Quantitative analysis of left-censored concentration data in environmental site characterization

By

Niloofar SHOARI

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTRÉAL, NOVEMBRE 22, 2016

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Jean-Sébastien Dubé, Thesis Supervisor
Department of Construction Engineering, École de technologie supérieure

Michel Rioux, President of the Board of Examiners
Department of Automated Manufacturing Engineering, École de technologie supérieure

François Brissette, Member of the jury
Department of Construction Engineering, École de technologie supérieure

Yannic A. Éthier, Member of the jury
Department of Construction Engineering, École de technologie supérieure

Nildari Basu, External Evaluator
Department of Natural Resource Sciences, McGill University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

ON NOVEMBRE 2, 2016

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGMENT

This dissertation would not have been possible were it not for the continuous encouragement, support of many individuals. First, I would specially like to thank my advisor, Prof. Jean-Sébestien Dubé, for his guidance, encouragement, and flexibility that allowed me to pursue my own ideas throughout this research. I am so grateful to be accepted in his group. I would also like to acknowledge my PhD Committee members: Prof. Yannic Ethier, Prof. Francois Brissette, Prof. Michel Rioux, and Prof. Nildari Basu for their valuable time in reviewing my thesis. I must thank Prof. Gabriel Lefebvre, Director of the Department of Construction Engineering, and Me Louis Marquis, the Secretary General, for their invaluable support and great understanding.

I owe an enormous debt of gratitude to Prof. Shoja'eddin Chenouri who hosted me at the University of Waterloo and gave me guidance and expert advice on statistical aspects of this dissertation. Many thanks go to Prof. Mohsen Pourahmadi who provided great support and advice during my visit at Texas A&M University. Thanks are also due to my dear friend, Mirela, for her positive attitude, wise words, and encouragements. She brought fun into the lab.

My heartfelt thanks are extended to my mother and brother and in particular my late father for their inestimable support and love. Finally, I would like to thank my husband, Shahram, for his endless love and support, and for giving me the confidence in my abilities to do anything that I set my mind to. No words can express how lucky I am to have you in my life.

L'ANALYSE QUANTITATIVE DES DONNÉES DE CONCENTRATIONS CENSURÉES EN CARACTÉRISATION ENVIRONNEMENTALE DES SITES CONTAMINÉS

Niloofar SHOARI

RÉSUMÉ

L'analyse statistique de concentrations des contaminants dans les sols, l'eau et l'air constitue une composante essentielle de la caractérisation des sites contaminés. Ce type d'analyse présente des défis attribuables à la présence d'observations non détectées ou censurées à gauche relatives à des mesures inférieures à une limite de détection. Il est nécessaire de prendre en compte les valeurs censurées dans un ensemble de mesures de concentrations parce qu'elles n'impliquent pas l'absence de contamination, mais le manque de précision des instruments de mesure. En effet, des traces de polluants dangereux peuvent constituer des risques pour la santé humaine et l'environnement. Même si une étude environnementale permet de fournir un échantillon représentatif de données de concentration conformément à des protocoles analytiques bien conçus et à des procédures de validation des données, des analyses statistiques inadéquates ne prenant pas en compte correctement les observations censurées peuvent ne pas refléter l'état réel du site. Manifestement, des mesures de réhabilitation basées sur une image faussée des conditions de contamination pourraient être inefficaces et non durable écologiquement et économiquement.

L'objectif principal de cette recherche vise à examiner en détail l'influence des concentrations non détectées sur les décisions découlant des études de caractérisation des sols contaminés. À cette fin, nous explorons différentes méthodes statistiques (i) pour estimer les statistiques descriptives (ii), pour quantifier l'incertitude sur les estimés, et (iii) pour analyser les éventuelles dépendances liées aux observations groupées, lesquelles peuvent être inhérentes aux techniques d'échantillonnage. Le remplacement de valeurs censurées par une constante choisie de façon arbitraire est une pratique courante tant chez les spécialistes que chez les chercheurs. En revanche, il existe un certain nombre de méthodes paramétriques et non paramétriques permettant de tirer des déductions à partir des données censurées et, par conséquent, offrir un aperçu plus exact du problème. Les méthodes paramétriques, comprenant les procédures basées sur le maximum de vraisemblance et la régression, évaluent les statistiques descriptives grâce à l'ajustement d'une distribution paramétrique aux données. Étant donnée l'asymétrie à droite des données de concentration, les distributions gamma, Weibull et log-normale constituent les modèles paramétriques les plus plausibles, ce dernier type étant le plus souvent utilisé dans les études environnementales. Les procédures non paramétriques telles que la méthode Kaplan-Meier, cependant, ne nécessitent aucune hypothèse de distribution.

La présente étude utilise un exercice exhaustif de simulations des données, où le type de distribution sous-jacent est connu, afin d'évaluer la performance des estimateurs paramétriques et non paramétriques. Les simulations comprennent un grand nombre de scénarios avec différents pourcentages de censure, tailles d'échantillons de données et degrés

d’asymétrie des données. Cette recherche met également en évidence l'importance d'examiner la robustesse des méthodes paramétriques contre une mauvaise spécification du modèle de distribution. En utilisant les données simulées, nous élucidons comment la substitution des valeurs censurées fausse les estimations et pourquoi cette approche devrait être écartée, même quand il s'agit de données où le pourcentage de censure est limité. Nous avons découvert que la méthode du maximum de vraisemblance reposant sur l'hypothèse de la loi log-normale est hautement sensible à l'asymétrie des données, à la taille de l'échantillonnage et au pourcentage des valeurs censurées. Alors que la méthode de maximum vraisemblance basée sur la distribution log-normale est principalement utilisée dans les études environnementales, notre avons constaté qu'il faut faire preuve de prudence en supposant une distribution log-normale. Nous recommandons plutôt l'estimateur du maximum de vraisemblance reposant sur une distribution gamma, ainsi que des méthodes fondées sur la régression (utilisant un modèle log-normal ou gamma) et la technique Kaplan-Meier. En ce qui concerne les incertitudes sur les estimations relatives aux données réelles de concentration, pour lesquelles la vraie structure des données est inconnue, nous évaluons la performance des estimateurs paramétriques et non paramétriques en employant une technique de «bootstrapping». Les conclusions tirées du bootstrapping de données réelles sont conformes avec celles déduites à partir des données simulées.

Une partie importante de cette recherche porte sur la présence d'une corrélation entre les concentrations, en lien avec des techniques d'échantillonnage. Nous fournissons un fondement statistique et conceptuel ainsi que les raisons d'appliquer des modèles à effets mixtes capables d'accommoder la dépendance entre les données tout en tenant compte des observations censurées. Les méthodes statistiques habituelles tiennent pour acquis que les échantillonnages de données de concentration sont indépendants. Cependant, dans les études de la caractérisation environnementale de sites, cette supposition sera probablement contredite parce que les observations de concentration obtenues, par exemple, du même trou de forage pourraient être corrélées. Cela peut ensuite affecter les procédures de détermination de nombre d’échantillons de sol. Ainsi, nous avons eu recours à des modèles à effets mixtes pour capturer d'éventuelles dépendances dans les données ainsi que la variabilité entre groupes. La pertinence de l'estimé de la variabilité inter-forage est attestée par la détermination du nombre optimal de trous de forage de même que d'échantillons devant être prélevées à chaque trou de forage. Le modèle à effets mixtes que nous proposons fournit un aperçu de l'étendue verticale de la contamination, ce qui peut être utile pour concevoir des stratégies d'assainissement.

Les conclusions de cette recherche doctorale aident à accroître la sensibilisation à l'importance des observations censurées auprès de la communauté scientifique, des professionnels de l’environnement, ainsi que des décideurs politiques. Cette thèse constitue une contribution à la littérature en améliorant notre compréhension des aspects comparatifs des diverses méthodes statistiques dans le contexte des études de caractérisation de sites ainsi qu’en proposant une uniformisation des recommandations concernant l’utilisation de ces méthodes. Elle s'annonce, par conséquent, très prometteuse en tant que ligne directrices à suivre pour les chercheurs, les spécialistes et les décideurs.

Mots-clés: observations censurées à gauche, caractérisation de sites, estimation du maximum de vraisemblance, régression sur les statistiques d'ordre, Kaplan-Meier, modèles à effets mixtes

QUANTITATIVE ANALYSIS OF LEFT-CENSORED CONCENTRATION DATA IN ENVIRONMENTAL SITE CHARACTERIZATION

Niloofar SHOARI

ABSTRACT

A key component of site characterization is the statistical analysis of contaminant concentrations in soil, water and air samples. Such analysis can pose challenges due to the presence of nondetects or left-censored observations, which are measurements smaller than a detection limit. Censored values should be accounted for because they do not imply the absence of contamination, but the insufficient accuracy of the measuring instruments. Indeed, trace levels of hazardous pollutants can pose risks to the human health and the environment. Even if an environmental investigation achieves a representative sample of concentration data according to sound analytical protocols and data validation procedures, improper statistical analyses that do not properly accommodate censored observations may not represent actual site conditions. Obviously, remedial designs based on a distorted view of the contamination condition could be ineffective and not sustainable environmentally and economically.

The main goal of this research is to scrutinize the impact of left-censored values on site characterization outcomes. To this end, we explore different statistical methods (i) to estimate descriptive statistics, (ii) to quantify uncertainty around estimates, and (iii) to examine potential dependencies across observations due to clustering as an inherent part of sampling techniques. Substituting censored values with an arbitrarily selected constant is commonly practiced by both practitioners and researchers. In contrast, there are a number of parametric and non-parametric methods that can be used to draw inferences from censored data, and therefore, provide a more realistic insight into a contamination problem. Parametric methods, such as maximum likelihood and regression-based procedures, estimate descriptive statistics through fitting a parametric distribution to data. Due to the right-skewed shape of concentration data, gamma, Weibull, and lognormal distributions are the most plausible parametric models, with the latter being the most commonly used in environmental studies. Non-parametric procedures such as the Kaplan-Meier method, however, do not require any distributional assumption.

This study employs a comprehensive data simulation exercise, in which the true underlying distribution is known, to evaluate the performance of parametric and non-parametric estimators based on a large number of scenarios differing in censoring percent, sample size, and data skewness. This research also highlights the importance of investigating the robustness of parametric methods against model misspecifications. Using simulated data, we elucidate how substituting censored observations provides biased estimates and why it should be avoided even for data with a small percentage of censoring. We found that the maximum likelihood method based on the lognormality assumption is highly sensitive to data skewness, sample size, and censoring percentage. While the lognormal maximum likelihood method is mainly used in environmental studies, our findings point out that caution should be exercised

in assuming a lognormal density distribution of data. Instead, we recommend the maximum likelihood estimator based on a gamma distribution, regression-based methods (using either a lognormal or gamma distribution), and the Kaplan-Meier technique. With respect to quantifying the uncertainty around estimates for real concentration data, in which the true structure of data is unknown, we evaluate the performance of parametric and non-parametric estimators employing a bootstrapping technique. The conclusions drawn from bootstrapping of real data are in accordance with those inferred from the simulated data.

An important part of this research investigates the presence of correlation, associated with sampling techniques, among concentration observations. We provide statistical and conceptual backgrounds as well as motivations for mixed effects models that are able to accommodate dependence across data points while accounting for censored observations. Standard statistical methods assume that samples of concentration data are independent. However, in environmental site characterization studies, this assumption is likely to be violated because concentration observations collected, for example, from the same borehole are presumably correlated. This can in turn affect sample size determination procedures. We therefore employ a mixed effects model to capture potential dependencies and between group variability in data. The relevance of the estimated between-borehole variability is explained in terms of determining the optimal number of boreholes as well as samples to be collected from each borehole. Our proposed mixed effects model also provides insights into the vertical extent of contamination that can be useful in designing remediation strategies.

The findings of this doctoral research help increase the awareness of the scientific community as well as practitioners, exposure assessors, and policy-makers about the importance of censored observations. Aiming at unification of the field, this thesis contributes to literature by improving our understanding of the comparative aspects of different statistical methods in the context of site characterization studies. It thus offers considerable promise as a guideline to researchers, practitioners, and decision-makers.

Keywords: left-censored observations, site characterization, maximum likelihood estimation, regression on order statistics, Kaplan-Meier, mixed effects model

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABREVIATIONS

95UCL	95% Upper Confidence Level
AD	Anderson-Darling
AIC	Akaike Information Criterion
BIC	Bayesian Iinformation Criterion
cdf	Cumulative Distribution Function
CvM	Cramer-von Mises
DL	Detection Limit
ECDF	Empirical Cumulative Distribution Function-based
EM	Expectation Maximization
GM	Geometric Mean
GROS	Gamma Regression on Order Statistics
GSD	Geometric Standard Deviation
GTC	Gestion des Terrains Contaminés
HP	Hollander and Proschan
IBC	Intra-Borehole Correlation
KM	Kaplan-Meier
KS	Kolmogorov-Smirnov
LOB	Limit Of Blank
LOD	Limit Of Detection
LOQ	Limit Of Quantification
MGF	Moment Generating Function

MLE	Maximum Likelihood Estimation
MLEs	Maximum Likelihood Estimates
MSE	Mean Square Error
pdf	Probability Density Function
PPWM	Partial Probability Weighted Moments
Q-Q	Quantile-Quantile plot
RPRT	Règlement sur la Protection et la Rehabilitation des Terrains
rMLE	robust Maximum Likelihood Estimation
ROS	Regression on Order Statistics
rROS	robust Regression on Order Statistics
Subs	Substitution

INTRODUCTION

To date, the Federal Contaminated Sites Inventory has listed over 22,000 contaminated or suspected contaminated sites from which 2,393 are located in Quebec. According to Quebec contaminated sites inventory, Système de gestion des terrains contaminés (GTC), 8,334 sites had been registered in the system in 2010. These sites are not just in remote areas. In Montreal, for example, 1,617 sites have been identified (Hébert & Bernard, 2013). A sustainable revitalization of contaminated sites requires a comprehensive characterization followed by the adoption of appropriate remediation technologies. Within this context, the main goal of a site characterization study is to determine the type, concentration, location and extent of contamination. To this end, Quebec guidance on site characterization (Ministère du Développement durable, de l'Environnement, de la Faune et des Parcs du Québec, 2003a) recommends following the three steps below.

