Environmental Research 134 (2014) 466–473

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Multiple imputation for assessment of exposures to drinking water contaminants: Evaluation with the Atrazine Monitoring Program

Rachael M. Jones [a,*], Leslie T. Stayner [b], Hakan Demirtas [b]

[a] Division of Environmental and Occupational Health Sciences, School of Public Health, University of Illinois at Chicago, 2121 W. Taylor St. (M/C 922), Chicago, IL 60612, United States
[b] Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, United States

ARTICLE INFO

Article history:
Received 4 June 2013
Received in revised form
12 May 2014
Accepted 30 July 2014
Available online 29 October 2014

Keywords:
Multiple imputation
Drinking water quality
Atrazine
Atrazine Monitoring Program

ABSTRACT

Background: Drinking water may contain pollutants that harm human health. The frequency of pollutant monitoring may occur quarterly, annually, or less frequently, depending upon the pollutant, the pollutant concentration, and community water system. However, birth and other health outcomes are associated with narrow time-windows of exposure. Infrequent monitoring impedes linkage between water quality and health outcomes for epidemiological analyses.
Objectives: To evaluate the performance of multiple imputation to fill in water quality values between measurements in community water systems (CWSs).
Methods: The multiple imputation method was implemented in a simulated setting using data from the Atrazine Monitoring Program (AMP, 2006–2009 in five Midwestern states). Values were deleted from the AMP data to leave one measurement per month. Four patterns reflecting drinking water monitoring regulations were used to delete months of data in each CWS: three patterns were missing at random and one pattern was missing not at random. Synthetic health outcome data were created using a linear and a Poisson exposure–response relationship with five levels of hypothesized association, respectively. The multiple imputation method was evaluated by comparing the exposure–response relationships estimated based on multiply imputed data with the hypothesized association.
Results: The four patterns deleted 65–92% months of atrazine observations in AMP data. Even with these high rates of missing information, our procedure was able to recover most of the missing information when the synthetic health outcome was included for missing at random patterns and for missing not at random patterns with low-to-moderate exposure–response relationships.
Conclusions: Multiple imputation appears to be an effective method for filling in water quality values between measurements.

1. Introduction

The quality of drinking water provided by community water systems (CWSs) is determined by periodic measurement of contaminants in the CWS's finished drinking water. Measurement frequencies are specified in the National Primary Drinking Water Regulations (NPDWR), promulgated by the US EPA under authority of the Safe Drinking Water Act. Measurement frequencies vary among contaminants, with contaminant concentration, and with features of the CWS. Typically, water quality is measured quarterly, annually, or once every several years. Drinking water quality data are contained in the Safe Drinking Water Information System (SDWIS), and are available to the public upon request from state agencies. Recently, SDIWS data have been integrated into the Environmental Public Health Tracking (EPHT) Network, which will increase their accessibility to the public and researchers.

SDWIS data have been utilized in several epidemiological studies in the United States (Rinksy et al., 2012; Ochoa-Acuna et al., 2009; Ward et al., 2007; Weyer et al., 2001; Munger et al., 1997), and similar data have been used in other countries (Migeot et al., 2013; Chang et al., 2010; Villanueva et al., 2005). Since the water quality measurements are relatively infrequent in a calendar year, epidemiological studies have generally averaged data across years and CWSs (Rinksy et al., 2012; Weyer et al., 2001) or restricted the study population to metropolitan areas served by a single CWS, with or without supplemental monitoring programs (Munger et al., 1997; Ochoa-Acuna et al., 2009; Chang et al., 2010). These approaches, however, have significant limitations that are particularly difficult to overcome in EPHT applications.

The objective of EPHT is to utilize existing data collection systems to survey and evaluate environmental health, including

[*] Corresponding author.
E-mail address: rjones25@uic.edu (R.M. Jones).

http://dx.doi.org/10.1016/j.envres.2014.07.027
0013-9351/© 2014 Elsevier Inc. All rights reserved.

changes in environmental quality, population health and linkages between environmental quality and health. In the EPHT framework, supplemental targeted data collection is not feasible. For example, additional water quality monitoring cannot be conducted to fill in gaps between regulatory-required monitoring. Cohort epidemiological studies require large populations, so the study cannot be restricted to small areas with the most robust data without losing power to detect associations. In addition, long-term averages may introduce significant exposure misclassification in the context of non-chronic health effects, such as adverse birth outcomes (e.g., low birth weight or pre-term birth) for which the critical time-window of exposure is measured in months (Rogers and Kavlock, 2008). These limitations led us to consider statistical methods by which to estimate drinking water quality between measurements.

The specific objective of this work is to evaluate whether multiple imputation is a feasible approach for estimating water quality in each calendar month when water quality is monitored quarterly, annually or less frequently. Multiple imputation is one of many statistical methods that can be used to fill in missing values, but was selected for evaluation for two reasons. Firstly, there has been relatively limited application and evaluation of the method for environmental data (Hopke et al., 2001; Le et al., 2007; Nieh et al., 2014). Though previous research has found multiple imputation to perform relatively well, even in the absence of structured time-series imputation models for time-series data (Hopke et al., 2001), the high frequency of missing data in the context of monthly drinking water quality warrants further evaluation. Secondly, multiple imputation has advantages relative to single value imputations (e.g., mean value substitution) and common likelihood-based methods (e.g., harmonic analysis or forecasting by Becker et al., 2006; Dilmaghani et al., 2007; Weerasinghe, 2010) used to fill in missing values in environmental data. Single-value imputations reduce variance in the water quality data: exploratory analysis in these data found that filling missing observations with CWS-specific annual average atrazine concentrations underestimated the exposure-response association, even when data are missing at random (data not shown). A particular advantage of multiple imputation with respect to likelihood-based methods is that the method identifies missing data as a source of random variation distinct from ordinary sampling variability, which maintains appropriately wide standard errors for inference (Demirtas and Hedeker, 2007).

