Published in final edited form as:
Sci Total Environ. 2020 August 15; 730: 139140. doi:10.1016/j.scitotenv.2020.139140.

Imputation methods for addressing missing data in short-term monitoring of air pollutants

Steven J. Hadeed[a,*], **Mary Kay O'Rourke**[a], **Jefferey L. Burgess**[a], **Robin B. Harris**[a], **Robert A. Canales**[b]

[a]The Mel and Enid Zuckerman College of Public Health, The University of Arizona, 1295 N. Martin Ave, Tucson, AZ 85724, USA

[b]Interdisciplinary Program in Applied Mathematics, The University of Arizona, 617 N. Santa Rita, Tucson, AZ 85721, USA

1. Introduction

Missing data are a problem frequently encountered in many fields of research, especially in environmental health sciences and occupational health research. Monitoring of environmental contaminants is a critical part of exposure sciences research and public health practice. Often, environmental monitors are used for compliance purposes by government entities, or for research applications. Environmental health researchers rely on the use of monitors to quantify concentrations of contaminants in the environment, and relate those concentrations with potential exposures and health effects.

Fine particulate matter ($PM_{2.5}$) is widely studied and is associated with significant adverse health effects. Over 7 million deaths per year are attributable to ambient and household air pollutants, including $PM_{2.5}$ (WHO, 2014). Monitoring $PM_{2.5}$ can occur across various spatial and temporal scales, ranging from fixed ambient monitors that run continuously as part of an air monitoring network, to citizen science projects using low-cost equipment, and household area and personal air monitors that may run for hours or days in a community or occupational setting. Regardless of the sampling approach, missing data may occur for a number of reasons when monitoring environmental contaminants. Understanding the pattern of missing data is important for guiding imputation methods that yield reliable estimates.

[*]Corresponding Author: Steven J. Hadeed, The Mel and Enid Zuckerman College of Public Health, The University of Arizona, 1295 N. Martin Ave, Tucson, AZ 85724, USA, hadeed@email.arizona.edu*(Steven Hadeed)*.

Credit Author Statement
Steven J. Hadeed: Data Curation, Writing - Original Draft, Writing - Review & Editing, Validation, Methodology, Formal analysis. **Mary Kay O'Rourke:** Writing - Review & Editing, Project administration, Supervision, Funding acquisition. **Jefferey L. Burgess:** Writing - Review & Editing, Project administration, Supervision, Funding acquisition. **Robin B. Harris:** Writing - Review & Editing, Project administration, Supervision, Funding acquisition. **Robert A. Canales:** Conceptualization, Methodology, Software, Validation, Writing - Review & Editing, Visualization

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Declaration of interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Missing at random (MAR) is frequently encountered in environmental health sciences studies. Under MAR, it may be possible to explain why the data are missing, since missing data are related to other data that are observed (Gomez-Carracedo et al., 2014; Little & Rubin, 2019; Quinteros et al., 2019). Under MAR, missing data can be estimated from other observed predictor variables. Gomez-Carracedo et al., (2014) argue that when air monitoring data are missing due to electrical power failure, the missingness follows the MAR pattern. Data that are missing not at random (MNAR) occur when the probability of an observation being missing is related to unobserved values (Donders et al., 2006; Lavarkas, 2008; Little & Rubin, 2002; McPherson et al., 2015). In the context of air monitoring, if a monitor shuts down due to a malfunction, such as high filter loading or extreme temperatures, then it may be considered MNAR.

Generally, the most widely used method for imputing missing data is unconditional mean imputation (Donders et al., 2006; Junger & De Leon, 2015; Junninen et al., 2004; Quinteros et al., 2019). However depending on the duration and type of missing data, this method of imputation may yield different results (Junger & De Leon, 2015). For instance, under MAR, high variance in regression coefficient estimates may result from mean imputation, whereas underestimated, but more reliable estimates of variance are obtained under Missing Completely at Random (MCAR) (Junger & De Leon, 2015). Median imputation is another simple method often appropriate for highly skewed data, and may yield better results compared to mean imputation (Junger & De Leon, 2015; Miettinen, 1985).

Univariate time-series imputation is another class of methods used in air pollution studies that accounts for the time series characteristics of real-time monitoring data (Mortiz et al., 2015). One common method is last observation carried forward (LOCF), which bridges data together by filling in gaps of missing data with the last observed value (Engles & Diehr, 2003; Plaia & Bondi, 2006). Hourly mean method is another approach for imputing missing hourly concentrations for a single fixed air monitoring site. This method uses observed hourly concentrations recorded at the same monitor over extended periods of time, often months or a year. Observed hourly averages collected at the same monitoring site are used to impute hours when the same monitor may be missing data (Li et al., 1999; Plaia & Bondi, 2006).