Phase I preliminary site characterization includes review of present and historical records, site visits, interviews, and identifying potential areas of contamination. If information obtained indicate any contamination evidence, phase II should be performed.

Phase II preliminary site characterization includes collection of field samples and analyzing them to confirm the nature as well as the horizontal and vertical extent of contamination.

Phase III exhaustive site characterization incorporates a series of actions for a more detailed characterization of a contaminated site if the result of phase II confirm the presence of pollution. These actions include further delineation of the impacted area, determining the volumes of contaminated material, and evaluating potential risks for human health and the environment.

Phase II and III always involve collection and chemical analysis of samples for contaminants concentrations. Given that the resultant concentration data sets are representative of site conditions, statistical analysis is used to decide whether or not the site is polluted and should undergo some remediation actions. Estimating descriptive statistics is the most important application of statistical analysis since they are employed in other statistical procedures;

some applications include quantifying the potential impact on human health and the environment, monitoring compliance with environmental standards, and devising/refining sampling strategy. Another important statistical analysis in site characterization can be the study of the association between contaminants and selected soil properties. Or, in the case that a human health risk assessment process is incorporated into the site characterization study, the relationship between the pollution and their adverse effects on health is of interest. Other more sophisticated applications of statistics include principal component analysis and identifying spatial and temporal patterns of contamination.

Even with technical advances in chemical analysis protocols and laboratory instrumentations, there remains a threshold below which contaminants concentrations are not precisely quantifiable. These concentrations are called left-censored (equivalently nondetects) and present a serious challenge in data analysis. The problem exacerbates when environmental scientists substitute censored observations with arbitrary constants before carrying out any statistical analysis. Helsel (2006) refers to substitution of nondetects as a "data fabrication" method because those measurements that are considered as highly unreliable are then treated as actually observed values. This approach diminishes data representativeness and provides biased results, potentially compromising human health and the environment and causing financial losses. However, substitution of censored values is commonly practiced because, as said by Helsel (2010a), "there is an incredibly strong pull for doing something simple and cheap."

Alternatively, researchers have exploited methodologies from survival analysis, which were originally developed for right-censored medical data. From the estimation point of view, the alternative methods to deal with left-censored data fall into two categories:

a) Parametric methods that fit a distribution to data through maximum likelihood or probability plotting. The estimates obtained from the maximum likelihood method (MLE) are those that maximize a likelihood function, which is a product of the probability density function (pdf) when an observation is detected and the cumulative distribution function (cdf) when an observation is censored. On the other hand, the

most popular probability plotting-based method is the regression on order statistics (ROS) which involves fitting a regression line to data on a probability plot. A novel extension of the ROS technique is based on a gamma assumption and is called gamma regression on order statistics (GROS). In the case that one of the above parametric methods is used to impute values for censored observations, the robust versions of MLE and ROS (rMLE and rROS, respectively) are obtained. Since the statistical distribution of concentration data is typically right-skewed, a lognormal distribution is often fitted to data. However, other similar distributions (such as Weibull and gamma) are occasionally encountered;

b) Non-parametric methods such as Kaplan-Meier (KM), which does not require any distributional assumption and uses only data ranks.

A number of simulation studies have been devised to assess the merits of these alternatives to substitution, but their sometimes contradictory conclusions still rule out recommending a single method as the preferred approach. This is the main reason for which nondetects are still substituted with arbitrary constants despite the fact that numerous publications provide recommendations against it. The findings of this doctoral research help increase the awareness of the scientific community as well as practitioners, exposure assessors, and policy-makers about the importance and benefits of considering censored data as such in quantitative analysis. Using data simulations and real data analysis, this thesis investigates the impact of left-censored values on different aspects of contaminated sites characterization. In particular, the focus has been on appropriate strategies to (i) estimate descriptive statistics and associated uncertainty, and (ii) to model dependency in concentration observations coming from the same borehole. Overall, this thesis illustrates best practices to handle left-censored concentration data that should be incorporated in the environmental policies and procedures toward a sustainable characterization of contaminated sites.

Characterization data with left-censored observations

Before proceeding with a detailed discussion on the impact of left-censored data, we define three key relevant terms: limit of blank (LOB), limit of detection (LOD), and limit of quantification (LOQ). This terminology is adopted from the Clinical and Laboratory Standard Institute (2004). The LOB is the highest expected concentration of a chemical when replicates of a blank sample are measured. The LOD is the minimum concentration of a chemical that can be distinguished from the absence of the chemical with a stated confidence limit. The LOD is estimated by preparing and analyzing a series of blank samples and using the mean and standard deviation of the replicates with some confidence factor. Instead of the LOD, some laboratories use the LOQ to report their analysis results. The LOQ is the lowest concentration at which the chemical can be reliably quantified. Throughout this thesis, we use the general term detection limit (DL) to refer to LOD or LOQ. Those concentration measurements below the DL are called left-censored or nondetects.

Two types of censoring are encountered: in type I censoring, which is the typical situation of environmental data, the censoring point is known (this is the DL in chemical analytical practice) and the number of censored data is random. In type II censoring, on the other hand, the number of censored observations is fixed in advance and the censoring point is a random variable. Type II censoring typically occurs in life-testing and reliability investigations.

Significance of left-censored data

Although a left-censored observation does not report an exact value of a chemical concentration, it still contains the information that the measurement falls somewhere between zero and DL. Considering the efforts and expenses dedicated to environmental data collection and analysis, it seems worthwhile to investigate more sophisticated statistical methods in order to extract the maximum amount of reliable information from left-censored data. It is crucial to acknowledge that left-censored concentrations do not necessarily insinuate the absence of contamination; rather they indicate that the precision of the

analytical instrument was too low to reliably quantify a concentration value. The importance of accounting for left-censored concentrations is highlighted when dealing with historical concentration data, where analytical instruments were still less powerful and DLs were higher. In addition, in the case of highly toxic contaminants such as dioxins and arsenic, even trace levels may pose risks to human health and the environment.

A wide range of management decisions can also be affected by left-censored data. In environmental studies, left-censored data impact not only the estimation of statistical parameters, but also the characterization of data distributions, inferential statistics (e.g., comparing the mean of two or more populations) (Finkelstein, 2008; Antweiler, 2015), the determination of correlation coefficients, the construction of regression models (Lynn, 2001; Schisterman, Vexler, Whitcomb & Liu, 2006). In addition to the environmental sciences, handling left-censored data has been a challenge in astronomy (Feigelson & Babu, 2012), occupational health (Succop, Clark, Chen & Galke, 2004; Hewett & Ganser, 2007), and food health (European Food Safety Authority, 2010).

CHAPTER 1

RESEARCH FOCUS AND OBJECTIVES

1.1 Objectives

The main objective of this PhD thesis is to address the issues associated with the presence of left-censored concentrations, which is a pervasive problem in environmental research. Within the context of characterization of contaminated soils, the specific objectives focus on two important aspects of statistical inferences. The first aspect is to identify appropriate strategies for estimating descriptive statistics of a soil population; these estimates are employed in decision-making process (e.g., compliance with a regulatory standard) or in improving the precision of a characterization study (e.g., determining the sample size). The second aspect highlights the importance of accounting for dependency among concentration observations while left-censored values are accommodated. Statistical analyses throughout this dissertation focus on quantifying the bias resulting from the substitution of left-censored observations with arbitrary constants. As substituting is a common approach to deal with left-censored concentration data among practitioners and researchers, we are interested in understanding and comparing the consequences of a characterization study when the substitution or alternative techniques are employed.

1.2 Synopsis

1.2.1 Evaluating the performance of different estimators based on simulated censored data (Chapters 3&4)

Despite proliferation of simulation studies that compare different statistical methods for analyzing censored data, yet there is a need for further investigations because of the following concerns:

a) Most previous simulation studies overlooked exploring the impact of data distribution skewness on the performance of the estimators under study. Indeed, failing to accounting for a wide range of data skewness might have led to the lack of general agreement between different studies. In fact, as mentioned by Singh, Maichle & Lee (2006), simulation results derived for low skewed data cannot be generalized for highly skewed data;
b) In previous simulations, artificial data were mainly generated from normal and lognormal distributions; and consequently, the parametric estimation methods (e.g., MLE) relied on these distributions. Given that no theoretical study supports the assumption that environmental concentration data are normally or lognormally distributed, there is a need for a comprehensive simulation framework that encompasses other distributions and explores the robustness of estimators against distribution misspecifications;
c) Previous simulation studies discouraged the substitution of censored values due to the lack of a theoretical basis. However, some of these studies report simulation scenarios where the performance of the substitution approach equals that of other alternative methods. Therefore, it is useful to understand reasons for which substitution may or may not result in biased estimates.

Given the above aspects, the main objective of chapter 3 is to investigate the properties of alternative statistical methods that can handle left-censored data. To this end, we design a comprehensive simulation study that compares the performance of the MLE, rROS, GROS, and KM estimators under different scenarios of percentage of censoring, sample size, and data skewness. In addition, this simulation study evaluates the robustness of the parametric methods (i.e., MLE, rROS, and GROS) to distributional misspecification. According to our simulations, the MLE method based on lognormal and Weibull distributions provides inflated estimates of the mean and standard deviation when data distribution is highly skewed and censoring percent is large. Relating to sample size, although current literature indicates that 50 observations are sufficient to guarantee reliable MLEs, our simulations show that more than 50 observations might be required in the case that the distribution is highly skewed. Among other finding, this chapter demonstrates that the methods of MLE (using gamma assumption), rROS, GROS, and KM should be considered for estimating descriptive statistics

of censored environmental data sets because of their robustness against distributional assumptions, censoring percent, and skewness.

The simulation study reported in chapter 4 of this thesis discusses inherent problems associated with the substitution of censored observations, the most commonly practiced approach. We illustrate that the performance of the substitution approach varies according to the population's distributional characteristics (such as coefficient of variation and skewness) that are unknown a priori. For the same reason, substitution of censored observations should be avoided even when the censoring percent is as low as 10%. For a general overview of the simulation framework used in chapters 3 and 4, Table 1.1 reports a summary of different parameters of the simulation study.

Table 1.1 An overview of the different parameters used in the simulation study

Data generating distributions	Lognormal Weibull Gamma Mixture lognormal Mixture Weibull Mixture gamma
True values of the mean and standard deviation	$\mu = 1,2,3,\ldots,10$ $\sigma = 0.5, 1.2, 1.9, 2.6, 3.3, 4$
Sample size	60, 120, 180, 240, 300, 360
Censoring percent	10%, 30%, 50%, 70%
Statistical methods	Substitution with DL/2 Maximum likelihood estimation (MLE) • Lognormal • Gamma • Weibull Robust regression on order statistics (rROS) Gamma regression on order statistics (GROS) Kaplan-Meier (KM)

1.2.2 Quantifying uncertainty of different estimators through bootstrapping (Chapter 5)

The results reported in chapter 4 showed that the substitution approach is not reliable for computing the mean and standard deviation of data when left-censored observations are encountered. With respect to alternative estimation techniques (i.e., MLE, rROS, GROS, and KM), some amount of uncertainty is always associated with the estimates. This uncertainty arises from the presence of left-censored concentrations as we do not have any knowledge regarding the true value of left-censored measurements.

We use a bootstrapping technique to provide uncertainty information along with the estimates of the mean and standard deviation obtained from the aforementioned alternative estimators. Unlike the analyses discussed in chapter 3 and 4 that were based on computer-generated data, the adopted methodology in chapter 5 allows making inferences based on real concentration data. Concentration data sets used in this research are from chemical analysis of soil samples collected for characterizing a brownfield site in Montreal, Canada.

We assume, as other bootstrapping applications, that the concentration data at hand is a representative sample of a soil population. The idea behind the bootstrapping is to take repeated draws with replacement from the actual concentration data and treat these draws (bootstrap samples) as possible random samples that could have been taken in the real world. Using the MLE, rROS, GROS, and KM estimators, we compute the statistics of interest (the mean and standard deviation in this thesis) for each bootstrap replicate. This yields an approximation to the distribution of the statistics provided by a given estimator that is used to calculate the uncertainty of that estimator in terms of confidence intervals. The abovementioned procedure is a non-parametric bootstrapping technique, which avoids making unnecessary assumptions about the distribution of concentration data.

The conclusions drawn from bootstrapping of real data are in accordance with those inferred from the simulated data. In general, the MLE method using the lognormal and Weibull distributional assumptions leads to the highest levels of uncertainty whereas the MLE under

gamma assumption, rROS, GROS, and KM produce less uncertainty. Moreover, the rROS, GROS, and KM estimators have small approximate biases. Calculating the mean and its 95% upper confidence level of real contaminant concentration data, we demonstrate that adopting an inappropriate statistical method results in imprecise estimates, which contribute to the global uncertainty in the outcomes.

