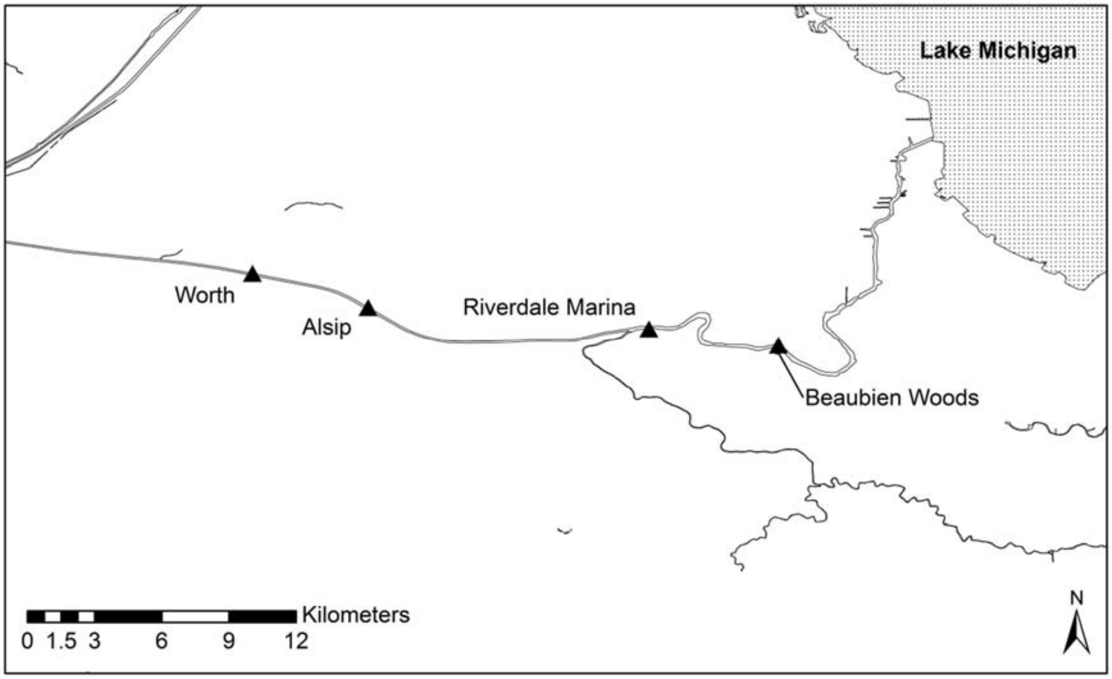

Figure 2. Map of study locations in the Cal-Sag Channel. The Cal-Sag Channel is in the far southern neighborhoods of the City of Chicago and adjacent suburbs, including Lemont, IL, Crestwood, IL, Blue Island, IL, and ends in Beaubien Woods Forest Preserve.