Multivariate time-series imputations are more complex methods that use predictor variables between observations to impute missing values. Some widely used methods include: regression imputations and predictive mean matching (PMM) (Rubin, 1986; Little, 1988), row mean method (RMM) (Engles & Diehr, 2003), multiple imputation chained equations (MICE) (Rubin, 1988), expectation-maximization (E-M) models (Dempster, 1977), random imputation (Moritz, 2015), and study specific imputations.

Many proposed imputation methods are specific to fixed central site air monitoring data, where periods of missingness are short, and are both preceded and followed by long periods of observed continuous readings (Junger & De Leon, 2015; Plaia & Bondi, 2006; Quinteros et al., 2019). Very little, if any, attention has been given to imputing missing real-time monitoring data of air pollutants on short time scales (<24 hours). Household indoor monitoring of air pollutants, specifically $PM_{2.5}$, is readily employed in developing nations, where solid fuel use for heating and cooking is prevalent. Often, these resource-limited areas lack the electrical power source to run modern air monitoring equipment, and instead, equipment must be operated on battery power. Depending on the sampling duration, the reliance on batteries may be insufficient to achieve the desired operating time, and may result in missing or incomplete data. For instance, if the study goal is to measure concentration of household $PM_{2.5}$ for 24-hours using a real-time monitor, battery or equipment failure may yield partial data with consecutive hours of missing data (e.g., 8 or 12 hours). This prevents the investigator from obtaining the desired 24-hour measurements and creates a unique pattern of missing data.

Missing data will also affect any summary statistics computed from real-time measurements. Examples of such statistics include peak hourly concentrations or 15-minute short-term exposure limits (STEL), often important in assessing acute health effects for specific pollutants. In epidemiological studies, pollutant concentrations are often reported in relationship to their potential health effects by averaging minute or hourly concentrations over a 24-hour period to yield a daily concentration. Premature shutdown of monitors due to equipment failure or battery power loss is a frequently encountered limitation in environmental field studies that is rarely acknowledged, or accounted for, at the analysis phase. However, when more than 25% of data is missing, daily average pollutant concentrations cannot be reliably computed (Plaia & Bondi, 2006).

Often when dealing with missing or incomplete data, analysts may be tempted to exclude these observations (case-wise deletion). Depending on the sample size and the degree of missingness, this may be appropriate. However, in studies with a small sample size, such as household and community-based studies, this may introduce bias and loss of power, and may not be a viable approach (Donders et al., 2006). Some researchers may decide to ignore this problem entirely, and they may be tempted to call an 8-hour concentration a 24-hour concentration, while others may exclude readings below a specific sampling duration threshold. However, relatively few environmental and health-based studies explicitly mention an inclusion or exclusion criterion in their analysis of air pollution data. Failing to recognize incomplete or missing data in community and household environmental health studies may lead to biased measurements and inconsistent findings across study environments.

Existing methods for imputing missing real-time environmental monitoring data over short time periods have yet to be explored. This paper examines methods for imputing consecutive periods of missing household air monitoring data based on various degrees of missing data when operating on a 24-hour time scale. Our goal is to provide guidance that is easy to implement and can be applicable to other study settings.

2. Methods

2.1. Data Source:

Data used in this evaluation included household outdoor concentration of $PM_{2.5}$ measured at 1-minute intervals using a pDR-1500 (ThermoFisher), yielding 1440 1-minute concentrations over 24-hours from 20 households from a rural Northern Arizona community that participated in a field study of indoor-outdoor air quality. Monitors were powered by 4 AA batteries at a flow rate of 1.51 L/min.

2.2. Type of Missing Data:

During the field study, incomplete data arose from premature shutdown of the pDR-1500, either from battery power loss or equipment failure occurring at some point during the 24-hour monitoring period. Outdoor monitors shut down more frequently during winter months, perhaps due to extreme temperatures or relative humidity beyond the operating range specified by the manufacturer. If monitor shutdown was a function of temperature or humidity, the data would be MNAR. However, exploring the conditions proximal to the shutdown of these monitors indicates the ambient conditions were within range of the manufacturers operating environment, suggesting MNAR was not present. Since MNAR was unlikely, we proceeded under the assumption our data was MAR, since data loss was due to battery power loss.

2.3. Pattern of Missing and Validation Dataset

In the field, once a monitor failed and shutdown, it stayed powered-off for the remainder of the sampling period. No attempt was made to restart or replace the batteries of the monitor, resulting in consecutive periods of missing data. To assess the accuracy of various imputation methods, a validation data set was developed that included the 20 households with complete 24-hour data. To recreate patterns of missingness encountered in the field, missing data were artificially created in these complete 24-hour samples. Starting at a designated time point, consecutive periods of missingness were created at four levels: 20% (288 minutes), 40% (576 minutes), 60% (864 minutes), and 80% (1152 minutes). This approach facilitated comparison of imputed concentrations with observed concentrations at each household.