1.2.3 Accounting for dependence in data in presence of left-censored concentrations (Chapter 6)

In this chapter we discuss that sampling strategies in environmental site characterizations result in concentration data with a nested structure. Under this aspect, observations are generated from different groupings in data, so that those nested in the same borehole may share similar traits. In fact, it is quite plausible to postulate that concentration measurements obtained from the same borehole are likely to be correlated due to some unmeasured known or unknown factors. Employing standard approaches, for which independence assumption is crucial, to analyze such data leads to unfounded conclusions. To tackle this issue, while accommodating left-censored observations, we propose a mixed effects model that accounts for data dependencies. It is thus possible to estimate between-borehole variability. In addition, we set the proposed model in a way that allows us to estimate the mean value of concentration of a given contaminant at different depths or type of material constituting the brownfield site.

A major implication of the adopted approach in the context of site characterization studies relates to determination of optimal sample size in terms of the number of required boreholes as well as the number of required samples per borehole. It should be highlighted that the current practice does not follow statistical approaches. Moreover, this chapter examines the vertical extent of contamination that can be useful in defining the remediation depth.

1.2.4 List of manuscripts

This dissertation includes 4 published manuscripts. Figure 1.1 presents the manuscripts and also their main findings. The manuscripts are listed as follows.

Manuscript (1): Shoari, Niloofar, Jean-Sebastien Dubé and Shoja'eddin Chenouri. (2015). Estimating the mean and standard deviation of environmental data with below detection limit observations: Considering highly skewed data and model misspecification. *Chemosphere*, *138*, 599-608.

Manuscript (2): Shoari, Niloofar, Jean-Sébastien Dubé and Shoja'eddin Chenouri. (2016). On the use of the substitution method in left-censored environmental data. *Human & ecological risk assessment*, *22* (2), 435-446.

Manuscript (3): Shoari, Niloofar and Jean-Sébastien Dubé, (2016). An investigation of the impact of left-censored soil contamination data on the uncertainty of descriptive statistical parameters». *Environmental Toxicology and Chemistry*. *35* (10), 2623-2631.

Manuscript (4): Shoari, Niloofar and Jean-Sébastien Dubé, (2017). Application of mixed effects models for characterizing contaminated sites. *Chemosphere*. *166*, 380-388.

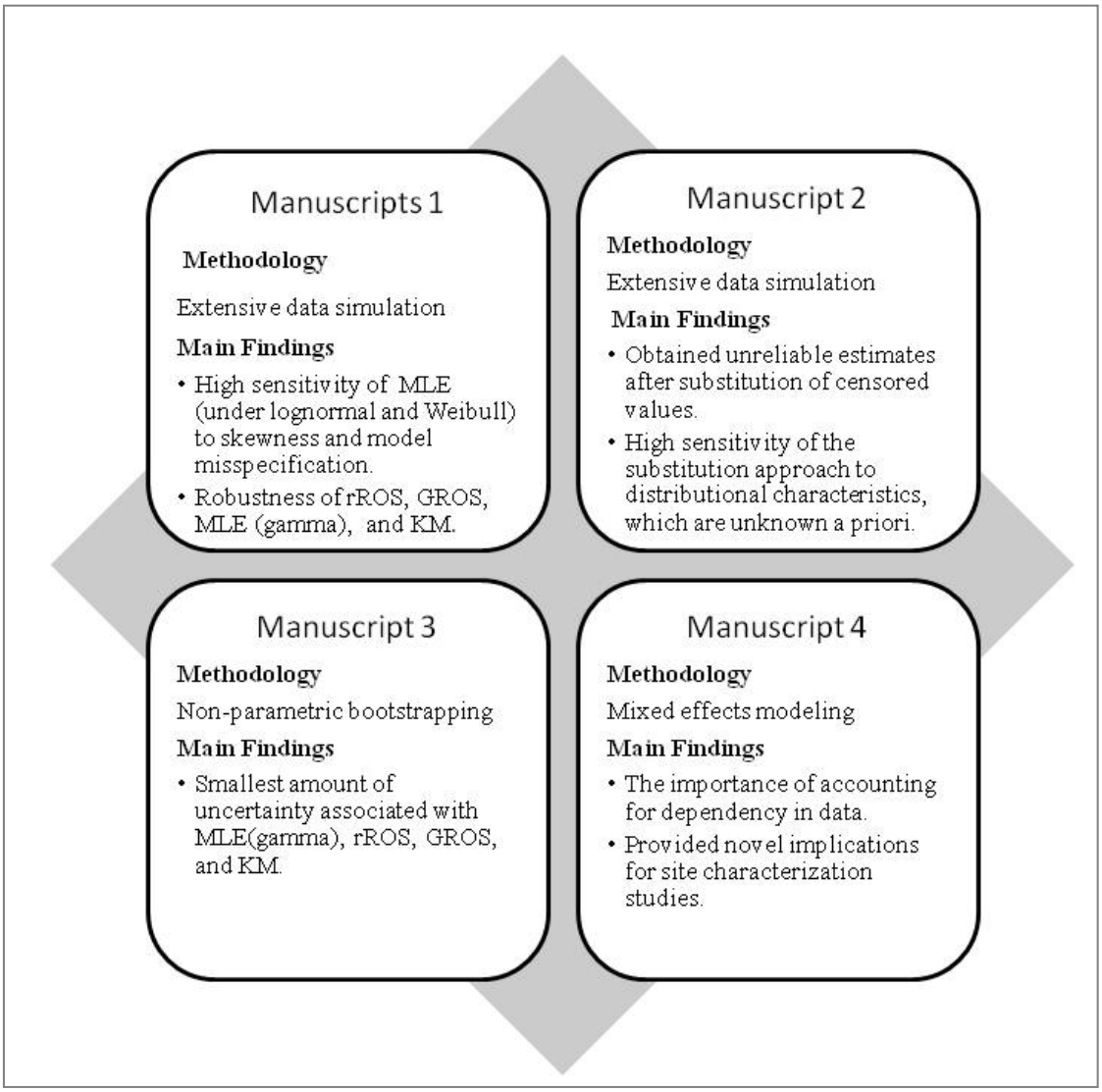


Figure 1.1 Manuscripts and their main findings

CHAPTER 2

LITERATURE REVIEW

The literature review of this thesis provides an overview of statistical methods that can be used for interpreting data containing left-censored observations. This chapter consists of two parts. The first part reviews publications that have focused on identifying appropriate strategies to accurately estimate the statistical parameters of left-censored data. The related concepts and mathematical formulations for different statistical methods are explained in Appendix I of the thesis. Moreover, major articles are organized in chronological order (Table 2.2) to provide a perspective on the developments over the past 30 years. In addition to parameter estimation, environmental studies may require performing regression analyses on censored data in order to investigate the relationship between a response variable (e.g., arsenic concentration in soil) and one or more explanatory variables (e.g., soil type). In this regard, the second part of the literature review gathers all studies that developed regression models while accounting for left-censored data.

2.1 Parameter estimation of left-censored data

Many publications use Monte Carlo experiments to explore and compare the performance of substitution with alternative estimators. Some relevant studies include Gilliom & Helsel (1986); Helsel & Cohn (1988); Newman, Dixon, Looney & Pinder (1989); She (1997); Singh & Nocerino (2002); Lubin et al. (2004), Hewett & Ganser (2007). All of the above studies share the same research design: Let θ be the true distributional parameter (e.g., mean or standard deviation) of the reference population, from which artificial data of size *n* were generated. Note that normal and lognormal distributions were typical in simulation studies. For a given censoring percentage, say c%, a censoring point was imposed at the c^{th} percentile of the reference population. To be precise, for data sets generated from distribution $F(\mu, \sigma^2)$, the censoring point was calculated as $F^{-1}(c;\ \mu, \sigma^2)$, where $F^{-1}(.)$ is the inverse cumulative distribution function. Within a set of simulations, substitution and alternative methods were

used to estimate $\hat{\theta}$, where $\hat{\theta}$ was the estimated statistical parameter of the simulated data. To investigate and compare the ability of different estimators in reproducing the correct values for θ, bias and/or mean square error (MSE) were utilized as the comparison criteria. The common conclusion of the published literature was that substituting censored data introduced an estimation bias and it should be avoided.

El-Shaarawi & Esterby (1992) provided analytical expressions for quantifying the bias due to substitution of censored values. However, the application of these expressions is limited because (i) they are valid only for normally and lognormally distributed data; and (ii) they require knowledge about the proportion of censoring, mean, and variance, which are usually unknown. Helsel (2005, 2006, and 2010b) consistently emphasized the unreliability of substitution and discussed how it would provide poor estimates for different statistical analyses (i.e., the mean, standard deviation, t-value, correlation coefficient, regression slope, p-value, etc.). Nevertheless, substitution remains a common practice in environmental studies (e.g., Farnham, Singh, Stetzenbach & Johannesson, 2002; Krapac et al., 2002; Sapkota, Heidler & Halden, 2007; Schäfer, Paschke, Vrana, Mueller & Liess, 2008; Higley, 2010; Hsu, Guo, Wang, Liaoand & Liao, 2011; Jones, 2011; Vassura, Passarini, Ferroni, Bernardi & Morselli, 2011; Watkins et al., 2016).

A few studies suggested the use of substitution of censored observations. Hornung & Reed (1990) suggested the substitution method whenever less than 50% of data were nondetects. Clarke (1998) advocated substitution of a constant rather than the MLE and ROS methods when data sets were small (with less than 10 observations). The failure of the parametric methods might have been due to small sample size because distributional properties could not be accurately established with only a few observations. In a comparative Monte Carlo simulation study, She (1997) reported that the estimates obtained after substituting censored data with DL/2 were sometimes as good as those provided by the KM estimator. Also, Hewett & Ganser (2007) reported simulation scenarios where substitution was recognized as the estimation method of choice. Although some studies reported good agreement between

the substitution and alternative methods, Leith et al. (2010) cautioned that this behavior should not be interpreted as evidence of equivalency between these methods.

Rather than using the simplistic substitution technique, researchers developed parametric procedures that use the observed values in combination with the information contained in the censored part. The two main categories of parametric procedures are based on maximum likelihood and probability plotting. The maximum likelihood methods can be traced back to the work of Cohen (1959; 1961) who developed a version of MLE that relied on look-up tables to estimate statistical parameters of censored data. The Cohen's MLE method has a drawback of being restricted to normally distributed data that contain a single DL, whereas concentration data are typically skewed and contain multiple DLs. Gilliom & Helsel (1986) considered estimating statistical parameters of singly censored (with only one DL) water quality data and conducted a comparative simulation study to compare the performance of substitution, MLE, and probability plotting procedures. Assuming that environmental data are lognormally distributed, their simulations suggested the MLE as the best estimator of different percentiles. However, the performance of the MLE method was not satisfactory for estimating the mean and standard deviation. Later, in a related study, Helsel & Cohn (1988) extended the work of Gilliom & Helsel (1986) and investigated the effect of the presence of multiple DLs on the performance of different estimators.

The MLE method for multiply censored and normally distributed data (or approximately normally distributed after log-transformation) was discussed by El-Shaarawi & Naderi (1991). Instead of using Cohen's look-up tables, they developed likelihood functions needed to estimate the mean and standard deviation of data. To employ the MLE method under the normality assumption, Shumway, Azari & Kayhanian (2002) suggested using a Box-Cox transformation to generate approximately normal data. The first problem with transformation is the transformation bias, which occurs when the estimates are back-transformed to the original scale (Helsel, 1990). To resolve this issue, Shumway et al. (2002) employed the Quenouille-Tukey Jackknife to improve the quality of estimates and to compensate for the transformation bias. The second problem is the ambiguity as to which transformation (e.g.,

logarithmic, square root, inverse, and arcsine) should be used. While most previous simulation studies used the MLE with lognormal assumption for lognormal or contaminated lognormal data generating distributions, European Food Safety Authority (2010) demonstrated the adequacy of the MLE method under Weibull and gamma assumption when applied to lognormal data and vice versa.

Hewett & Ganser (2007) considered a comprehensive simulation study that aimed at identifying an "omnibus" method for estimating the mean and 95th percentile of exposure data sets containing nondetects. Their study incorporated several simulation scenarios using computer-generated data from lognormal and contaminated[1] lognormal distributions with different censoring percentages. The estimation methods examined were substitution, several variations on the MLE and ROS, non-parametric quantile and KM. No single method showed superiority across all simulation scenarios although the MLE-based techniques generally performed well. However, their study did not address confidence intervals.

Despite numerous researchers tended to favor the MLE method, the results of some investigations (for instance, Lee & Helsel, 2007 and Jain & Wang, 2008) indicated the limited ability of this method when dealing with small data sets and large censoring percent. These investigations agreed that the MLE method may not show some of its desirable properties (consistency, efficiency, and asymptotic normality) for small data sets (with <50 uncensored values as reported in Helsel, 2005) with large amount of censoring (Helsel, 2012).

A parametric method based on probability plotting was discussed in Travis & Land (1990). This method assumes that observations (or log-transformed observations) below and above the DL are normally distributed. It fits a regression line on probability plot of data and the intercept and slope of this line provide the estimates of the mean and standard deviation,

[1] A contaminated lognormal distribution is a combination of two or more lognormal distributions.

respectively. Although censored observations are treated as unknown, their percentile values are accounted for. This method is commonly known as regression on order statistics (ROS).

An extension of the above mentioned fully parametric methods (i.e., MLE and ROS) are referred to as "imputation" or "robust" procedures, in the hope to have estimators that are both reliable and easier to implement. In these methods the observation above the DL are combined with imputed values for observation below the DL and thus standard statistical methods can be applied. The censored observations are imputed using some initial estimates obtained from MLE or ROS. Two popular examples include robust MLE (rMLE) proposed by Kroll & Stedinger (1996) and robust ROS (rROS) developed by Helsel & Cohn (1988). Hewett & Ganser (2007) discussed that the rMLE and rROS slightly outperform their fully parametric counterparts. For example, when dealing with small data sets or when data distribution does not exactly match the assumed distribution, the rROS approach outperforms the MLE. Specific applications of the rROS was reported by Baccarelli et al. (2005) to estimate mean levels of dioxin in marine water samples and by Röösli et al. (2008) to estimate the mean and different quantiles of radiofrequency measurements subject to censoring.

