Chemosphere 191 (2018) 412–416

Contents lists available at ScienceDirect

Chemosphere

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

Evaluation of methods for managing censored results when calculating the geometric mean

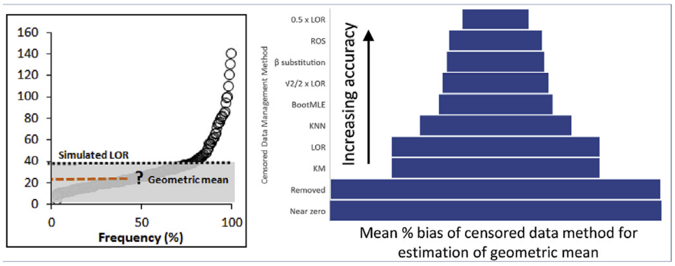
Hannah G. Mikkonen [a, b, e], Bradley O. Clarke [b, c], Raghava Dasika [d], Christian J. Wallis [e], Suzie M. Reichman [a, b, *]

[a] *School of Engineering, RMIT University, GPO Box 2476, Melbourne, Australia*
[b] *Centre for Environmental Sustainability and Remediation, RMIT University, Victoria, Australia*
[c] *School of Science, RMIT University, GPO Box 2476, Melbourne, Australia*
[d] *Australian Contaminated Land Consultants Association, Victoria, Australia*
[e] *CDM Smith, Richmond, Victoria, Australia*

HIGHLIGHTS

- Diverse range of censored data methods evaluated for deriving the geometric mean.
- Censored data methods tested on real soil datasets.
- Substitution of half the limit of reporting amongst most accurate methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 August 2017
Received in revised form
4 October 2017
Accepted 6 October 2017
Available online 8 October 2017

Handling Editor: A. Gies

Keywords:
Censored data
Geometric mean
Limit of reporting
Nearest neighbour imputation
Soil
Environmental analysis

ABSTRACT

Currently, there are conflicting views on the best statistical methods for managing censored environmental data. The method commonly applied by environmental science researchers and professionals is to substitute half the limit of reporting for derivation of summary statistics. This approach has been criticised by some researchers, raising questions around the interpretation of historical scientific data. This study evaluated four complete soil datasets, at three levels of simulated censorship, to test the accuracy of a range of censored data management methods for calculation of the geometric mean. The methods assessed included removal of censored results, substitution of a fixed value (near zero, half the limit of reporting and the limit of reporting), substitution by nearest neighbour imputation, maximum likelihood estimation, regression on order substitution and Kaplan-Meier/survival analysis. This is the first time such a comprehensive range of censored data management methods have been applied to assess the accuracy of calculation of the geometric mean. The results of this study show that, for describing the geometric mean, the simple method of substitution of half the limit of reporting is comparable or more accurate than alternative censored data management methods, including nearest neighbour imputation methods.

1. Introduction

Environmental datasets often include results below the limit of reporting (LOR) that are referred to as "censored" results.

[*] Corresponding author. School of Engineering, RMIT University, GPO Box 2476, Melbourne, Australia.
E-mail address: suzie.reichman@rmit.edu.au (S.M. Reichman).

https://doi.org/10.1016/j.chemosphere.2017.10.038
0045-6535/© 2017 Elsevier Ltd. All rights reserved.

Commonly recommended statistical methods for interpretation of environmental data (viz., geometric mean) cannot be undertaken without making an assumption about the censored values. Over the last 60 years researchers have developed and evaluated many different statistical approaches for the management of censored results (Cohen, 1957; Peto and Peto, 1972; Gilbert, 1987; Ganser and Hewett, 2010), with the preferred approach often differing between disciplines and depending on the statistical analysis being conducted. There remains no consensus on which censored data analysis method is best suited for calculation of the geometric mean.

Scientists and environmental assessors commonly use statistical approaches to describe the condition of the environment. Derivation of the geometric mean $(n\sqrt{(x_1\ .\ x_2 \ldots . x_n)}\ )$ in preference to the arithmatic mean $((x_1+x_2\ \ldots\ +x_n\ )/n)$, has been recommended to describe the average or the central tendency of elemental concentrations (Reimann et al., 2008). Given that environmental data is often log-normally distributed, the use of the geometric mean normalizes the data being averaged, and is therefore a good approximation of the median concentration (Reimann et al., 2008). Conversely, the arithmatic mean can be skewed away from the median due to the presence of outliers and anomalus results.