2.4. Imputation Methods

Univariate, univariate time-series, and multivariate time-series methods were used to impute missing data at each of the four levels of missingness. Univariate methods were those methods using partially-observed data within each household to impute the remaining missing values, without considering the time-series nature of the data. Univariate imputation methods included mean, median, and random imputation to impute 1-minute values to yield complete 24-hours (1,440 minutes) of data at each home. For example, mean and median imputation used the mean and median, respectively, of the partially-observed data from within each household to impute missing 1-minute concentrations within the same household. Random imputation replaced missing concentrations with values that were sampled, with replacement, from the partially-observed data within each household. One thousand iterations were performed, and 1-minute averages of the 1000 iterations were used to impute missing concentrations.

Univariate time-series methods considered some time-series characteristics of the partially-observed data within each household to impute missing values. The LOCF method replaced missing data points with the value of the last observed concentration. Imputation via first-order Markov chains assumed a concentration at any time point was dependent only on the previous value and used a transitional probability matrix to generate future values based on the last observed concentration (Canales, 2004). Similar to the random imputation, 1,000 iterations were performed for each household and 1-minute averages were used for imputation. Finally, Kalman filters were used to fit autoregressive integrated moving average (ARIMA) models to predict missing values based on trends of previously observed measures (Moritz et al., 2015). From here on, we define univariate methods as methods that also include univariate time-series methods.

Multivariate time-series imputation was another class of methods that accounted for the time-series nature of the data, but a set of observed predictor variables was used to impute missing concentrations. Here information between households was used to predict missing data within each home. Categorical predictors used for multivariate imputation included geographic location (3-categories) and heating fuel type (3-categories). Homes were geographically grouped into three districts under the assumption that homes in the same district were spatially representative of one another. Based on questionnaire data, homes were assigned one of three indoor heating fuel types – electric-gas, coal-wood, or a combination. Fuel types were potential sources of outdoor $PM_{2.5}$ during the winter season when these fuels are used primarily for indoor heating. Additional continuous predictors included ambient temperature and relative humidity recorded from a co-located monitor logged at 1-minute intervals.

RMM is a multivariate time-series method that can be used to impute missing concentrations for a single monitoring site using observed concentrations recorded at nearby surrounding monitoring stations (Engles & Diehr, 2003; Plaia & Bondi, 2006). Time-matched mean concentrations from surrounding monitors were used to impute missing values (homes located 0.25 to 20 miles apart). For example, 9:00 AM mean concentrations from observed sites were used to impute missing concentration at another single monitoring site that was missing data at 9:00 AM. We also performed two variants of RMM: one by grouping homes geographically within a 2.5 mile radius (RMM-L), and another by grouping homes by heating fuel type (RMM-F). We were unable to group homes by both location and fuel type due to small sample size.

PMM is a form of hot-deck imputation that imputes missing values based on linear regression coefficients from observed values. For this method, missing values were selected randomly from a set of donors that were matched on an observed set of predictors. PMM was preferred over simple regression imputation because imputed values are drawn from a range of observed values, eliminating the possibility of imputing unrealistic values, such as negative concentrations (Van Buuren, 2018). Several variants of PMM were constructed using different combinations of predictor variables (i.e., ambient temperature, ambient humidity, fuel type, geographic location). Ten imputations and 5-nearest neighbors were used for each PMM imputation (Van Buuren, 2018), and 1-minute averages across the 10 iterations were used to impute missing 1-minute concentrations.

All methods were implemented in the R programming (R Core Team, 2017) language using the imputeTS (Moritz et al., 2015), mice (Buuren & Groothuis-Oudshoorn, 2010), and markovchain (Spedicato et al., 2016) packages.

2.5. Error Metrics:

Several metrics were used to evaluate the performance of imputation methods by comparing observed and imputed concentrations across four levels of missingness (i.e., 20%, 40%, 60%, 80% missing). Health based studies are often concerned with a daily average concentration that can be associated with a health effect. Absolute bias (AB) and percent absolute error in means (PAEM) were two metrics used to evaluate the errors between observed and imputed 24-hour mean concentrations within each household. The AB calculates the absolute difference between observed $(\ddot{x})$ and predicted 24-hour mean concentration $(\hat{x})$.

$$AB = \left| \ddot{x} - \hat{x} \right|$$

Similarly, PAEM measures the percent difference between the observed 24-hour mean concentration $(\ddot{x})$ and imputed 24-hour mean concentrations $(\hat{x})$ and is an easily interpretable metric.

$$PAEM = \left| \frac{\ddot{x} - \hat{x}}{\ddot{x}} \right| \cdot 100$$

Values of AB and PAEM equal to zero indicate no difference between observed and estimated mean values.

Time-series analysts may often be concerned with examining the temporal patterns of real-time data over a finer time scale, such as minutes or seconds. When missing data are encountered, imputation methods that yield reliable minute-by-minute or second-by-second estimates are desired. Within each household, the minute-by-minute differences in observed and imputed concentrations were evaluated using three metrics.