The advantages of robust procedures were reported in Huybrechts, Thas, Dewulfand & Van Langenhove (2002) where they identified two problems associated with fully parametric methods: First, the presence of outliers may falsify the lognormality assumption and may results in highly biased estimates. Secondly, even if data happens to be lognormally distributed, the estimates of the mean and standard deviation suffer from the back-transformation bias. They pointed out that robust parametric methods are not very sensitive to departures from the assumed distribution. Moreover, the censored observations are predicted and can be directly back-transformed to original scale avoiding the transformation bias.

In addition to the rROS and rMLE, the environmental literature reports other ad-hoc imputation techniques that are less frequently used. For data X censored at DL, the imputed

values are generally conditional expected values of X, given nondetects are smaller than DL, that is $E[X|X < DL]$. Some relevant papers in this regard are listed below:

- Lynn (2001) developed an imputation technique based on maximum likelihood. The imputed values were random draws from a normal distribution whose parameters were maximum likelihood estimates (MLEs);
- Succop et al. (2004) employed the MLE method to derive initial estimates of the sample mean and standard deviation. These estimates were used to impute censored observations, which they called "the most provable value";
- Lubin et al. (2004) carried out a multiple imputation procedure where Tobit regression followed by a non-parametric bootstrapping was used to estimate the mean and standard deviation of lognormally distributed data. These estimates were used to construct a lognormal distribution and the imputed values were random draws from that distribution. Comparing the performance of Lubin's against Lynn's imputation method, Jain et al. (2008) showed the superiority of the Lubin's method for censoring percent larger than 20%;
- Aboueissa & Stoline (2004) proposed a new imputation technique that performed as well as the MLE method. Their method employed information regarding the number of observations below and above DL as well as the estimates of the mean and standard deviation of the uncensored part of the data. However, the application of their methodology is limited to low skewed normal and lognormal data subjected to a single DL with a censoring percent of less than 50%;
- Krishnamoorthy, Mallick and Mathew (2009) proposed a methodology valid for data that can be represented by a normal distribution. An appropriate transformation such as lognormal or cube root transformation may be necessary to be able to employ their technique. This imputation method uses initial estimates of the mean and standard deviation based on the uncensored data, with some adjustments to compensate for parameter overestimation since only above-DL values are considered. Through simulation studies and real data examples, the authors demonstrated that their imputation technique worked well for small to moderately large sample sizes;

- Ganser & Hewett (2010) proposed the β-substitution method which consists of substituting censored observations with a data-dependent β factor multiplied by the DL. They demonstrated that their proposed substitution technique performed equally well when compared to the MLE method, particularly in simulation scenarios with small sample sizes. They did not include KM when comparing the performance of their new method. To complement this study, Huynh et al. (2014) devised a simulation framework to evaluate the performance of β-substitution against its competitors, MLE and KM. They concluded that the β-substitution method performed as well as or better than MLE and KM methods for data from lognormal and contaminated lognormal distributions. When data contain multiple DLs, this methodology suffers from a drawback in that the average of DLs is considered in the algorithm as if the data had a single DL.

Another promising estimation method falls under the category of non-parametric techniques that do not require any parametric assumption about the data; all that matters is the relative rank of observations. As the distribution of left-censored concentration data is complex and often unknown, She (1997) favored the non-parametric KM estimator. For estimating the 95% upper confidence level (95UCL) of censored concentration data sets, the simulation study by Singh et al. (2006) considered the impact of censoring percent as well as the degree of skewness on the performance of different statistical methods. Overall, they advocated the KM estimator on the basis that parametric methods relying on lognormal distribution assumption resulted in unrealistically inflated estimates. Importantly, they noted that an estimation method may perform differently depending on whether data are low or highly skewed. Antweiler & Taylor (2008) questioned the reliability of research studies based on data generated from known distributions as it might have been caused the preference of parametric estimators. They used a more precise laboratory instrument and re-measured the contaminants concentrations of the samples that had previously provided censored data, making it possible to attribute a concentration value to censored measurements. This resulted in having two concentration data sets for each contaminant, one with censored observation and the other one without them. They applied the substitution (with zero, DL, and DL/2, a random number between zero and the DL), rROS, MLE (under normal and lognormal

assumption), and KM methods to estimate statistical parameters of data and compared these to the true values. They concluded that generally the KM method estimated the mean, standard deviation, 25^{th}, 50^{th}, and 75^{th} percentiles with less error. The simulation results of European Food Safety Authority (2010) also suggested the KM method when the underlying distribution of data was not easily identified and particularly, the censoring percentage was lower than 50%. Some sample applications of KM can be found in Pajek, Kubala-Kukuś, Banaś, Braziewicz & Majewska (2004); Helsel (2010b); and Barghi, Choi, Kwon, Lee & Chang (2016). However, the KM method is not recommended for data sets with only one DL and whenever the smallest observation is a nondetect (Hewett & Ganser, 2007).

Despite attempts of researchers to encourage the use of alternative estimators, we still encounter studies that avoid using them. The computational complexity of implementing alternative estimators is often one of the hurdles. However, increasing availability of standard software programs has resolved this problem. For example, the computation of the MLE method through Microsoft Excel Solver Tool was made available by Finkelstein & Verma (2001). Flynn (2010) presented an estimation technique that was also simply implemented in an Excel worksheet and claimed that the mean and standard deviation estimates provided by their methodology were comparable to those obtained from the restricted MLE method. This methodology imputes censored values by maximizing the Shapiro-Wilk statistic such that a normal distribution is produced. The rough assumption of this estimation technique is that data or transformed data follow a normal distribution. In two companion papers by Lee & Helsel (2005; 2007), S-language software implementations for the rROS and KM methods are explained.

The second reason that prevents using the alternative estimators is the lack of clarity as to what is the best course of action to take in the presence of left-censored data. In an attempt to unify the opinions, Helsel (2012) reviewed several papers on the performance of various methods for estimating statistical parameters of data and gave a concise summary of the results as reported in Table 2.1.

Table 2.1 Summary of recommended methods for estimating the statistical parameters [2]

	Sample size	
Censoring percent	**<50 observations**	**>50 observations**
<50%	Imputation or KM/Turnbull	Imputation or KM/Turnbull
50%-80%	rMLE, rROS, multiple imputation	MLE, multiple imputation
>80%	Report only % above a meaningful threshold	May report high sample percentiles (90^{th}, 95^{th})

In addition to employing appropriate estimation technique(s), it is crucial to identify adequate methods for constructing confidence intervals as these indicate the uncertainty in the estimates. However, the majority of the above-mentioned investigators did not address this issue. Assuming that data (or transformed data) follows a normal distribution, confidence intervals around the mean based on the Student's t-statistic are computed as $(\hat{\mu} - t_{\alpha/2,(n-1)}\sqrt{\hat{\sigma}^2/n}, \hat{\mu} + t_{(1-\alpha/2),(n-1)}\sqrt{\hat{\sigma}^2/n})$, where $\hat{\mu}$ and $\hat{\sigma}^2$ may be computed using any of the parametric estimators such as MLE or an extension of it. Singh et al. (2006) and Helsel (2012) discussed that these parametric intervals are highly sensitive to the normality assumption and if applied for skewed data sets, the estimated confidence intervals may be biased (and sometimes unrealistic, for example, in the case that negative lower confidence levels are estimated). Another shortcoming with this confidence interval is the lack of clarity about whether n represents the total number of observations, or only the number of uncensored observations. Bootstrapping (Efron, 1981) is a compelling method for computing confidence limits around the statistic of interest (mean, median, percentile, etc.). This method consists of sampling with replacement from the original data for *B* times and calculating the statistic of interest for each draw. Doing so, one obtains *B* estimates of the statistic (for example, mean), which are used to describe the probability distribution of that statistic. This

[2] Adopted from Helsel (2012), page 93

probability distribution serves as a basis for calculating the confidence interval. Being a non-parametric technique, bootstrapping has the advantage of not relying on the normality assumption of data. Frey & Zhao (2004) fit lognormal, Weibull, and gamma distributions using MLE to estimate the mean concentration of censored urban air toxics data. They also employed the bootstrap method to calculate the uncertainty around the estimated mean. Their paper showed that the range of uncertainty increased with increasing censoring percent and coefficient of variation (coefficient of variation was used as an indicator of data variability), and on the other hand, decreased when sample size got larger. Singh et al. (2006) considered several estimation methods including Tiku's method, Scheneider's approximate UCL method, Student t-statistic, Land's H-statistic, Chebyshev inequality, and different versions of bootstrapping to calculate the 95UCL of the mean. They concluded that the KM estimator followed by Chebyshev, student's t-statistic, or bootstrap provided good estimates of 95UCL. A review of prior publications that evaluated the performance of different estimators is reported in chronological order, as illustrated in Table 2.2.

2.1.1 Current norms of environmental agencies on censored data

The guidelines issued by USEPA (2000) advocate substitution of censored data by half of the DL when less than 15% of data is censored. However, Helsel (2006) states that the 15% cut-off value is simply based on judgment rather than any peer-reviewed publication. If 15%-50% of data are censored, USEPA (2000) recommends using the MLE, trimmed, or winsorized mean and standard deviation. For data sets with more than 50% censoring, a percentile larger than the censoring percent can be used, instead of the mean value, to represent contamination level. This guideline cautions practitioners when using the MLE method for small data sets ($n<20$), as it may produce biased results. Although not discussed in details, this document gives some recommendations on which statistical parameter to use for different censoring percent and coefficient of variation. Noticeable is that after 6 years, another document issued by USEPA (2006) incorporates the same elements of the prior guidance on how to handle censored data. The Appendix of the Local Limits Development Guidance (USEPA, 2004) recognizes that substitution of censored data results in biased

estimates and encourages the use of rROS and MLE techniques. In a report published by Oak Ridge National Laboratory, Frome & Wambach (2005) recommends MLE as the first method of choice and KM when the data distribution is hard to identify.

Canadian Federal and provincial government documents related to site characterization are strongly based on the above-mentioned USEPA guidelines for handling left-censored data. Within the context of risk-based site characterization, the guidance document "human health detailed quantitative risk assessment" provided by Health Canada (2010) accepts substitution for low censoring amounts; however, no threshold for censoring percent is reported. In addition, this document recommends the rROS method for modest to large data sets and the MLE method only for large data sets without giving any indicative value about the sample size. Similarly, Canadian Council of Ministers of the Environment (2016) suggests the use of substitution as long as censoring percent is less than 10%. For higher censoring percent, this guidance recommends one of the MLE, rROS, or KM estimation techniques. Surprisingly, the problem of left-censoring has not been mentioned in "Guide de caractérisation des terrains" (Ministère du Développement durable, de l'Environnement, de la Faune et des Parcs du Québec, 2003a).

Table 2.2 Review of prior publications (in chronological order) on the performance of estimators for left-censored data

year	Reference	Methodology	Estimators	Considered distributions	Distributional parameters	Preferred estimators
1986	Gilliom & Helsel	Monte Carlo simulations	subs, ROS, MLE	lognormal, delta, contaminated lognormal, gamma	μ=1 σ=0.25, 0.5, 1, 2	ROS for estimating percentiles; MLE for estimating the mean and standard deviation
1988	Helsel & Cohn	Monte Carlo simulations	subs, ROS, rROS, MLE	lognormal, delta, contaminated lognormal, gamma	μ=1 σ=0.25, 0.5, 1, 2	rROS
1989	Newman et al.	3 water quality data sets subject to artificial censoring	subs, ROS, MLE, restricted MLE, bias-corrected MLE	Normal, Lognormal	μ=18.3, σ=3.83; μ =997, σ=19.9; μ= 5.21, σ=2.16	Restricted MLE when the underlying distribution is known; ROS when the underlying distribution cannot be identified
1990	Hass & Scheff	Monte Carlo simulations	subs, ROS, rROS, Cohen's MLE, restricted MLE, bias corrected MLE	normal contaminated normal	μ=0,σ=1; μ=(1,-1),σ=1; μ=0,σ=1,5	Bias-corrected and restricted MLE
1992	El-Shaarawi & Esterby	Developed analytical expressions	-	normal lognormal	μ=1,2,3,…,10 σ=1	Quantifying the bias of substitution depends on a variety of parameters; MLE is preferred.
1996	Kroll & Stedinger	Monte Carlo simulations	MLE, rMLE, ROS, PPWM	contaminated lognormal, gamma, lognormal, Delta, Weibull, log-Pearson III	μ=1 σ=0.25, 0.5, 1, 2	rMLE