This study compared the accuracy of censored data management methods for calculation of the geometric mean using four soil datasets and three levels of simulated censorship. Further, this study included comparison of K nearest neighbour (KNN) imputation. Although KNN imputation methods are increasingly being used for statistical interpretation of environmental data (de Caritat and Cooper, 2011; Grunsky et al., 2014; Harris and Grunsky, 2015; Makvandi et al., 2016), to the researchers best knowledge, comparison of the accuracy of KNN imputation methods with other techniques for management of censored data for calculation of the geometric mean has not been previously published.

2. Methods

2.1. The dataset

Four datasets were chosen for evaluation. The datasets were manganese (Mn) and zinc (Zn) concentrations measured in soils derived from Quaternary basalt parent materials in Greater Melbourne, Victoria, Australia from a soil survey and an open-source dataset (Mikkonen et al., 2017). These datasets were selected as no Zn or Mn concentrations were below the LOR of 5 mg/kg and a wide variety of replicates were present, with *n* ranging between 27 and 194 (Table 1). The number of samples in each dataset was considered typical of environmental soil datasets used for evaluation of soil contamination at a site assessment scale. Simulated LORs (Table 1) were applied to the Zn and Mn datasets at approximately 35%, 60% and 80% censoring, respectively. The percentage of censored results varied between datasets because the percentage of results censored was restricted to the nearest result.

2.2. Censored data analysis methods

The censored data analysis methods tested were substitution, maximum likelihood estimation (MLE), regression on order substitution (ROS), nonparametric, nearest neighbour imputation and removal of censored data (Table 2).

The distribution of Mn and Zn concentrations for the survey and open-source datasets was assessed against normal, lognormal, Weibull and loglogistic distribution using statistical package Minitab 17 (Minitab, 2010). Zinc concentrations were closest to a lognormal or loglogistic distribution and Mn concentrations were closest to normal distribution and weibull distribution (Supplementary Material). As such, where statistical methods required assumption of data distribution (i.e. for ROS and MLE), Zn concentrations were log transformed whereas Mn concentrations were assumed to be normally distributed, and thus not transformed.

2.3. Calculation of the geometric mean and geometric standard deviation

The geometric mean was calculated by taking the *nth* root of the product of *n* numbers $(n\sqrt{(x_1\ .\ x_2 \ldots . x_n)})$, except for KM and ROS where, due to the nature of the ranked data (where by all numbers are rounded), the geometric mean was calculated by back transforming, the mean of log transformed data. The geometric mean was calculated for Mn and Zn concentrations, using each of the censored data management techniques, at the three simulated levels of censoring (approximately 35%, 60% and 80%). The estimated geometric mean after censoring (E) was compared to the measured geometric mean with no censuring (M), using the percentage bias equation below (Equation (1)):

$$\text{Percentage difference} = (\text{E-M})\ \text{x}\ 100/\text{M} \qquad (1)$$

In addition, the geometric standard deviation (GSD) was calculated for complete datasets using the exponential of the standard deviation of log transformed data, as shown in Table 1.

3. Results and discussion

The accuracy of the censored data management methods to estimate the measured geometric mean generally decreased with increased censorship (Fig. 1), with the mean percentage bias for each substitution method (except near zero) and the KM method more than doubling when censorship increased from 35% to 55–60% (Table 3). However, accurate estimates (less than 5% bias) of the geometric mean were still achieved for Zn, for the open-source dataset, using the ROS and 0.5 × LOR methods, even when censoring was high, at 80%.

Trends of the accuracy of the tested methods were comparable for the Zn dataset and Mn datasets, indicating that the distribution of data (log-normal and normal distribution) did not greatly change

Table 1
Summary statistics (mg/kg) and percentage of censored results for Zn and Mn concentrations, under three simulated limits of reporting (LOR), for survey (S) and open-source (OS) soil datasets.

Element	Data Source	*n*	Measured LOR (mg/kg)	Measured GM (mg/kg)	Measured GSD (mg/kg)	Simulated LOR (mg/kg)			Simulated % of results censored		
						Low	Mod	High	Low	Mod	High
Zn	OS	194	5	25.94	0.61	20	30	40	37.1	60	79
	S	40	5	29.6	0.77	20	30	65	35	55	80
Mn	OS	27	5	330.53	1.68	340	370	500	37	55	77
	S	41	5	270.6	3.55	200	550	750	34	59	82

Notes: GM = Geometric mean, GSD = Geometric Standard Deviation.