The coefficient of determination (R^2) is commonly used as a goodness of fit metric for evaluating models. R^2 was calculated by squaring the correlation coefficient between two variables and describes the variance accounted for between the observed and the predicted concentrations (Quinteros et al., 2019).

$$R^2 = \left(\frac{\sum_{i=1}^{n} (x_i - \hat{x}) * (\dot{x}_i - \ddot{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \hat{x})^2} * \sqrt{\sum_{i=1}^{n} (\dot{x}_i - \ddot{x})^2}} \right)^2$$

Where x_i and $\dot{x}_i$ are the ith observation for the imputed and observed datasets, and $\hat{x}$ and $\ddot{x}$ are the means for the imputed and observed datasets (Quinteros et al., 2019). Although widely used, R^2 is limited in its ability to account for size differences between observed and imputed values (Junninen et al., 2004; Willmot et al., 1985). Despite this limitation, R^2 was used as a comparable metric across studies.

Root mean square error (RMSE) is another metric for determining the error between imputed and actual values and is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \dot{x}_i)^2}$$

Although utilized in several studies comparing imputation methods (Junninen et al., 2004; Moritz et al., 2015; Moritz & Bartz-Beielstein, 2017; Quinteros et al., 2019) this metric may not be appropriate for data with large differences (Junger & De Leon, 2015; Moritz et al., 2015).

Mean absolute error (MAE) on the other hand, is a common metric for determining errors between imputed and observed values that is less affected by large differences in data (Junger & De Leon, 2015; Moritz et al., 2015).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid x_i - \dot{x}_i \mid$$

3. Results

A total of 20 household with complete 24-hour (1440-minute readings) monitoring data for $PM_{2.5}$ were used in this analysis. Imputation methods were implemented for each level of missingness and performance metrics were used to evaluate the accuracy of estimates by comparing observed and imputed concentrations within each household (Table 1).

Differences between observed and imputed 24-hour mean concentrations were evaluated using AB and PAEM (Table 1). For all 20 homes, mean, random, and Markov imputation methods consistently resulted in the lowest AB and PAEM across all levels of missingness. At 20% missing, these three methods yielded means that differed from observed means by roughly 2.0 μg/m^3. Differences between observed and imputed means at 40% missing were around 2.5 μg/m^3, differences ranged from 3.3-4.1 μg/m^3 at 60% missing, and imputed means differed from observed means by less than 7.0 μg/m^3 at 80% missing. Kalman filters and median imputation performed moderately well at 20% and 40% missingness, however AB and PAEM increased at 60-80% missing. Multivariate methods of PMM and RMM resulted in the highest AB and PAEM, whose performances decreased substantially as missingness increased.

Correlation between 1-minute observed and 1-minute imputed concentrations was assessed using R^2. High R^2 values indicated a high level of correlation between observed and predicted values. Relatively higher R^2 values were observed for mean, median, LOCF, Kalman filters, random, and Markov imputation, with mean, random and Markov methods yielding the greatest R^2 (Table 1). At 20% missing, R^2 values for these three methods were around 0.65, and at 40% missingness R^2 values were about 0.58. R^2 decreased substantially to around 0.35 and 0.11 at 60% and 80% missingness, respectively. Multivariate methods of RMM and PMM had the lowest R^2 values for all levels of missingness.

The RMSE metrics showed good performance for mean, median, random, Markov and Kalman filters. Much like other error metrics, RMSE increased as the level of missingness increased (Table 1). Kalman filters and median were high performing at 20% and 40% missing, however at 60-80% missing, Markov, random and mean imputation yielded the lowest RMSE values.

In comparison to RMSE, a metric less sensitive to extremes in the predicted values is MAE, where the lower the value, the better the performance. At 20-60% missing, median and Kalman filters performed best and yielded the lowest MAE values of 2.45 and 2.46 at 20% missing, 4.10 and 4.30 at 40% missing, and 7.46 and 7.72 at 60% missing, respectively. LOCF performed better than anticipated at 20-40% missing, but performance deteriorated as duration of missingness increased, which would be expected based on the nature of this method. Markov performed moderately well at 60% missing, however at 80% missing Markov, median, and mean imputation had the lowest MAE values (Table 1).

The worst performing methods were the two models that incorporated geographic and household fuel characteristics, RMM and PMM. They consistently yielded high errors across all levels of missingness, and in some instances resulted in MAE and RMSE values 2-3 times higher than their univariate counterparts (Table 1).

Based on our results, we list the top three imputation methods by performance metrics for each level of missingness (Table 2). Markov, random, and mean imputation provided the best estimates of 24-hour mean concentrations of $PM_{2.5}$ across all levels of missingness. When evaluating error metrics minute-by-minute, Kalman filters, median, and Markov methods performed well at low levels of missingness (20-40%). However, at higher levels of missingness (60-80%), Markov, random, median and mean imputation performed best on average (Table 2). Nonetheless, across all metrics, Markov seems to be the better performing imputation method for use with $PM_{2.5}$ data from these rural households.