Year	Reference	Methodology	Estimators	Considered distributions	Distributional parameters	Preferred estimators
1997	She	Monte Carlo simulations	subs, ROS, MLE,KM	lognormal, gamma	μ=1 σ=0.25,0.5,1,2	KM
1998	Clarke	Monte Carlo simulations	subs, ROS, MLE	normal , lognormal, gamma	μ=1 σ=0.1,0.5,1,2	Subs
2002	Shumway et al.	Monte Carlo simulations	rROS, MLE	lognormal, gamma	log: (μ=2.77,σ=0.75) gamma:(μ=4 ,σ=2.83)	MLE followed by Jackknife to reduce transformation bias
2002	Singh & Nocerino	Monte Carlo simulations	subs, rROS EM algorithm, Cohen's MLE, restricted MLE, unbiased MLE	normal	μ=5, σ=2	restricted MLE
2002	Huybrechts et al.	3 water quality data sets subject to artificial censoring	Cohen's MLE, bias-corrected and restricted MLE, ROS, rROS	lognormal	μ=12.9, σ=17.8; μ=26.7, σ=35; μ=199, σ=473	robust bias-corrected and restricted MLE; rROS
2005	Baccarelli et al.	Dioxin data	subs, MLE, ROS, rROS	lognormal	-	rROS
2006	Singh et al.	Monte Carlo simulations	subs, MLE, bias-corrected MLE, restricted MLE, EM algorithm, delta method, ROS, rROS, KM, winsorization	normal, lognormal, gamma	Normal:(μ=100, σ=30) Log:(μ=5,σ=0.75,1.5,2) Gamma:(α=0.5,0.75,2; β=100)	KM followed by bootstrap for 95UCL
2007	Hewett & Ganser	Monte Carlo simulations	subs, ROS, rROS, MLE, rMLE, succop imputation	lognormal, contaminated lognormal	GM=1,GSD=1.2-4	MLE

year	Reference	Methodology	Estimators	Considered distributions	Distributional parameters	Preferred estimators
2008	Antweiler & Taylor	Concentrations of inorganic compounds	subs, MLE, ROS, rROS, KM, subs with Instrument-generated data	lognormal normal	$\frac{\sigma}{\mu} < 5.4$	KM
2008	Jain et al.	Actual data set subject to artificial censoring	Lubin's method; Lynn's method	lognormal	$\frac{\sigma}{\mu} < 1.6$	Lubin's method
2010	European Food Safety Authority	Monte Carlo simulations	subs, ROS,MLE, KM	lognormal, gamma, contaminated lognotmal, lognormal with zero values.	Log: (μ=1, σ=2); Gamma:(α=1.07,β=0.70)	KM for <50% censoring; otherwise, MLE
2014	Huynh et al.	Monte Carlo simulations	β-subs, MLE, KM	lognormal, contaminated lognormal	GM=1,GSD=2,3,4,5	β-subs

EM: Expectation maximization; KM: Kaplan Meier; MLE: Maximum likelihood estimation, rMLE: robust maximum likelihood estimation; PPWM: partial probability weighted moments; ROS: regression on order statistics, rROS: robust regression on order statistics; Subs: substitution GM: Geometric mean; GSD: Geometric standard deviation

Remarks

The literature review reveals that previous studies reached different conclusions about appropriate analysis of left-censored data. We believe that the following shortcomings led to such an inconsistency:

a) Failure to investigate the impact of data skewness: the MLE method wins when data are generated from low to medium skewed distributions as in Shumway et al. (2002) and Hewett & Ganser (2007);
b) Failure to explore the robustness of different methods to departure from an assumed distribution: the general attitude in previous studies has been to generate data from a lognormal distribution and to employ parametric estimators that rely on lognormality without investigating the consequences when real data do not closely adheres to the distributional assumption;
c) Failure to consider and compare the uncertainty intervals provided by different estimators.

2.2 Modeling of concentration data containing left-censored observations

Previous studies investigating the effect of censored data on developing regression models led to the general conclusion that substituting nondetects with a single constant produces biased and misleading estimates (Helsel, 1990; Thompson & Nelson, 2003; Lubin et al., 2004; Helsel, 2005; Eastoe, Halsall, Heffernan & Hung, 2006; Jin, Hein, Deddens & Hines, 2011; Helsel, 2012). For example, Eastoe et al. (2006) showed that substitution of censored concentrations of semi-volatile organic compounds confounded the year-on-year mean trend of pollution, specifically when the censoring percent was higher than 50%.

Appropriate procedures while investigating the relationship between a response variable containing nondetects (i.e., concentration data) and explanatory variable(s) is broadly classified into parametric and non-parametric. With respect to parametric methods, the most common procedure has been Tobit regression (Tobin, 1958), which is based on the

assumption of normal error distribution. However, the performance of Tobit regression is not satisfactory when assumptions related to normality and uniformity of errors are violated (Austin, Escobar & Kopec, 2000), and the censoring percent is large (>30% according to Uh, Hartgers, Yazdanbakhsh & Houwing-Duistermaat, 2008). Another popular parametric method for analyzing censored data is imputing nondetects and combining them with uncensored data before performing the modeling analysis. Simulation studies have shown that the multiple imputation method produces unbiased estimates of regression parameters (Liu, Lu, Kolpin & Meeker, 1997; Lubin et al., 2004; Uh et al., 2008). Despite good properties of the imputation method, Lubin et al. (2004) pointed out that this technique was not necessary when individual values for nondetects were not needed, which in that case they recommended using Tobit regression. Among non-parametric techniques for left-censored data, the literature review points out to Buckley-James regression (Buckley & James, 1979), Schmitt's weighted least square regression (Schmitt, 1985), least absolute deviations regression (Powell, 1984), and Theil-Sen regression (Sen, 1968 and Theil, 1992), among others. The latter is particularly interesting as it accommodates censoring in both response and explanatory variables (i.e., doubly censored data).

The essential assumption of censored regression models is independency between observations, which may not necessarily be true when observations reside in groups. For example, for assessing the exposure of workers to air pollutants at a workplace, Peretz, Goren, Smid & Kromhout (2002) collected repeated measurements of inhalable particulates from a randomly selected number of workers. The measurements obtained from the same worker formed a group and thus were likely to be correlated. Another example of data collection process that induced correlation was reported in Bogner, Gaul, Kolb, Schmiedinger & Huwe (2010): for the purpose of investigating the significant factors affecting water flow process in a forest soil, Bogner et al. (2010) collected soil samples from different areas of the forest situated approximately 50 *m* apart (i.e., plots) and at different depths (i.e., horizons). Obviously, observations collected from the same plot (or horizon) were likely to be more related to each other than to the observations from different plots (or horizons). Applying standard regression models to the above-mentioned examples does not

necessarily give consistent results because the independency assumption is not satisfied. As a matter of fact, biased estimates with erroneously narrow confidence intervals may be obtained (Kreft & De Leeuw, 1998), implying that a regression parameter is significant while actually is not.

A promising approach of fitting regression models to data where some degree of dependency is suspected is mixed effects models. In addition to producing the least biased estimates of regression parameters, mixed effects models enable estimation of within- and between-group variance components, whereas simple regression models only provide a global variance. Some studies in the field of exposure assessment (e.g., European Food Safety Authority, 2010 and Jin et al., 2011) and epidemiology (e.g., Thiebaut & Jacqmin-Gadda, 2004; Twisk & Rijmen, 2009; Vaida & Liu, 2009) employed mixed effects models, while incorporating nondetects in the models. However, to our knowledge, the benefits of these models in the field of environmental engineering and in particular for site characterization studies have not been explored.

Remarks

The unique property of mixed effects models is the inclusion of both fixed and random effects. While fixed effects describe the average relationship between a response and explanatory variables, random effects accounts for inherent heterogeneity in response due to different groups. Literature review revealed that most environmental data analyses considered either (i) censored linear regression models, ignoring the correlation between measurements or (ii) mixed effects models, ignoring below DL observations.

CHAPTER 3

ESTIMATING THE MEAN AND STANDARD DEVIATION OF ENVIRONMENTAL DATA WITH BELOW DETECTION LIMIT OBSERVATIONS: CONSIDERING HIGHLY SKEWED DATA AND MODEL MISSPECIFICATION

Niloofar Shoari[a,b], Jean-Sébastien Dubé[a], Shoja'eddin Chenouri[b]

[a] Department of Construction Engineering, École de technologie supérieure, Université du Québec, 1100, rue Notre-Dame Ouest, Montréal, Québec, Canada, H3C 1K3.
[b] Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1.

This article has been published in *Chemosphere* in July, 2015.

3.1 Abstract

In environmental studies, concentration measurements frequently fall below detection limits of measuring instruments, resulting in left-censored data. Some studies employ parametric methods such as the maximum likelihood estimator (MLE), robust regression on order statistic (rROS), and gamma regression on order statistic (GROS), while others suggest a non-parametric approach, the Kaplan-Meier method (KM). Using examples of real data from a soil characterization study in Montreal, we highlight the need for additional investigations that aim at unifying the existing literature. A number of studies have examined this issue; however, those considering data skewness and model misspecification are rare. These aspects are investigated in this paper through simulations. Among other findings, results show that for low skewed data, the performance of different statistical methods is comparable, regardless of the censoring percentage and sample size. For highly skewed data, the performance of the MLE method under lognormal and Weibull distributions is questionable; particularly, when the sample size is small or censoring percentage is high. In such conditions, MLE under gamma distribution, rROS, GROS, and KM are less sensitive to skewness. Related to model misspecification, MLE based on lognormal and Weibull distributions provides poor estimates when the true distribution of data is misspecified. However, the methods of rROS, GROS, and MLE under gamma distribution are generally

robust to model misspecifications regardless of skewness, sample size, and censoring percentage. Since the characteristics of environmental data (e.g., type of distribution and skewness) are unknown a priori, we suggest using MLE based on gamma distribution, rROS and GROS.

3.2 Introduction

It is often necessary to estimate statistical parameters of contaminant concentration distributions. For example, in contaminated site characterization, this helps us to determine the average level of contamination of a remediation unit or to make statistical inferences to differentiate contaminated soil layers. Complications occur when the contaminant concentrations cannot be quantified because the precision of the laboratory instrument is not sufficient to distinguish the presence of the contaminant from the background noise. As a result, qualitative information is obtained since all we know is that the concentration lies between zero and the detection limit (DL) of measuring instruments (El-Shaarawi & Piegorsch, 2002; Ofungwu, 2014). A measurement that is less than the DL is called a left-censored data point. Furthermore, the concentration data might contain multiple DLs due to the use of different measuring instruments, analytical methods, or combining data sets with different DLs (Jin et al., 2011; He, 2013).

In survival analysis, there are several statistical methods to accommodate right-censored data that can be adapted to address the problem of left-censoring in environmental studies. The most common methods to handle left-censored data include (i) the Maximum Likelihood estimator (MLE), (ii) methods based on Regression on Order Statistics (ROS), and (iii) Kaplan–Meier (KM) procedure. The MLE and ROS-based methods are parametric approaches that assume a predetermined distribution for the data, whereas the KM method is a non-parametric approach and does not require any distributional assumption. The two common versions of ROS are the robust ROS (rROS) and gamma ROS (GROS) methods that rely on lognormal and gamma assumptions, respectively.

Although several studies try to offer guidelines about how to deal with left-censored data through Monte Carlo simulations (Singh et al., 2006; Helsel, 2010b; Helsel, 2012), there has been no general agreement on an appropriate strategy. Literature review reveals that, in addition to sample size (Annan, Liu and Zhang, 2009; Gardner, 2012) and percentage of censoring (Kroll & Stedinger, 1996; Huynh et al., 2014), skewness of the underlying distribution influences the performance of the methods (USEPA, 2006). To our knowledge, only a few studies consider skewness when assessing the performance of the statistical methods in estimating the distributional parameters. For example, USEPA (2006) guidelines state that conclusions derived for low skewed distributions cannot be generalized to moderately and highly skewed ones. We believe that the reason for which the conclusions of previous studies are not in general agreement is the fact that the impact of skewness was overlooked. In fact, the comparative simulations that were based on low to moderately skewed distributions or the simulations in which the results were averaged over a wide range of distributions generally argue in favor of the MLE method under lognormal assumption (Lynn, 2001; Shumway et al., 2002; Hewett & Ganser, 2007; Jain et al., 2008; European Food Safety Authority, 2010). On the other hand, studies that include more skewed distributions report poor performance of MLE under lognormal assumption (Gilliom & Helsel, 1986; Helsel & Cohn, 1988).

In addition to the issue of skewness mentioned earlier, there is an issue regarding the performance of the parametric methods in the case of misspecified distributions. The common practice in environmental literature is to assume that data are lognormally distributed and to use the MLE and rROS methods based on this assumption (El-Shaarawi, 1989; Huybrechts et al., 2002; Baccarelli et al., 2005; Caudill et al., 2007; Leith et al., 2010). It is crucial to know how these methods behave if the underlying parametric model is misspecified. This occurs because

a) There is no evidence that all environmental data are actually lognormal;
b) There is not any straightforward extension of goodness-of-fit tests to establish the true underlying distribution of a given environmental data set due to the presence of left-censored observations.

Unfortunately, comprehensive studies that examine the robustness of the parametric estimators in the case of model misspecification are rather rare. Although the MLE method under lognormal assumption has been widely studied (for example, Gilliom & Helsel, 1986; She, 1997; Shumway et al., 2002; Hewett & Ganser, 2007; among others), only a few environmental studies have attempted to investigate the performance of MLE under Weibull and gamma assumptions (Schmoyeri, Beauchamp, Brandt & Hoffman, 1996; European Food Safety Authority, 2010).

This paper aims at unifying the existing literature on environmental data analysis in the presence of left-censored data by addressing the above mentioned issues. To infer conclusions applicable to more realistic scenarios, we investigate the robustness of the methods under study to variations in data skewness and departures from a distributional assumption. This is key in the analysis of concentration data as neither the underlying distribution nor the skewness is exactly known a priori. We employ an extensive simulation exercise to evaluate the performance of the MLE, rROS, GROS, and KM methods in estimating distributional parameters in simulation scenarios based on different levels of skewness and data generating distributions. The particular objective of this work is to address the issue of the robustness of the parametric methods (i.e., MLE, rROS and GROS). This is achieved by:

a) Investigating the robustness of MLE and rROS based on lognormal assumption when the data are generated from Weibull, gamma, and some mixture distributions;
b) Investigating the robustness of MLE under Weibull assumption when the data are generated from lognormal, gamma, and some mixture distributions;
c) Investigating the robustness of MLE and GROS based on gamma assumption when the data are generated from lognormal, Weibull, and some mixture distributions.