The specific multiple imputation method employed in this study was Multivariate Imputation by Chained Equations (MICE) in which unique imputation models are specified for each variable with missing values and imputed sequentially (van Buuren, 2012). This method of multiple imputation has not been previously applied to environmental data, to our knowledge. The MICE algorithm is implemented as follows (White et al., 2011). Consider a data set with $x_1, x_2, \ldots x_k$ variables. Initially, all missing values are replaced by random sampling from observed values. For variable x_i with missing values, x_i is regressed on the other variables using the specified imputation model, and the missing values in x_i are replaced by draws from the posterior predictive distribution of x_i. In one cycle, this is repeated for all of the x_i, $i = \{1, 2, \ldots, k\}$, variables with missing values; and the cycle is repeated several times to produce an imputed data set. This process is repeated m times, to give m imputed data sets. MICE has the advantage of not requiring a joint distribution, such that different types of variables can be multiply imputed using appropriate regression models and subsets of the variables.

The objective was approached through simulation. To allow the performance of multiple imputation to be evaluated relative to real data, the simulation was performed using data from the Atrazine Monitoring Program (AMP), rather than from SDWIS. The AMP is a special program managed by the US EPA, which measures the concentration of atrazine in CWSs known to be heavily impacted by atrazine approximately every two weeks. There are thousands of CWS in the Midwest (Jones et al., 2014) of which 90 CWSs in Illinois, Indiana, Iowa, Missouri and Ohio participated in the AMP at some time during the years 2006–2009. Owing to the relatively high-frequency of measurement in the AMP CWSs, it was possible to create missing values by deleting measurements in these data, and to compare inferences made from multiply imputed data to inferences from the real data.

Table 1
Summary of AMP data used in the simulation study. The geometric mean (GM) and the geometric standard deviation (GSD) were estimated using the method of maximum likelihood with consideration for censoring at detection limits 0.05 μg/L (2007–2009) and 0.1 μg/L (2006).

		Estimated				
Chemical	Water Type	GM (μg/L)	GSD	Maximum (μg/L)	N	Percent Censored
Atrazine	Finished	0.159	4.29	17.8	3325	18
Atrazine	Raw	0.338	4.10	34.0	3282	8
Simazine	Finished	0.002	8.82	15.98	3325	68
Simazine	Raw	0.048	7.89	21.6	3282	55
DIA	Finished	0.034	4.30	2.51	3325	66
DEA	Finished	0.062	3.39	2.20	3325	51

2. Methods

2.1. Atrazine Monitoring Program data

The AMP monitors the concentration of atrazine and related chemicals in CWSs vulnerable to atrazine pollution. Data are available online from the EPA Office of Pesticide Programs. The AMP continues monitoring campaigns conducted by private companies and professional associations (Graziano et al., 2006). Participating CWSs monitor finished (treated) and raw (untreated) drinking water approximately every two weeks: frequency may increase in the spring and summer, and decrease in the winter. In general, atrazine concentrations decrease over time within each CWS, until the CWS is removed from the AMP. The AMP data include the CWS name, and the county and state in which each CWS is located.

For the years 2006–2009, 90 CWSs in Illinois, Indiana, Iowa, Missouri and Ohio ever participated in the AMP. Not every CWS participated in the AMP every year 2006–2009. The years 2006–2009 and CWSs in Midwestern states were selected because they fall into the time period and geographical area of interest for our larger study, and have consistent analytical detection limits in each CWS. During this period, the AMP consistently included the agents atrazine, simazine, deethlatrazine (DEA), and deisopropylatrazine(DIA). Simazine is a related pesticide, while DEA and DIA are metabolites of atrazine and simazine. The limit of detection for all chemicals in 2006 was 0.1 μg/L , and in 2007–2009 was 0.05 μg/L.

The AMP data were prepared as follows. Random deletion was used to retain only one value in each calendar month for each CWS. In each calendar year, a CWS was excluded if there were fewer than 6 months with observations; if 6–11 months had observations rows were added to indicate missing values. As a result, there were 289 CWS-years, and 143 rows (4.2%) with missing observations of chemicals in finished water and 186 rows (5.4%) with missing observations in raw water. This approach was taken because elimination of CWS-years with observations missing in any month would reduce the data substantially, and introduce bias if the missing observations were not missing at random.

The data used in the simulation study are summarized in Table 1. Lognormality of the chemical concentrations is indicated by the high GSD values in Table 1, and was confirmed by quartile–quartile plots (not shown). Many of chemicals are

Table 2
Pearson's correlation of log-transformed chemical concentrations of AMP data used in the simulation study. Atrazine and simazine were measured in finished (F) and raw (R) drinking water. All correlation tests have $p < 0.05$.

Chemicals	Atrazine-F	Atrazine-R	Simazine-F	Simazine-R	DIA	DEA
Atrazine-F	1.00					
Atrazine-R	0.70	1.00				
Simazine-F	0.20	0.04	1.00			
Simazine-R	0.11	0.12	0.57	1.00		
DIA	0.44	0.24	0.47	0.40	1.00	
DEA	0.72	0.49	0.18	0.08	0.47	1.00