4. Discussion

Missing data are often encountered when performing short-term monitoring of air pollutants with real-time monitors, especially in resource-limited areas. We explored approaches for handling missing data in this context. Our results, across several metrics, show univariate methods that impute missing values based on incomplete data observed within households performed best. Markov, random, and mean imputations were the best performing methods that yielded 24-hour mean estimates with the lowest error and highest R^2 values. Minute-by-minute imputation methods had mixed performance by percent missing, but Markov appears to be the most promising approach. On a case-by-case basis, Kalman filter imputation performed exceptionally well in data with strong trends, and may be a viable option for imputing time-series data of this nature.

The multivariate methods RMM and PMM were expected to perform well because these methods impute missing values from observed concentrations at households sharing similar characteristics. Surprisingly however, these methods performed poorly across all levels of missingness. One possible explanation of the superior performance of univariate imputation and the low performance of multivariate imputation may be due to households being significantly different from one another. These differences may not be captured by the predictor variables used in our analysis. For example, homes located in similar geographic locations, ambient conditions, and heating fuel types may differ in outdoor concentration by other unaccounted factors. This may include idling vehicles and human activities that generate or reduce PM concentrations at some households and not others. That is, households may have been so different from one another that the predictor variables used in multivariate imputation were unable to account for these differences. This would explain why univariate methods using partial data observed within each household yielded better estimates of PM concentrations, compared to multivariate methods using characteristics between homes.

Relatively good performance of mean imputation, even at high levels of missingness, was an unexpected finding since this method can yield biased estimates under MAR and tends to be discouraged (Quinteros et al., 2019). One reason for the success of mean imputation may be that the partially-observed data within each household was fairly representative of concentrations throughout the rest of the day.

Random and Markov imputation performed exceptionally well at high and low levels of missingness. The success of these two methods may be attributed to the high number of iterations used to generate 1-minute concentrations that were then used to impute missing values within households. Both methods utilized partial data observed within each home to predict concentrations missing within the same household, which may be effective for homes or monitoring stations located in areas that are completely different from one another.

Markov imputation was a novel approach implemented in this analysis. The probabilistic nature of 1st order Markov chains to impute values based on a logical order from the previous time step may explain this method's success. Compared to the random method, Markov imputation takes into account some aspects of the structural nature of the time series, and computes the probability of concentrations based on previously observed values.

One drawback of univariate imputation occurs when the partially observed data are fairly homogenous and do not contain patterns or extreme events that are expected during the time of equipment failure. Under these conditions, imputed values will fail to capture expected diurnal or temporal events. This could have implications for occupational settings, where specific worker tasks might be associated with high pollutant concentrations. If these high pollutant tasks are unobserved, univariate imputation may be unable to account for them. Imputed values are dependent on partially observed values, which can lead to under- or over-estimation. Despite this limitation, univariate methods appear to be viable options for imputing missing data across highly heterogeneous samples and populations.

Imputation of ambient real-time monitoring data over short time periods have not been adequately explored. Previous methods for imputing missing air pollution data are specific to fixed ambient air monitors, designed to run continuously for extended periods of time. These proposed methods are specific to the pattern of missing data, which is often for short periods that are preceded and followed by observed data; often occurring in a single station that is part of an existing air monitoring network. In the context of air monitoring, there are only a few comparable studies.

For example, Quinteros et al. (2019) imputed missing data for ambient air monitoring stations in Temuco, Chile that ran continuously from 2009 to 2014. Missing datasets artificially constructed at different levels of missingness were used to evaluate the performance of conditional and unconditional mean imputation, PMM k-nearest neighbor, multiple imputation (MI), and Bayesian principal component analysis imputation. Of these methods, PMM and MI performed best. These methods incorporated several meteorological (wind speed, temperature, humidity, precipitation) and temporal variables (day of week, month) for predicting ambient $PM_{2.5}$ concentrations that were absent in our analysis, which may explain differences in performances. Additionally, the continuous readings over several years permitted seasonal and diurnal trends to be accounted for in their analysis, potentially improving PMM and MI performance. Due to the short sampling duration in our analysis, we were unable to incorporate strong temporal trends in our imputations, which may have contributed to poor performances of PMM. Although not acknowledged by Quinteros et al. (2019), unconditional and conditional mean imputation had modest performance compared to PMM and MI in terms of R^2, MAE, Bias, and Index of Agreement using the validation dataset.