Careful collection and chemical analysis of environmental samples leads to obtaining concentration data sets that are representative of the actual contamination level of the sampling location. However, extracting correct information contained in the data and estimating the contamination level at the scale of a remediation unit or the site is possible using adequate statistical methods. Decisions made upon appropriate statistical methods

protect human health and environment, optimize the allocation of financial resources and save time and effort. The conclusions of this study are applicable to any process that include contaminant quantification such as environmental monitoring and risk assessment.

3.3 Estimation techniques

In this section, we briefly describe the most common statistical methods for analyzing left-censored data. These are maximum likelihood estimation, methods based on regression on order statistics, and Kaplan–Meier methods.

Maximum Likelihood estimation (MLE) utilizes a likelihood function to estimate the distributional parameters. The likelihood function describes the likelihood of observed data, given any member of an assumed parametric family of distributions. In this method, the distributional parameter θ (e.g., the mean and standard deviation) is estimated by maximizing the likelihood function with respect to these parameters. Let $y_1, y_2, \dots, y_n$ be some observations (i.e., contaminant concentrations) and let $\boldsymbol{DL} = (DL_1, \dots, DL_n)$ denote the vector of censoring points (detection limits). The observed concentration data consist of pairs (x_i, δ_i) where $x_i = \max(y_i, DL_i)$ and $\delta_i = I(y_i \geq DL_i)$, meaning that $\delta_i = 1$ if $y_i \geq DL_i$ (in that case $x_i = y_i$) and $\delta_i = 0$ if $y_i < DL_i$ (in that case $x_i = DL_i$) for any $i = 1, \dots, n$. For a random sample of size n, the likelihood contribution from the i^{th} observation is expressed as the probability density function $f(x_i; \theta)$, if the observation is not censored, and as the cumulative density function $F(x_i; \theta)$ if it is left-censored. For a full sample of n observations, the likelihood function is given by

$$LF \propto \prod_{i=1}^{n} f(x_i; \theta)^{\delta_i} F(x_i; \theta)^{1-\delta_i} \tag{3.1}$$

Under mild regularity conditions which are not mentioned here, maximum likelihood estimators are asymptotically normally distributed (Knight, 2000). This means that, for large enough samples, the histograms of MLEs, under repeated sampling, should resemble the

curve of a probability density function of normal distribution. In addition, it is known that MLEs are statistically consistent estimators of the respective population parameters, meaning that as the sample size tends to infinity, the MLE estimates become closer and closer to the true values of the population parameters (Lawless, 2003). In order to use the asymptotic normality of MLEs, a sufficiently large sample size is required. The adequate sample size depends on the underlying assumptions on the population. In addition, the sample size should be larger when data sets consist of left-censored observations. Based on simulation studies, Perez & Lefante (1997) concluded that the larger the variability of data or the percentage of censoring, the larger the sample size required. Further discussions on the use of the maximum likelihood method with censored data can be found in Kuttatharmmakul, Smeyers-Verbeke, Massart, Coomans & Noack (2000) and Lee & Wang (2003).

Robust Regression on Order Statistics (rROS) assumes that data distribution is lognormal. Under this assumption, the scatter plot of the ordered logarithm of the uncensored observations against the quantiles of the normal distribution should show a straight line. The intercept and the slope of the regression line yield an estimate of the mean and standard deviation, respectively. These estimates are employed to predict censored observations (Helsel, 2012). The predicted values are combined with the observed values resulting in a complete data set for which usual methods can be used to estimate statistical parameters. Although the lognormal distribution has been the most commonly used model in environmental studies, Singh, Singh & Iaci (2002) computed upper confidence limits based upon a gamma distribution and concluded that gamma distribution is more appropriate to model uncensored environmental data. Consequently, Singh & Singh (2013) developed the ROS method based on gamma distribution (GROS) and included its implementation in ProUCL (version 5.0.00), statistical Software for analysis of environmental data with left-censored observations.

Gamma Regression on Order Statistics (GROS) fits a regression line to the scatter plot of the ordered uncensored observations against gamma quantiles. Note that one has to estimate the shape and scale parameters of the gamma distribution based on the uncensored

observations in order to compute the respective gamma quantiles. The censored observations are then predicted using the intercept and slope of the regression line and combined with the observed values resulting in a complete data set.

Kaplan–Meier (KM) is a non-parametric technique meaning that it does not rely on a parametric distributional assumption. This method estimates the cumulative distribution function, $F(x) = P(X \leq x)$ non-parametrically. Recall that $F(x)$ is the probability of observing a concentration less than or equal to a certain value x. The resulting estimate of $F(x)$ is a step function, each step corresponding to an uncensored observation.

3.4 Demonstration of the problem

As mentioned earlier, conflicting opinions exist on selecting a suitable technique for handling left-censored data. While some researchers advocated the use of the MLE method, others preferred either the rROS or a non-parametric approach. In this section, we estimate the mean and standard deviation of distribution of real concentration data for four contaminants (pyrene, fluoranthene, acenaphthylene, naphthalene) measured in soil samples obtained from a site characterization study conducted in Montreal, Canada. In addition to the methods of rROS, KM, and MLE under lognormal distribution, which are commonly studied in the environmental literature, we include the MLE method based on Weibull and gamma distributions as well as the ROS method that relies on gamma assumption.

Table 3.1 shows how different methods can provide quite different estimates. Referring to the literature recommendations about dealing with left-censored environmental data (e.g., USEPA, 2006; Helsel, 2012), the MLE method under lognormal assumption should provide good estimates with the sample size and percentage of censoring reported in Table 3.1. However, interestingly, the estimates provided by lognormal MLE are unreasonably larger compared to the estimates provided by other methods (i.e., rROS, GROS, MLE under Weibull and gamma assumptions, and KM). This is particularly clear when estimating the

standard deviation of the soil population under study. We believe the reason of the discrepancy between our example and the literature can be the following.

a) The literature recommendations do not take into account the impact of skewness on the performance of the methods;
b) Assuming the lognormality for environmental data as a default and computing the estimates based on such assumption is incorrect.

The example reported in Table 3.1 clearly demonstrates the necessity for additional simulations to investigate the behavior of different estimators for a wide range of data skewness. Also, we need to explore the robustness of the estimators to distribution misspecification given the fact that the underlying distribution of real data is unknown.

Table 3.1 The estimates of the mean and standard deviation of some concentration data from a characterization study (3 significant digits)

Contaminant	n	% Censoring	MLE (lognormal)	MLE (gamma)	MLE (Weibull)	rROS (lognormal)	GROS (gamma)	KM
Estimation of the mean (mg/kg)								
Pyrene	62	21	66.9	18.4	17.9	18.4	18.3	18.4
Fluoranthene	62	21	89.3	22.5	21.7	22.5	22.5	22.6
Acenaphthylene	62	53	1.51	1.02	0.97	1.02	1.01	1.07
Naphthalene	60	55	5.40	1.81	2.02	1.82	1.8	1.91
Estimation of the standard deviation (mg/kg)								
Pyrene	62	21	$3.63*10^3$	39.3	63.4	65.1	65.1	65.2
Fluoranthene	62	21	$5.80*10^3$	49.1	79.2	85.3	85.3	85.4
Acenaphthylene	62	53	26.7	2.54	3.75	3.32	3.33	3.34
Naphthalene	60	55	$3.60*10^2$	4.97	10.6	5.63	5.64	5.66

3.5 Methodology

Monte Carlo simulations were used to assess the performance of MLE, rROS, GROS and KM methods under a variety of conditions, including sample size, degree of skewness, percentage of censoring, and model misspecification. We simulated data by generating samples from a set of distributions and by allocating a given percentage of data as left-censored. Environmental literature states that concentration data are often right-skewed, therefore, we used lognormal, gamma, Weibull and some mixture distributions that assimilate such data. The mixture distributions are used to investigate the effects of departures from an assumed distribution (i.e., model misspecification) on the estimates. We assumed that the mean of each distribution equals to one ($\mu = 1$) and the standard deviation (σ) takes any of the values 0.5, 1.2, 1.9, 2.6, 3.3, 4. The mixture distributions were considered to have two components: the first component is one of the above mentioned distributions, and the second component belongs to the same distributional family but with $\mu = 3$ and $\sigma = 0.5$. Data sets generated from mixture distributions consisted of a proportion of 0.75 of the first component and 0.25 of the second one (contaminant distribution). For each combination of σ and the type of the distribution, the skewness (γ) was computed, formulas are given in Appendix III. Figure 3.1 shows the shape of distributions for different σ as well as the corresponding amount of skewness. Note that, for a fixed μ, as σ increases, the distributions become more skewed. Based on the shape of distributions (Figure 3.1), we set a subjective criterion that distributions with σ<1 are referred to low skewed distributions, $1 < \sigma < 2$ to moderately skewed distributions, and $\sigma > 2$ to highly skewed distributions. Each generated data set consisted of $n = 60$ observations and contained 30%, 50% and 70% left-censored values. The censoring scheme was defined by computing four censoring points corresponding to four different quantiles of the data generating distribution so that scenarios with the desired percentage of censoring were obtained.

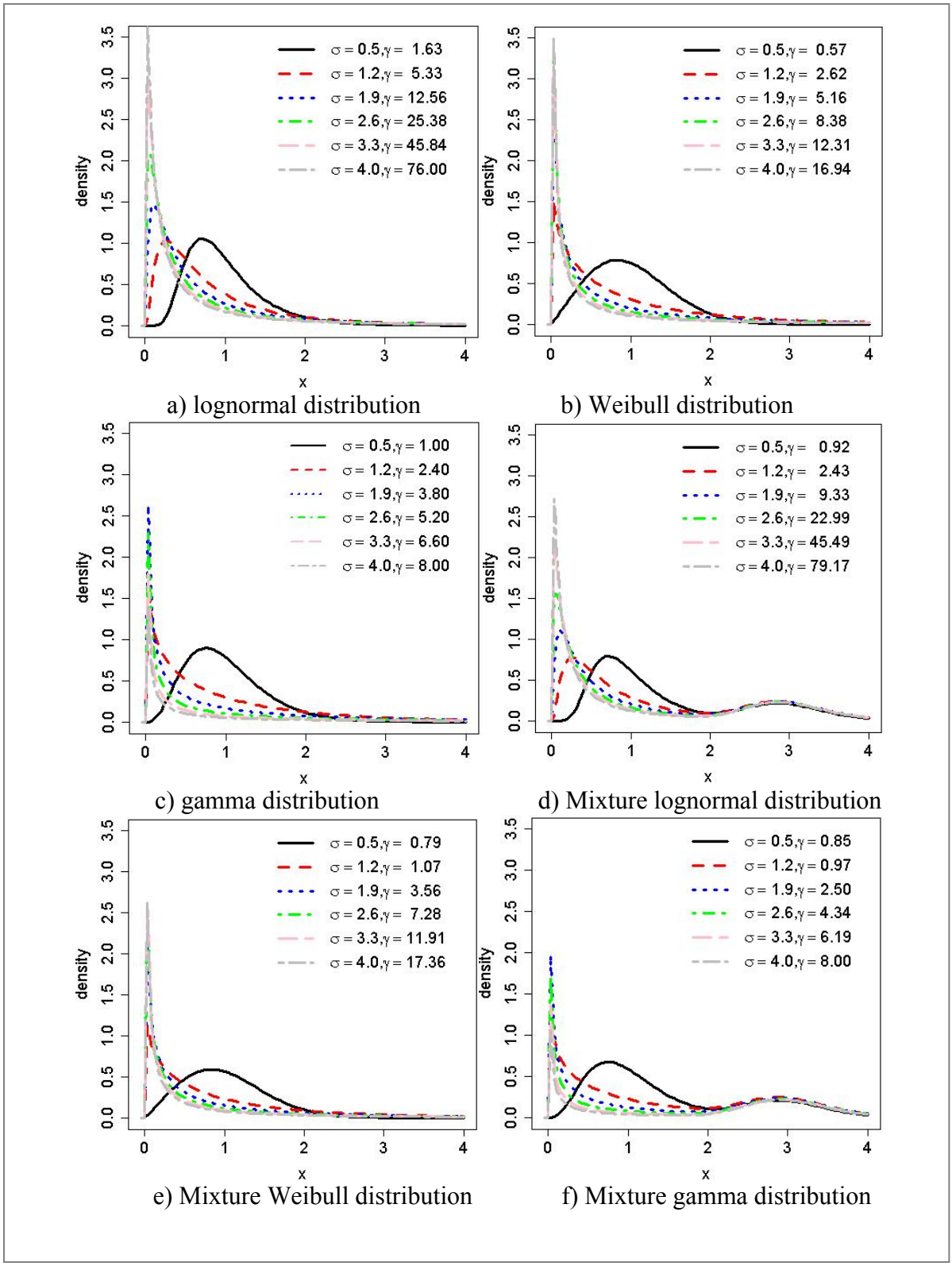


Figure 3.1 Density plots for different distributions with different degrees of skewness reported in Table 3.1

The mean and standard deviation of each distribution were estimated from the simulated data using the parametric MLE (under lognormal, Weibull and gamma distributions), rROS, GROS and non-parametric KM methods.

