Chemosphere 93 (2013) 1701–1709

Contents lists available at SciVerse ScienceDirect

Chemosphere

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

Method of median semi-variance for the analysis of left-censored data: Comparison with other techniques using environmental data

Hugo José Oliveira Zoffoli[a], Carlos Alberto Alves Varella[b], Nelson Moura Brasil do Amaral-Sobrinho[a], Everaldo Zonta[a], Alfredo Tolón-Becerra[c,*]

[a] *Department of Soils, Rio de Janeiro Federal Rural University, Seropédica, RJ, Brazil*
[b] *Department of Rural Engineering, Rio de Janeiro Federal Rural University, Seropédica, RJ, Brazil*
[c] *Department of Rural Engineering, University of Almería, Almería, Spain*

HIGHLIGHTS

- The new method of median semi-variance is an adequate option for the easy analysis of censored data.
- The parametric methods showed slightly biased behavior with good accuracy.
- The simple substitution $L/2$, Inter and $L/\sqrt{2}$ methods can be used with caution under specific conditions.

ARTICLE INFO

Article history:
Received 24 December 2012
Received in revised form 12 May 2013
Accepted 20 May 2013
Available online 5 July 2013

Keywords:
Data imputation
Detection limit
Trace elements
Non-detected data

ABSTRACT

In environmental monitoring, variables with analytically non-detected values are commonly encountered. For the statistical evaluation of these data, most of the methods that produce a less biased performance require specific computer programs. In this paper, a statistical method based on the median semi-variance (SemiV) is proposed to estimate the position and spread statistics in a dataset with single left-censoring. The performances of the SemiV method and 12 other statistical methods are evaluated using real and complete datasets. The performances of all the methods are influenced by the percentage of censored data. In general, the simple substitution and deletion methods showed biased performance, with exceptions for $L/2$, Inter and $L/\sqrt{2}$ methods that can be used with caution under specific conditions. In general, the SemiV method and other parametric methods showed similar performances and were less biased than other methods. The SemiV method is a simple and accurate procedure that can be used in the analysis of datasets with less than 50% of left-censored data.

1. Introduction

When assessing environmental data, it is common to encounter variables with values less than the analytical detection limit (*DL*). These data are called left-censored in the literature and can show a single *DL* value for all observations or multiple *DL* (Blackwood, 1991; She, 1997; Kuttatharmmakul et al., 2000; Thompson and Nelson, 2003; Hewett and Ganser, 2007). In this paper we consider only one single *DL* value, condition which is relevant in assessing the performance of the methods, since the performance of some methods vary depending on the data display single or multiple *DL* (EFSA, 2010).

The best practice to reduce the censored data is to focus efforts on reducing the *DL* (de Solla et al., 2012); however, if the dataset cannot be reanalyzed to eliminate censored data, appropriate statistical procedures must be used. The most common procedures in studies with censored environmental data involve the simple substitution of the censored values with a fraction of the *DL*. While these procedures have been criticized by some authors (Helsel, 2006; Baize et al., 2009) because they have little statistical underpinning, they have been widely used because of their simplicity and ease of operation. Many other procedures have been proposed to manipulate censored data (Helsel and Cohn, 1988; El-Shaarawi et al., 1989; She, 1997; Kuttatharmmakul et al., 2000; Lubin et al., 2004; Krishnamoorthy et al., 2009; Flynn, 2010; Ganser and Hewett, 2010; Chen et al., 2011), and the adoption of these procedures in the evaluation of environmental data has been steadily increasing (Leith et al., 2010; Wood et al., 2010). Greater use of these methods by the scientific community has been facilitated by the development of widely available computer programs to implement them (Flynn, 2010; Ganser and Hewett, 2010).

[*] Corresponding author. Address: Department of Rural Engineering, University of Almería, Ctra. Sacramento s/n, La Cañada de San Urbano, 04120 Almería, Spain. Tel.: +34 950015952; fax: +34 950015491.

E-mail addresses: zoffolihjo@yahoo.com.br (H.J.O. Zoffoli), varella@ufrrj.br (C.A.A. Varella), nelmoura@ufrrj.br (N.M.B. do Amaral-Sobrinho), ezonta@ufrrj.br (E. Zonta), atolon@ual.es (A. Tolón-Becerra).

0045-6535/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.chemosphere.2013.05.041

Recently, Flynn (2010) developed a technique for estimating the mean and standard deviation of censored data based on maximizing the Shapiro–Wilk statistic using the "Solver" function of Microsoft® Excel. Ganser and Hewett (2010) also developed a technique known as Beta-Substitution that can be performed with the use of common electronic spreadsheets. Neither the Flynn method nor the Beta-Substitution method have been compared with a greater diversity of statistical methods and neither have been evaluated with real and complete data (i.e., environmental or experimental data with absence of censored values), which is necessary to prove their validity (Antweiler and Taylor, 2008).

The objective of this paper was to propose a new method that is simple and easy to execute using electronic spreadsheets based on median semi-variance (SemiV) of normally distributed data or transformed to normality to estimate position and spread statistics in datasets with less than 50% of single left-censoring. The performances of several statistical methods were also compared using a real and complete dataset with artificial adjustments to simulate censored values.

2. Materials and methods

2.1. Datasets

A total of 57 datasets containing variables from four environmental sample types collected in the field (Table 1) were used to evaluate the performance of 12 statistical procedures in the estimation of the arithmetic mean and standard deviation.