Single imputation methods for imputing missing concentrations of PM_{10} from a network of ambient air monitors were also assessed in Palermo, Sicily. Plaia & Bondi (2006) proposed a site-dependent effect method (SDEM) for imputing monitoring data based on state-space information observed at each monitoring site, and compared their method with other single imputation methods that included hour mean method (Li et al., 1999), RMM (Engles & Diehr, 2003), last-next method (Engles & Diehr, 2003), and multiple imputation (Rubin, 1996; Shafer, 1997). They argued that missingness pattern and time-site specific information must be considered for selecting an appropriate imputation method. At various lengths of missingness (2, 4-6, 8-24, >24 hours), their site specific SDEM model performed best in terms of correlation, index of agreement, root mean square deviation (RMSD), and mean absolute difference (MAD). Good performance was also observed for row mean imputation, however multiple imputation and last-next method were found to perform poorest across all methods. The authors noted that heterogeneity in concentrations observed between monitoring sites led to overestimation of imputed values between some stations. This may explain why RMM and PMM performed poorly in our analysis, again suggesting homes may have been very different from one another. However, RMM may perform well and be applicable to occupational settings where multiple workers are monitored simultaneously, and when pollutant concentrations are not expected to differ significantly between workers performing similar tasks.

Junger & De Leon (2015) further explored imputation of missing ambient air monitoring data at various levels of missingness, ranging from 5-40%. Daily concentrations were collected for 366 days from 10 ambient stations in Sao Paulo, Brazil and were highly correlated with each other. Twelve imputation methods were evaluated, which included both univariate and multivariate methods (complete case analysis, mean, median, nearest neighbor, EM models, ARIMA, general additive models, spline models). Under low levels of missingness, conditional mean performed well, whereas unconditional imputation (median, mean) performed poorly even at low levels of missingness. Multivariate methods (conditional mean and EM-models) performed exceptionally well, even as the frequency and duration of missing data increased. Our findings differed from those of Junger & De Leon (2015), as unconditional mean imputation performed well at low and high levels of missingness, whereas multivariate methods performed exceptionally poor at all levels of missingness. The authors attributed improved performance to incorporating a temporal component and the high correlation among monitoring stations. The absence of lengthy monitoring data from multiple sites in our analysis may explain the differences in imputation performances, as the context and setting in which these methods were applied differed.

Imputation methods have also been explored in longitudinal health studies. Engles & Diehr (2003) evaluated several methods for imputing missing data when missingness occurred at various stages (baseline, follow-up) of the longitudinal Cardiovascular Health Study. Simple imputation methods, such as row mean, median, new observation carried backwards, and last-next method, yielded estimates with the smallest RMSD and MAD. More advanced hot deck, regression with error, and column mean imputation methods performed poorly in terms of RMSD and MAD. The authors found that imputation methods that used person-specific longitudinal data to impute missing values within the same individual performed better than methods that used no person specific information. Although our study was looking at variability among houses rather than individual people, similar findings were seen in our analysis. More advanced multivariate imputation methods that impute concentrations based on information observed at other monitoring sites performed poorly.

This is the first short-term pollution study to evaluate various methods for imputing missing data that are frequently encountered in environmental health sciences and occupational health research. Differences in imputation performances between our study and previous studies may be explained by the unique nature and duration of our missing data. The short sampling time (24 hours) limits our capacity to assess any long-term temporal trend in $PM_{2.5}$ concentration. Additionally, missing data were not preceded and followed by periods of observed data, a common feature for most imputation methods proposed and assessed to date. Creating consecutive periods of missing data starting from a fixed time point, rather than performing multiple runs with randomly selected start points at various times of the day, may have been a major limitation.

Another limitation of our study was the relatively small sample size (n=20) and the potential effects outliers can have on imputation performances. Additionally, our findings are specific to outdoor $PM_{2.5}$ concentrations in a rural coal and wood burning community with relatively few ambient sources and may not be applicable to densely populated urbanized settings. Application of these methods to indoor environments may be possible, provided that sufficient time activity data, personal and environmental predictor variables are collected. Imputation methods presented here may also be limited to $PM_{2.5}$, and results may differ based on specific contaminants, sampling approach (i.e., active, passive, real-time), pollutant sources, sampling duration, research setting, and co-collected predictor variables. Extreme caution should always be taken before imputing missing data.

In our context, univariate imputation methods that use partially-observed data within each subject yielded reliable estimates of missing concentrations. However, imputation methods presented here may perform differently depending on the type of data and study application. Selection of an appropriate imputation method should be driven by the pattern and duration of missing data, research objective, co-collected variables, and study environment. Nonetheless, we are confident that the Markov, random, and mean imputation techniques should be considered for dealing with missing data of any real-time air monitor (fixed ambient stations, light scattering laser technology, or low-cost air monitors).

Unfortunately, a large proportion of the global population exposed to hazardous air pollutants reside in low-income communities of developing nations and experience the greatest burden of disease (WHO, 2014). The use of advanced environmental monitoring equipment in resource-limited areas that lack the modern infrastructure to power these monitors will continue to be a challenge for researchers working in these areas. Imputation offers a possible solution to this challenge, however very little attention has been given to developing or evaluating existing methods for imputing missing data for real-time environmental monitoring in community and occupational research settings.