To compare different estimation methods, we assessed the performance of estimators using the mean square error (MSE). For each simulation scenario, 1000 samples of size $n = 60$ were drawn. Let $\hat{\theta}_i$ be the estimate of θ (either μ or σ of a population) based on the i^{th} sample of size $n, i = 1,2, \ldots 1000$, the Monte Carlo approximation of MSE of $\hat{\theta}$ is given by

$$\widehat{MSE}(\hat{\theta}) = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\theta}_i - \theta)^2 \qquad (3.2)$$

Furthermore, to evaluate the robustness of rROS and lognormal MLE, which are based on lognormal assumption, against model misspecification, analyses were done based on data generated from gamma, Weibull and mixture distributions. In the same way, the robustness of GROS and MLE under Weibull and gamma assumptions were assessed by analyzing data that were not generated from Weibull and gamma distributions, respectively.

3.6 Results

Simulation results show that the skewness, percentage of censoring, and sample size have an impact on the performance of the methods. The impact of these parameters is discussed below. Figure 3.2 and Figure 3.3 show the MSEs of various methods in estimating the mean and standard deviation for different simulation scenarios, respectively. The y-axes are in log-scale, whereas the *x*-axes are in linear scale.

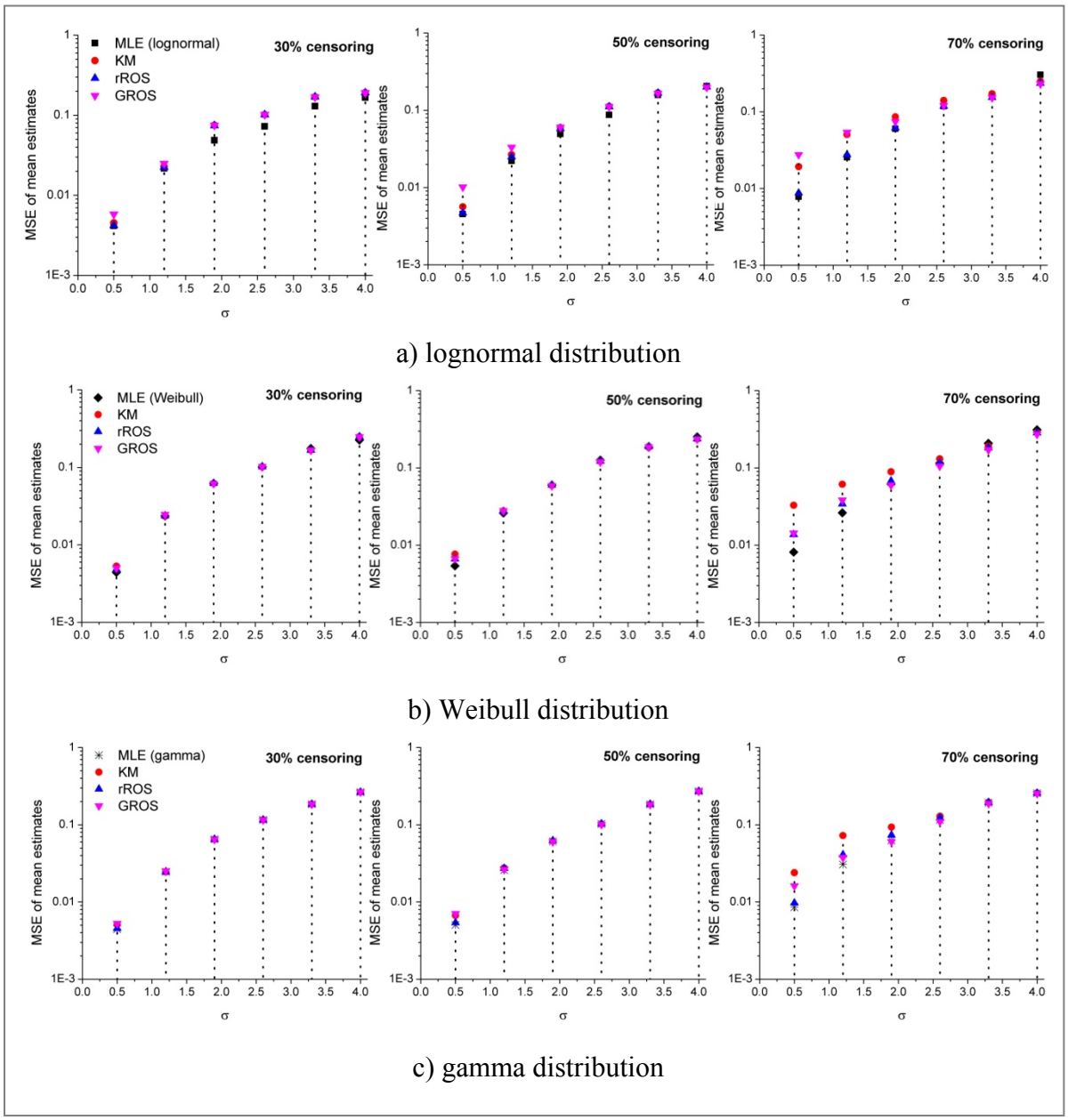


Figure 3.2 The MSE of mean estimates obtained by several methods for $\mu = 1$, σ=0.5,1.2,1.9,2.6,3.3,4

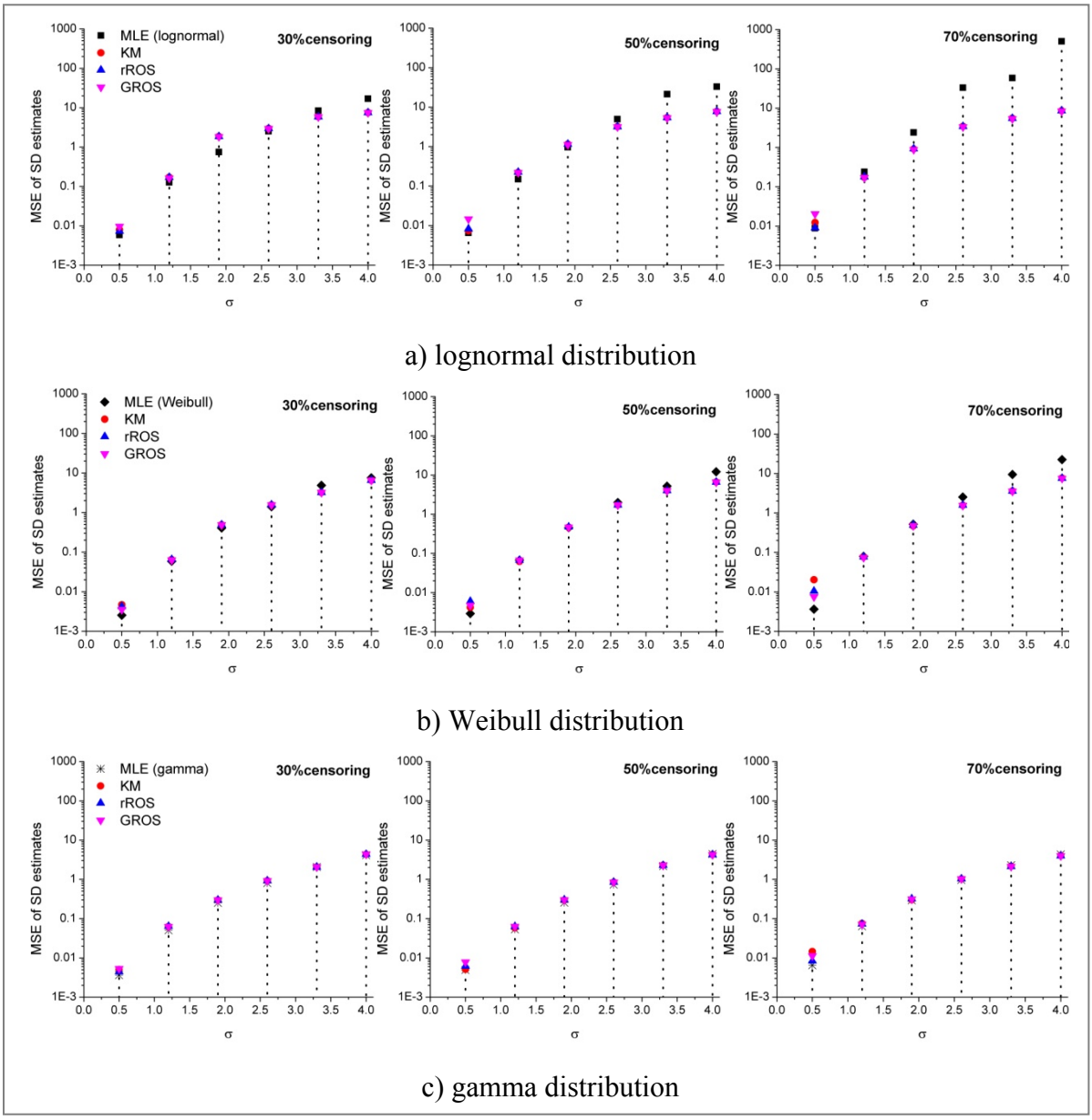


Figure 3.3 The MSE of standard deviation estimates obtained by several methods for $\mu = 1$, σ=0.5,1.2,1.9,2.6,3.3,4

3.6.1 The impact of skewness

Estimation of the mean: Figure 3.2 shows that the MLE, rROS, GROS and KM methods generally produce comparable MSE values in simulation scenarios with 30% and 50% censoring. Note that because of the similarity of the MSE values, these points may seem to overlap. For the mean estimation, the impact of skewness comes into play in cases with 70%

censoring. For example, in scenarios of lognormal and Weibull distributions with 70% censoring, the MLE method tends to produce slightly larger MSEs for highly skewed data (σ =3.3 and 4). In scenarios with <50% censoring, the performance of the KM method is either comparable or slightly worse than the MLE method. However, when the censoring percent increases to 70%, the performance of KM can dramatically deteriorate. As the skewness increases, the performance of the MLE, rROS and GROS methods starts to deteriorate and in some cases can be even worse than the KM method. For example, in the case of lognormal distribution with 70% censoring, the performance of MLE is even worse than KM.

Estimation of the standard deviation: For estimating the standard deviation, the impact of skewness is more pronounced. Figure 3.3a and b depict the MSE values in log-scale produced by different estimators based on data sets generated from lognormal and Weibull distributions. In both cases, the MLE method performs evidently better for cases characterized by low skewed distributions. See the superiority of the MLE for all scenarios with σ=0.5. On the other hand, for moderately to highly skewed distributions, the MLE method performs poorly compared to the rROS, GROS and KM methods. An important observation inferred from Figure 3.3c is that, contrary to the cases of lognormal and Weibull distributions, MLEs under gamma distribution are robust to variations in skewness of distributions. Simulation results also show that the rROS and GROS methods are robust to skewness regardless of the percentage of censoring and type of the underlying data generating distribution. Note that these two methods generally provide similar MSEs across all simulation scenarios.

3.6.2 The impact of the percentage of censoring and sample size

It is important to note that the effect of skewness on the performance of the methods described in this paper should be studied along with other factors such as the percentage of censoring and sample size. For example, in the case of estimating the standard deviation based on data generated from lognormal and Weibull distributions with 30%, 50% and 70%

censoring, the MLE method produces the lowest MSEs as long as σ is equal or below 2.6, 1.9, and 1.2, respectively (Figure 3.3a and b). This implies that the smaller the percentage of censoring, the less sensitive the MLE to skewness.

In addition to the skewness of distributions and percentage of censoring, the simulation results highlight the importance of the sample size on the performance of the methods used in this study. Environmental literature often considers MLE as a reliable estimator when the sample size contains at least 50 observations (Helsel, 2006; Helsel, 2012). However, it has come to our attention that this statement is only valid when the underlying distribution has low skewness. Even if the rule of thumb of having data sets with at least 50 observations is respected in our simulation study, we observe that the MLE method produces poor estimates of the standard deviation in scenarios based on highly skewed distributions. For example, see the MSE values obtained by MLE in scenarios of lognormal distribution, 30% censoring, and $\sigma \geq 3.3$ in Figure 3.3a. It is worth mentioning that, under some fairly general conditions, MLE is asymptotically consistent, and normally distributed with the rate of convergence that depends on the shape of the distribution. This is clearly evident in our simulation study showing that the required sample size to use the asymptotic properties of MLE strongly depends on the type of the data generating distributions and their skewness. To illustrate this, we focus on three simulation scenarios in which MLE performs poorly (discussed previously in Figure 3.3a and b):

a) lognormal distribution with 60 observations, σ=3.3 and 50% censoring;
b) lognormal distribution with 60 observations, σ=4 and 50% censoring;
c) Weibull distribution with 60 observations, σ=4 and 50% censoring.

Simulations were repeated for these critical scenarios except that the sample size gradually increased up to 360 while other simulation parameters were maintained as before. Table 3.2 shows the impact of skewness on the consistency of the MLE method. The MSE values in bold represent the smallest MSE and thus the preferred method. In the case of the lognormal population with $\sigma = 3.3$, MLE under lognormality produces the worst estimate of the standard deviation (with MSE=21.279) compared to the rROS, GROS, and KM methods for

sample size as large as 60 observations, and is the best estimator (with MSE=4.257) when the sample size increases to 120. For the same distribution, when σ increases to 4 the minimum sample size for which the MLE under lognormality outperforms rROS, GROS and KM is 180. Based on simulations for the critical scenario of Weibull distribution, the sample size of $n = 180$ is sufficiently large to obtain reliable MLEs. Figure 3.4 illustrates the asymptotic normality property of MLEs. We observe that in data sets with 60 observations generated from lognormal and Weibull distributions, the histograms of the MLEs are skewed however they approach normal distribution as the sample size increases to 360. On the other hand, when data come from gamma distributions, 60 observations are sufficient for approximate normality to hold.