The data were evaluated both in their raw form (original scale) and in their logarithmic transformed form (logarithmic scale).

A log–normal distribution was assumed for all variables, even if some variables did not achieve normality by the Shapiro–Wilk test (Shapiro and Wilk, 1965) when transformed to a natural logarithm (Table 1). The log–normal distribution was chosen because it is the most commonly reported distribution in environmental data (Helsel, 2005), although other distributions such as Weibul and Gamma are also common (EFSA, 2010).

Each sample type had a different number of observations and measured variables. The selected variables showed a wide variation in the geometric standard deviation (GSD) estimated with complete data (GSD is the antilog of the standard deviation of the logarithmic transformation data); variation was also observed in the *p*-value characterizing the normality by the Shapiro–Wilk test on the original scale (SWp) and with log-transformed data (SWpLn) (Table 1).

The samples labeled "Waste" were the tailings (drill cuttings) generated in the drilling and prospecting of oil wells onshore in Bahia state in the northeastern region of Brazil. The samples were collected from 16 different drilling depths (between 0 and 2853 m). Heavy metals were extracted by aqua regia and quantified by inductively coupled plasma atomic emission spectroscopy (ICP-OES). The organic compounds were measured by gas chromatography (Amaral-Sobrinho et al., 2011).

Plant samples represented different plant parts collected from 108 tobacco farms located in 38 municipalities of Brazil's southern region. The parts collected were leaves from the X position (lower plant portion), B position (upper plant portion) and roots. Heavy metals were extracted by digestion with microwave heating in a $HNO_3 + H_2O_2$ solution and measured by inductively coupled plasma mass spectrometry (ICP-MS) (Amaral-Sobrinho et al., 2010).

Table 1
Dataset characteristics.

Sample type/variable	GSD	SWp	SWpLn	Sample type/variable	GSD	SWp	SWpLn
Waste (n = 16)				*Plant (n = 108)*			
Al (aqua regia)	1.6	0.126	0.073	Cu leaf x	2.1	<0.001	0.144
Ba (aqua regia)	2.4	0.056	0.132	Fe leaf x	2.2	<0.001	0.914
B (aqua regia)	1.4	0.967	0.487	Mn leaf x	3.0	<0.001	0.266
Pb (aqua regia)	1.7	0.111	0.850	Ni leaf x	1.7	<0.001	0.034
Co (aqua regia)	1.5	0.194	0.207	Zn leaf x	2.0	<0.001	0.271
Cu (aqua regia)	1.5	0.567	0.039	Cd leaf x	1.9	<0.001	0.556
Cr (aqua regia)	1.4	0.106	0.015	Pb leaf x	1.8	<0.001	0.381
Fe (aqua regia)	1.5	0.513	0.615	Cu leaf b	2.8	<0.001	0.017
Mn (aqua regia)	1.4	0.446	0.033	Fe leaf b	1.8	<0.001	0.181
Ni (aqua regia)	1.5	0.389	0.208	Mn leaf b	2.9	<0.001	0.237
V (aqua regia)	1.4	0.356	0.295	Ni leaf b	1.7	<0.001	0.064
Zn (aqua regia)	1.5	0.029	0.442	Zn leaf b	1.8	<0.001	0.993
Naphthalene	2.0	<0.001	0.901	Pb leaf b	1.8	<0.001	0.306
Acenaphthylene	2.4	<0.001	0.258	Cu root	1.7	<0.001	<0.001
Phenanthrene	2.0	0.009	0.876	Fe root	1.7	<0.001	0.845
Anthracene	1.8	0.003	0.723	Mn root	2.4	<0.001	0.043
Fluoranthene	2.0	0.012	0.482	Zn root	1.5	<0.001	0.473
Pyrene	2.0	0.006	0.925	Pb root	1.8	<0.001	0.863
Benzoanthracene	2.3	0.003	0.953	*Sediment (n = 81)*			
Chrysene	1.9	0.031	0.826	Al soluble in H_2O	2.1	<0.001	0.682
n-Alkane	1.7	0.108	0.327	Al (aqua regia)	2.7	<0.001	<0.001
UCM	1.7	0.012	0.437	Ba (aqua regia)	1.8	<0.001	0.009
TPH	1.7	0.030	0.277	Co (aqua regia)	1.6	<0.001	0.555
Soil (n = 216)				Fe (aqua regia)	1.5	<0.001	0.170
Al (aqua regia)	2.3	<0.001	0.223	Zn (aqua regia)	1.8	<0.001	0.203
Co (aqua regia)	2.7	<0.001	0.438	Cr (aqua regia)	2.7	<0.001	<0.001
Cr (aqua regia)	2.5	<0.001	0.004	Ni (aqua regia)	1.8	0.002	0.044
Fe (aqua regia)	2.3	<0.001	0.053	V (aqua regia)	2.5	<0.001	0.009
Mn (aqua regia)	3.2	<0.001	0.035	Pb (aqua regia)	2.7	<0.001	<0.001
				Mn (aqua regia)	2.2	<0.001	<0.001

GSD = geometric standard deviation.
SWp = *p*-value for the Shapiro–Wilk test for normality in the original scale.
SWpLn = *p*-value for the Shapiro–Wilk test for normality in the log scale.
UCM = branched and cyclic alkanes + by-products of the hydrocarbon transformations.
TPH = total petroleum hydrocarbon.