5. Conclusion

In summary, we found Markov, random, and mean imputation performed best at providing 24-hour mean concentrations. Minute-by-minute imputation had mixed performance by metric and percent missing, however Markov imputation appears to be the better approach. Univariate imputation seems to provide a reliable solution to addressing missing data in real-time monitors operating over short periods, especially in heterogeneous environments and study populations. These methods are easy to implement and can be applied in various fields that encounter similar patterns of missing data. Our findings may be applicable to environmental health sciences and occupational health studies that rely heavily on environmental monitors to collect concentrations of contaminants in real-time in order to determine potential exposures and health hazards. Despite our findings, further research is needed to examine and identify imputation methods that are generalizable across a range of scenarios.

Acknowledgements:

This work was funded by NIEHS: P50ES026089, P30 ES006694, T32 ES007091, and EPA: R836151.

"Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number P50ES026089. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health."

"This publication was developed under Assistance Agreements No. 836151 awarded by the U.S. Environmental Protection Agency to The University of Arizona. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication."

References

1. Buuren SV, & Groothuis-Oudshoorn K (2010). mice: Multivariate imputation by chained equations in R. A. Journal of statistical software, 1–68.

2. Canales RA (2004). The cumulative and aggregate simulation of exposure framework (Order No. 3128357). Available from ProQuest Dissertations & Theses Global. (305124116).

3. Dempster AP, Laird NM, & Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1–22.

4. Donders ART, Van Der Heijden GJ, Stijnen T, & Moons KG (2006). A gentle introduction to imputation of missing values. Journal of clinical epidemiology, 59(10), 1087–1091. [PubMed: 16980149]

5. Engels JM, & Diehr P (2003). Imputation of missing longitudinal data: a comparison of methods. Journal of clinical epidemiology, 56(10), 968–976. [PubMed: 14568628]

6. Gómez-Carracedo MP, Andrade JM, López-Mahía P, Muniategui S, & Prada D (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemometrics and Intelligent Laboratory Systems, 134, 23–33.

7. Junger WL, & De Leon AP (2015). Imputation of missing data in time series for air pollutants. Atmospheric Environment, 102, 96–104

8. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, & Kolehmainen M (2004). Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 38(18), 2895–2907.

9. Lavrakas PJ (2008). Encyclopedia of survey research methods Thousand Oaks, CA: Sage Publications, Inc. doi: 10.4135/9781412963947

10. Li KH, Le ND, Sun L, & Zidek JV (1999). Spatial–temporal models for ambient hourly PM10 in Vancouver. Environmetrics: The official journal of the International Environmetrics Society, 10(3), 321–338.

11. Little RJ (1988). Missing-data adjustments in large surveys. Journal of Business & Economic Statistics, 6(3), 287–296.

12. Little R, & Rubin D (2002). Statistical Analysis with Missing Data. Newy York: John Wiley & Sons, Incorporated.

13. Little RJ, & Rubin DB (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

14. McPherson S, Barbosa-Leiker C, Mamey MR, McDonell M, Enders CK, & Roll J (2015). A 'missing not at random' (MNAR) and 'missing at random' (MAR) growth model comparison with a buprenorphine/naloxone clinical trial. Addiction (Abingdon, England), 110(1), 51–58. doi:10.1111/add.12714

15. Miettinen OS (1985). Theoretical epidemiology: principles of occurrence research in medicine (pp. 69–73). New York: Wiley.

16. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, & Stork J (2015). Comparison of different methods for univariate time series imputation in R.

17. Moritz S, & Bartz-Beielstein T (2017). imputeTS: time series missing value imputation in R. The R Journal, 9(1), 207–218.

18. Plaia A, & Bondi AL (2006). Single imputation method of missing values in environmental pollution data sets. Atmospheric Environment, 40(38), 7316–7330.

19. Quinteros ME, Lu S, Blazquez C, Cárdenas-R JP, Ossa X, Delgado-Saborit JM, … & Ruiz-Rudolph P (2019). Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. Atmospheric environment, 200, 40–49.

20. R Core Team (2017). R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria URL https://www.R-project.org/

21. Rubin D (1986). Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. Journal of Business & Economic Statistics, 4(1), 87–94.

22. Rubin DB (1988). An overview of multiple imputation. In Proceedings of the survey research methods section of the American statistical association (pp. 79–84).

23. Rubin DB (1996). Multiple imputation after 18+ years. Journal of the American statistical Association, 91(434), 473–489.

24. Schafer JL (1997). Analysis of incomplete multivariate data. Chapman and Hall/CRC.

25. Spedicato GA, Kang TS, Yalamanchi SB, Yadav D, & Cordón I (2016). The markovchain package: a package for easily handling Discrete Markov Chains in R. Accessed Dec.

26. Van Buuren S (2018). Flexible imputation of missing data. Chapman and Hall/CRC.

27. Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, … & Rowe CM (1985). Statistics for the evaluation and comparison of models. Journal of Geophysical Research: Oceans, 90(C5), 8995–9005.