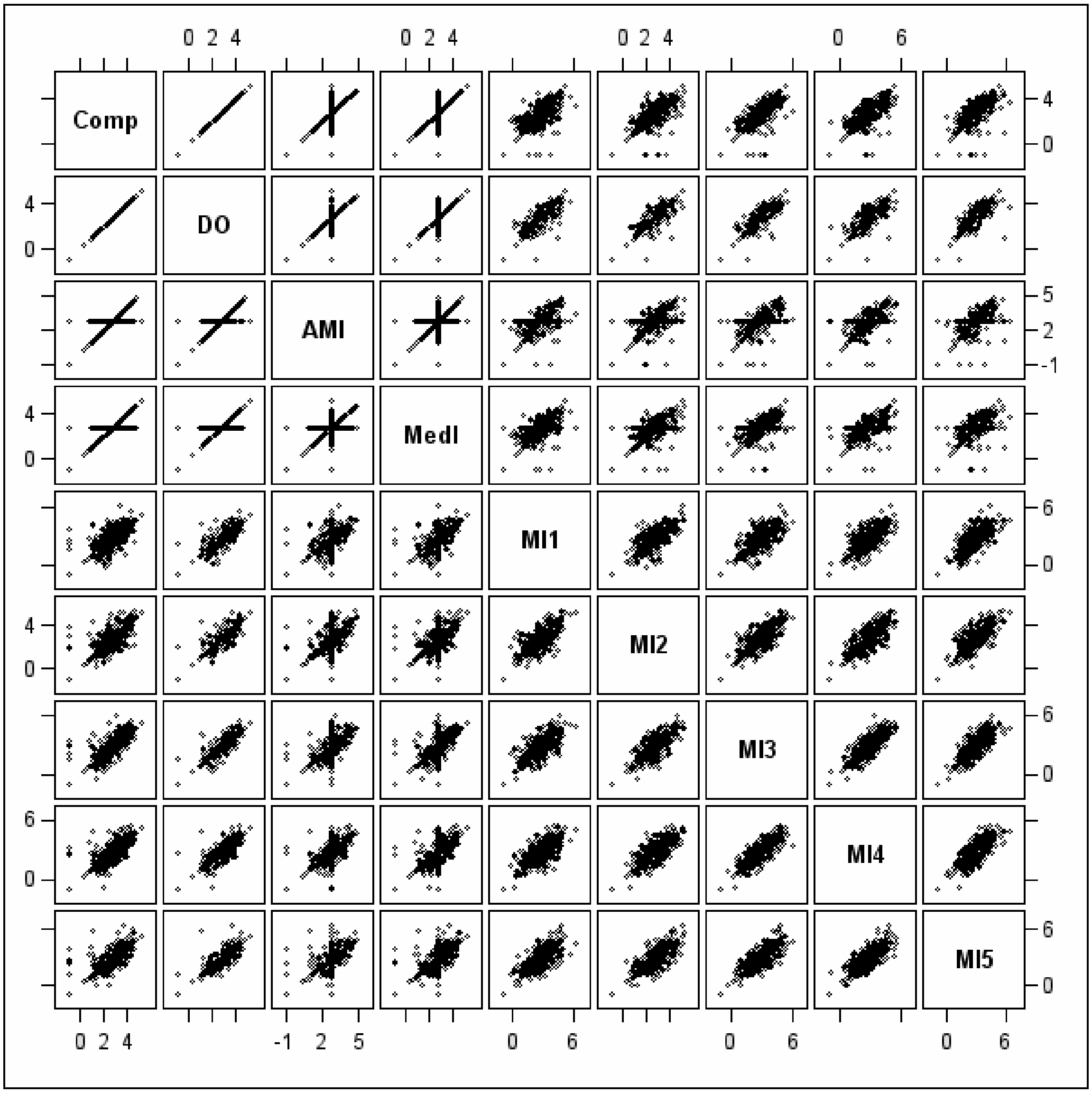

Figure 3. Scatter plot matrix of log10 E. coli densities imputed using data omission (DO), arithmetic mean imputation (AMI), median imputation (MedI), and multiple imputataion (MI1-MI5) methods in comparison to the complete data set of no missing log10 E. coli data (Comp).

AUTHOR INFORMATION

Corresponding Author

*Chiping Nieh.

Email: cnieh@health-ra.com

Address: Health Research and Analysis, 5410 Edson Lane, Suite 320, Rockville, MD 20852-3107.

Present Addresses

Health Research and Analysis, 5410 Edson Lane, Suite 320, Rockville, MD 20852-3107

Funding Sources

This research and researchers were supported by the Metropolitan Water Reclamation District of Greater Chicago.

ACKNOWLEDGMENT

The authors thank the Metropolitan Water Reclamation District of Greater Chicago for funding this project and kindly providing relevant data.

ABBREVIATIONS

AMI, arithmetic mean imputation; CAWS, Chicago Area Waterways System; CFU, colony-forming unit; CHEERS, Chicago Health, Environmental Exposure, and Recreation Study; CSO, combined sewer outfall; DA, data augmentation; DO, data omission; MAR, missing at random; MCAR, missing completely at random; MedI, median imputation; MI, multiple imputation; MNAR, missing not at random; PFU, plaque-forming unit; U.S. EPA, U.S. Environmental Protection Agency; VIF, variance inflation factor.

REFERENCES

(1) *Modeling water quality trend in long term time series*. SAS Institute Inc., Proceedings of the Thirty-One Annual SAS® Users Group International Conference: Reno, NV, www2.sas.com/proceedings/sugi31/205-31.pdf.

(2) *Long-term water quality trends in Virginia's Waterways*; VWRRC Special Report No. SR11-1998; Virginia Water Resources Research Center: Blacksburg, Virginia, 1998; vwrrc.vt.edu/pdfs/specialreports/sr111998.pdf.

(3) Enders, C. K. A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic Medicine*. **2006**, *68*, 427-436.

(4) Little, R. J.; Rubin, D. B. Statistical Analysis with Missing Data. Hoboken, NJ,

Wiley, 2 edition, 2002.

(5) Rubin, D. B. Multiple Imputation of Nonresponse in Surveys; Wiley: New York, USA, 1987.

(6) Whitman, R. L.; Nevers, M. B. Summer *E. coli* patterns and responses along 23 Chicago beaches. *Environ. Sc. Technol*. **2008**, *42*(24), 9217–9224; DOI 10.1021/es8019758.

(7) Bezuidenhout, C. C.; Mthembu, N.; Puckree, T.; Lin, J. Microbiological evaluation of the Mhlathuze River, KwaZulu-Natal (RSA). *Water SA*. **2002**, *28*(3), 281-286.

(8) Nevers, M. B.; Whitman, R. L. Efficacy of monitoring and empirical predictive modeling at improving public health protection at Chicago beaches. *Water Research*. **2011**, *45*, 1659–1668; DOI 10.1016/j.waters.2010.12.010.

(9) Hui, D.; Wan, S.; Su, B.; Katul, G.; Monson, R.; Luo, Y. Gap-filing missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agricultural and Forest Meteorology*. **2004**, *121*(1-2), 93-111; DOI 10.1016/S0168-1923(03)00158-8.