28. World Health Organization. (2014). Burden of disease from the joint effects of Household and Ambient Air Pollution for 2012. Public health, environmental and social determinants of health (PHE).

Highlights:

- Various methods for imputing short-term real-time air monitoring data were assessed.
- Markov, random, & mean imputation were best at providing daily mean concentrations.
- Minute-by-minute imputation had mixed performance by metric and percent missing.
- Univariate imputation may provide a reliable solution to missing real-time data.

Table 1:

Imputation Method Performance Metrics by Level of Missingness (n=20)

	Error Metrics	Mean	Median	LOCF	Random	Markov	Kalman	RMM	RMM L	RMM F	PMM RT	PMM RTF	PMM RTL	PMM RTFL	PMM FL
20%	AB	1.82	1.90	2.69	1.82	1.71	1.98	4.42	4.33	4.45	4.98	4.94	5.25	4.19	3.81
	PAEM	14.56	14.97	22.74	14.55	14.36	15.91	55.23	39.44	52.74	61.22	72.82	70.63	59.37	37.93
	R^2	0.65	0.64	0.60	0.65	0.65	0.65	0.36	0.43	0.35	0.32	0.34	0.39	0.39	0.47
	MAE	3.18	2.45	2.82	3.18	3.26	2.46	6.00	6.42	6.07	7.14	7.04	7.07	6.36	5.59
	RMSE	14.91	14.34	15.53	14.93	14.76	14.42	22.07	23.83	23.63	25.17	24.47	24.26	22.38	21.22
40%	AB	2.48	2.77	2.70	2.49	2.42	2.75	8.12	8.12	7.56	12.83	12.21	11.86	12.32	7.59
	PAEM	18.26	19.34	20.86	18.27	17.97	19.90	103.11	75.60	87.66	252.35	189.51	159.81	176.63	76.13
	R^2	0.58	0.57	0.57	0.58	0.57	0.58	0.30	0.32	0.29	0.22	0.25	0.29	0.28	0.35
	MAE	5.79	4.10	4.79	5.79	5.80	4.30	10.90	11.80	10.94	16.61	15.88	15.73	16.18	10.93
	RMSE	17.78	17.39	17.43	17.79	17.58	17.35	27.30	29.69	28.66	36.77	33.82	34.05	33.51	27.95
60%	AB	4.07	6.31	10.81	4.07	3.27	6.87	14.15	14.00	13.07	17.22	14.84	10.00	13.29	12.37
	PAEM	25.01	35.19	93.02	25.00	22.99	44.70	177.94	131.30	146.13	251.02	309.10	194.04	226.48	129.73
	R^2	0.36	0.35	0.34	0.36	0.37	0.35	0.16	0.15	0.14	0.11	0.11	0.11	0.11	0.21
	MAE	8.51	7.46	14.34	8.51	7.92	7.72	17.81	19.08	17.72	23.67	22.36	18.33	21.27	17.24
	RMSE	24.21	24.64	31.04	24.21	23.63	24.83	36.67	39.71	38.16	46.08	45.80	40.76	42.95	35.46
80%	AB	6.18	7.80	8.71	6.34	6.86	7.59	19.43	17.98	18.42	13.93	13.71	19.14	17.70	13.08
	PAEM	34.22	46.40	51.84	34.21	36.72	52.43	247.70	178.51	206.02	185.13	94.54	156.55	152.11	99.67
	R^2	0.11	0.10	0.10	0.11	0.11	0.09	0.04	0.04	0.03	0.03	0.04	0.05	0.06	0.06
	MAE	10.22	10.05	10.68	10.23	9.90	10.91	24.78	25.88	25.01	20.78	18.29	23.91	22.41	19.78
	RMSE	29.05	29.56	29.94	29.06	29.13	29.76	44.18	47.02	46.20	44.05	39.19	46.69	45.34	39.28

Absolute Bias (AB), Percent Absolute Error in Means (PAEM), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Row Mean Method (RMM), Predictive Mean Matching (PMM), Location (L), Fuel Type (F), Relative Humidity (R), Temperature (T)

Table 2:

Recommended Imputation Methods by Level of Missingness and Performance Metric[*]

Error Metrics	Percent Missing			
	20%	40%	60%	80%
AB	Markov Random Mean	Markov Mean Random	Markov Random Mean	Mean Random Markov
PAEM	Markov Random Mean	Markov Mean Random	Markov Random Mean	Random Mean Markov
R^2	Markov Random Kalman	Kalman Random Mean	Markov Random Mean	Markov Random Mean
MAE	Median Kalman LOCF	Median Kalman LOCF	Median Kalman Markov	Markov Median Mean
RMSE	Median Kalman Markov	Kalman Median LOCF	Markov Random Mean	Mean Random Markov

[*] Top three methods listed in ranking order

Absolute Bias (AB), Percent Absolute Error in Means (PAEM), Mean Absolute Error (MAE), Root Mean Square Error (RMSE)