(10) Mercer, T. G.; Frostick, L. E.; Walmsley, A. D. Recovering incomplete data using Statistical Multiple Imputation (SMI): A case study in environmental chemistry. *Talanta*. **2011**, *85*(5), 2599-2604.

(11) Simeonov, V.; Stratis, J. A.; Samara, C.; Zachariadis, G.; Voutsa, D.; Anthemidis, A.; Sofoniou, M.; Kouimtzis, Th. Assessment of the surface water quality in Northern Greece. *Water Research*. **2003**, *37*, 4119-4124.

(12) Startori, N.; Salvan, A.; Thomaseth, K. Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. *Computational Statistics & Data Analysis*. **2005**, *49*(3), 937-953; DOI 10.1016/j.csda.2004.06.013.

(13) Zhou, X.; Eckert, G.; Tierney, W. Multiple imputation in public health research. *Statist. Med*. **2001**, *20*(9-10), 15-30; DOI 10.1002/sim.689.

(14) Burns, R. A.; Butterworth, P.; Kiely, K. M.; Bielak, A. A.M.; Luszcz, M. A.; Mitchell, P.; Christensen, H.; Sanden, C. V.; Anstey, K. J. Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. *J. Clin. Epidemiol*. **2010**, *64*(7), 787-793; DOI 10.1016/S0168-1923(03)00158-8.

(15) Dorevitch, S.; Doi, M.; Hsu, F.; Lin, K.; Roberts, J. D.; Liu, L. C.; Gladding, R.; Vannoy, E.; Li, H.; Javor, M.; Scheff, P. A. A comparison of rapid and conventional measures of indicator bacteria as predictors of waterborne protozoan pathogen presence and density. *Journal of Environmental Monitoring.* **2011**, 13, 2427-2435; DOI 10.1039/c1em10379b.

(16) Jones, R. M., Stayner, L. T., Demirtas, H. Multiple Imputation for Assessment of Exposures to Drinking Water Contaminants: Evaluation with the Atrazine Monitoring Program. Environmental Research. (Pending)

(17) Jones, R.; Liu, L.; Dorevitch, S. Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: Data-driven methods for variable selection. *Environ. Monit. Assess.* **2012**, 1-12; DOI 10.1007/s10661-012-2716-8.

(18) Rubin, D. B. Inference and missing data. *Biometrika*, **1976**, *63,* 581–592.

(19) Schafer, J. L. Multiple imputation: a primer. *Stat. Methods Med. Res.* **1999**, *8*(1), 3–15.

(20) Rosenbaum, P.R.; Rubin,D.B. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983, 70, 41–55.

(21) Schafer, J. L. Analysis of Incomplete Multivariate Data. Chapman & Hall, London, 1997.

(22) Allison, P. D. Multiple miputation for missing data: a cautionary tale. Sociological Methods and Research. 2000, 28, 301-309.

(23) Yuan, Y. C. Multiple Imputation for Missing Data: Concepts and New Development. SAS Institute Inc., 1700 Rockville Pike, Suite 600, Rockville, MD 20852, 1.0 edition.

(24) Collins, L. M.; Schafer, J. L.; Kam, C.-H. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 2001, 6, 330–351.

(25) Kutner, M. A.; Nachtsheim, C.; Neter, J. Applied Linear Regression Models, 4th, ed.; McGraw-Hill, Irwin, 2004.

(26) Rubin, D. B. Multiple imputation after 18+ years. Journal of the American Statistical Association. 1996, 91(434), 473-489.

(27) Demirtas, H. Simulation driven inferences for multiply imputed longitudinal datasets. Statistica Neerlandica. 2004, 58(4), 466-482.

(28) Olinsky, A.; Chen, S.; Harlow, L. The comparative efficacy of imputation methods for missing data in structural equation modeling. Europe Journal of Operational Research. 2003, 151, 53-79.

(29) Hopke, P.; Liu, C.; Rubin, D. Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic. Biometrics. 2001. 57; 22-33.

(30) Héraud-Bousquet, V.; Larsen, C.; Carpenter, J.; Desenclos, J.; Strat, Y.L. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. BMC Medical Research Methodology. 2012, 12(1), 73; DOI: 10.1186/1471-2288-12-73.

Longitudinal studies of microbial water quality are subject to missing observations. Though a variety of statistical techniques are available for the analysis of data with missing values, there are circumstances in which complete data are required. This study has evaluated multiple imputation against data deletion, mean imputation, and median imputation for filling in missing microbial water quality data. The results have demonstrated the use of multiple imputation can restore the preferred sample size and provide statistical inferences with less bias than other traditional imputation methods in filling in missing microbial water quality data. Additionally, our findings suggest the possible use of multiple imputation to design a less costly longitudinal water quality study by planning sample collection to support data imputation.

A comparison of imputation techniques for handling missing values in microbial surface water quality data.