Table 3.2 Mean square error (MSE) in estimating the standard deviation for lognormal data with 50% censoring

Sample size	MLE	KM	rROS	GROS
lognormal ($\sigma = 3.3$)				
60	21.279	5.427	5.398	**5.339**
120	**4.257**	5.12	5.103	5.06
180	**1.943**	3.221	3.223	3.181
240	**1.536**	2.576	2.581	2.543
300	**1.232**	3.072	3.072	3.042
360	**0.808**	2.98	2.981	2.954
lognormal ($\sigma = 4$)				
60	32.875	7.761	7.737	**7.667**
120	10.664	9.602	9.563	**9.51**
180	**4.389**	8.123	8.105	8.06
240	**2.709**	27.57	27.482	27.439
300	**2.369**	5.882	5.879	5.841
360	**1.715**	5.076	5.077	5.037
Weibull ($\sigma = 4$)				
60	12.154	6.614	6.583	**6.574**
120	4.938	4.87	4.854	**4.848**
180	**2.957**	3.437	3.433	3.428
240	**1.975**	3.391	3.387	3.383
300	**1.514**	2.906	2.903	2.899
360	**1.212**	2.856	2.853	2.85

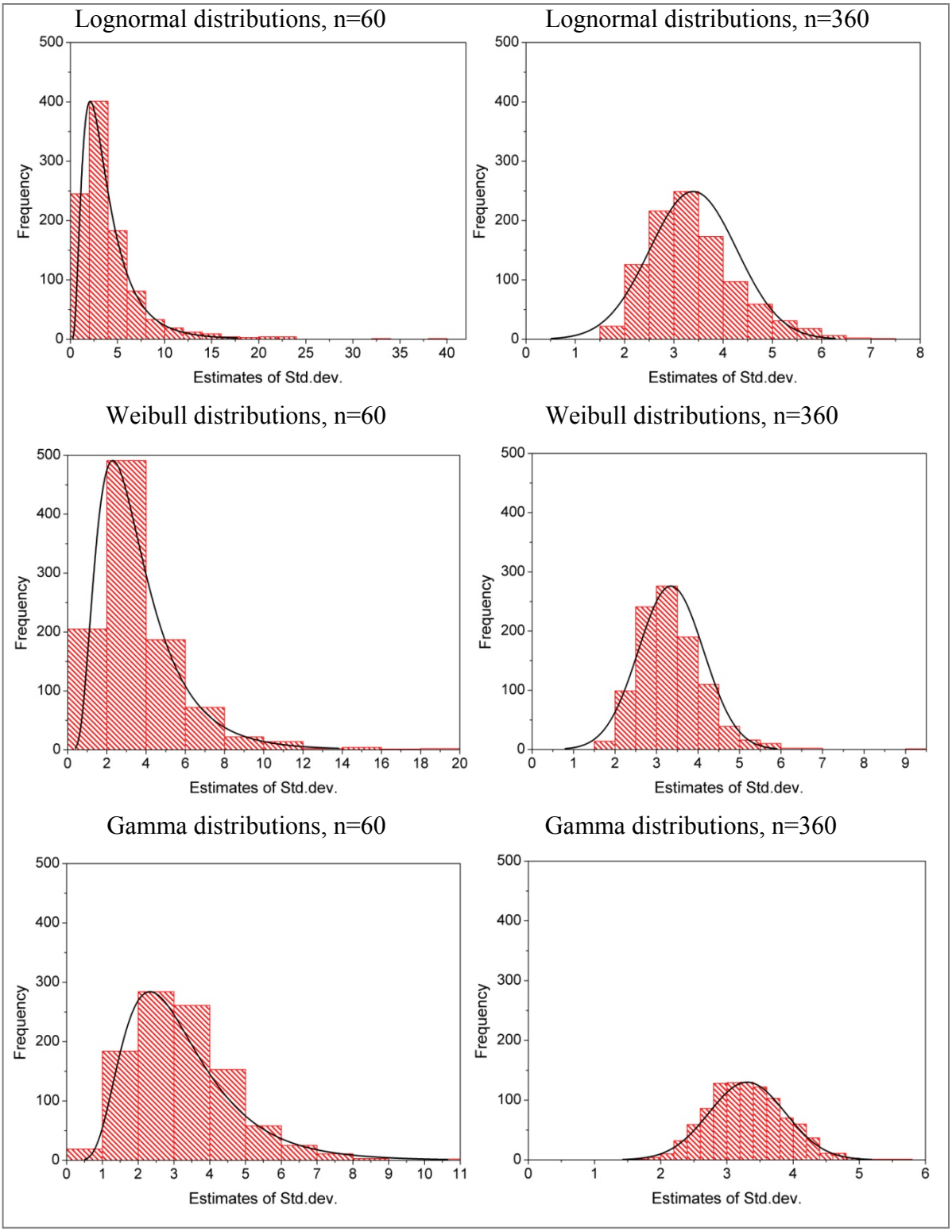


Figure 3.4 Histogram of the standard deviation of lognormal, Weibull, and gamma distributions with σ =3.3

3.6.3 The impact of distributional misspecification

Identifying the underlying distribution of environmental data is not straightforward due to the presence of left-censored observations. This limitation can lead researchers to incorrectly assume lognormality for the majority of environmental data and to use parametric methods based on this assumption. Use of the popular parametric methods for analyzing left-censored data under lognormality (i.e., MLE, rROS) would lead to biased estimates and potentially misleading inferences. Besides, Singh et al. (2002) suggest assuming a gamma distribution for uncensored environmental data and employ the ROS method based on such assumption (i.e., the GROS method) (Singh & Singh, 2013). In this section, we explore the robustness of MLE, rROS and GROS to distributional misspecification and show that MLE based on gamma assumption is generally a robust estimator. Table 3.3 shows the MSEs obtained by the MLE, rROS and GROS estimators in the misspecified settings with 50% censoring; the MSE values obtained with 30% and 70% censoring are available in Appendix III. In Table 3.3, the extremely large MSEs (values larger than 1000) are replaced by a star. For each simulation scenario, one can get the relative error of each parametric method in case of misspecified distribution by

$$\text{Percentage of error for method } i = \frac{MSE_i - MSE^*}{MSE^*} \tag{3.3}$$

where MSE_i is the MSE of method i and MSE^* is the smallest MSE obtained from MLE, rROS, GROS and KM. Note that method i refers to any parametric method used in this paper such as rROS, GROS, and MLE under lognormal, gamma, Weibull distributions. As can be seen in Table 3.3, the MLE method under lognormal assumption leads to very large MSEs when the simulated data are generated from moderately to highly skewed Weibull, gamma and mixture distributions. The poor performance of lognormal MLE is clearly evident when comparing the average values of the percentage of error for each method. The MLE under Weibull assumption performs well in a few scenarios where the underlying distribution is a lognormal or a mixture of lognormal or Weibull distributions. Otherwise, using the MLE under Weibull assumption leads to large MSEs especially when the underlying distributions

are highly skewed. On the contrary to lognormal and Weibull assumptions for the MLE method, the MLEs obtained under gamma distribution seem to be less affected by model misspecification. In fact, the MLE under the gamma distribution generally provides reasonably small MSEs and percentages of errors, regardless of the type of underlying distribution and its skewness. Moreover, our simulation results show that the performance of rROS and GROS are comparable since similar MSEs are obtained by these methods. Although the methods of rROS and GROS generally perform well, the average values of the percentage of error reported in Table 3.3 suggest that the performance of these methods deteriorate when data come from a lognormal or a mixture of lognormal distributions. Based on the discussions above, we conclude the following:

- The MLE method under lognormal and Weibull assumptions provide good estimates only in a few simulation scenarios and thus are not robust estimators;
- The MLE method under gamma assumption, followed by the rROS and GROS methods, are fairly robust estimators, regardless of the percentage of censoring, underlying distribution of data and skewness.

Table 3.3 The MSE of the mean and standard deviation produced by rROS, GROS and MLE under different distributional assumptions and model misspecification in scenarios with 50% censoring

		MSE of the mean estimates					MSE of the standard deviation estimates				
True dist.	**Parameters ($\mu = 1$)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**
Weibull	$\sigma = 0.5$	0.006	NA	0.005	0.007	0.007	0.005	NA	0.004	0.006	0.005
	$\sigma = 1.2$	0.051	NA	0.026	0.027	0.028	1.166	NA	0.054	0.067	0.067
	$\sigma = 1.9$	0.340	NA	0.059	0.060	0.059	89.998	NA	0.312	0.474	0.466
	$\sigma = 2.6$	2.147	NA	0.120	0.122	0.119	*	NA	1.084	1.721	1.711
	$\sigma = 3.3$	8.611	NA	0.187	0.189	0.185	*	NA	2.432	3.988	3.979
	$\sigma = 4$	42.261	NA	0.236	0.238	0.235	*	NA	4.750	6.583	6.574
Percentage of error		*	-	1%	6%	5%	*	-	4%	57%	49%
gamma	$\sigma = 0.5$	0.005	0.005	NA	0.005	NA	0.006	0.005	NA	0.006	NA
	$\sigma = 1.2$	0.060	0.026	NA	0.027	NA	1.592	0.069	NA	0.063	NA
	$\sigma = 1.9$	5.785	0.096	NA	0.062	NA	*	2.059	NA	0.301	NA
	$\sigma = 2.6$	*	1.655	NA	0.103	NA	*	303.905	NA	0.846	NA
	$\sigma = 3.3$	*	612.686	NA	0.184	NA	*	*	NA	2.264	NA
	$\sigma = 4$	*	*	NA	0.271	NA	*	*	NA	4.246	NA
Percentage of error		*	*	-	3%	-	*	*	-	15%	-
lognormal	$\sigma = 0.5$	NA	0.007	0.005	NA	0.01	NA	0.007	0.006	NA	0.015
	$\sigma = 1.2$	NA	0.024	0.025	NA	0.033	NA	0.111	0.111	NA	0.220
	$\sigma = 1.9$	NA	0.044	0.056	NA	0.060	NA	0.534	0.593	NA	1.119
	$\sigma = 2.6$	NA	0.066	0.108	NA	0.111	NA	1.507	1.777	NA	3.151
	$\sigma = 3.3$	NA	0.094	0.164	NA	0.165	NA	3.063	3.607	NA	5.339
	$\sigma = 4$	NA	0.115	0.197	NA	0.198	NA	5.394	6.176	NA	7.667
Percentage of error		-	10%	43%	-	71%	-	3%	10%	-	96%

(Continued)

		MSE of the mean estimates					MSE of the standard deviation estimates				
True dist.	**Parameters ($\mu = 1$)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**
Mixture Weibull	$\sigma = 1.00$	0.216	0.239	0.224	0.212	0.229	0.344	0.322	0.349	0.39	
	$\sigma = 1.38$	0.262	0.301	0.298	0.287	0.356	0.158	0.223	0.232	0.224	
	$\sigma = 1.88$	0.319	0.42	0.404	0.415	0.448	4.084	0.522	0.567	0.576	
	$\sigma = 2.43$	0.396	0.542	0.500	0.529	0.524	59.858	1.240	1.389	1.524	
	$\sigma = 3.00$	0.567	0.645	0.577	0.615	0.59	*	2.496	2.747	2.881	
	$\sigma = 3.58$	1.195	0.72	0.646	0.689	0.656	*	4.278	4.694	5.422	
Percentage of error		16%	22%	14%	17%	22%	*	7%	16%	23%	15%
Mixture gamma	$\sigma = 1.00$	0.026	0.023	0.023	0.025	0.034	0.083	0.010	0.019	0.008	0.008
	$\sigma = 1.38$	0.114	0.040	0.04	0.050	0.056	2.407	0.046	0.063	0.044	0.049
	$\sigma = 1.88$	4.067	0.077	0.064	0.069	0.063	*	1.009	0.263	0.191	0.221
	$\sigma = 2.43$	*	0.668	0.109	0.117	0.092	*	40.863	0.667	0.644	0.720
	$\sigma = 3.00$	*	13.783	0.155	0.162	0.129	*	*	1.385	1.529	1.626
	$\sigma = 3.58$	*	860.356	0.240	0.246	0.202	*	*	2.595	3.602	3.765
Percentage of error		*	*	10%	20%	15%	*	*	63%	25%	36%
Mixture lognormal	$\sigma = 1.00$	0.026	0.024	0.023	0.024	0.034	0.091	0.012	0.021	0.007	0.009
	$\sigma = 1.38$	0.060	0.036	0.037	0.041	0.044	0.614	0.050	0.045	0.107	0.115
	$\sigma = 1.88$	0.13	0.055	0.059	0.064	0.055	2.880	0.210	0.168	0.589	0.597
	$\sigma = 2.43$	0.225	0.082	0.137	0.145	0.129	8.388	0.649	0.960	4.553	4.557
	$\sigma = 3.00$	0.383	0.093	0.132	0.140	0.114	26.14	1.146	1.339	4.030	4.026
	$\sigma = 3.58$	0.565	0.120	0.184	0.193	0.162	46.173	2.201	2.73	6.686	6.678
Percentage of error		180%	2%	30%	39%	33%	*	18%	49%	242%	250%

NA: Not Applicable

3.7 Summary and conclusions

This paper evaluates the performance of the most common statistical methods for handling left-censored data. The methods under study are MLE, rROS, GROS and KM. Our simulation study emphasizes the importance of including skewness, percentage of censoring and sample size when evaluating the performance the aforementioned statistical methods. Some of the highlights are as follows:

- Impact of skewness: we observe that in the case of low skewed data, the performance of the MLE, rROS, GROS and KM methods are comparable although the MLE method provides slightly better estimates. When dealing with highly skewed data, the performance of the MLE method drastically deteriorates. For example, the simulations show that the MLE method under lognormal and Weibull assumption provides poor estimates in simulation scenarios with moderately to highly skewed distributions even though the simulated data were generated from lognormal and Weibull distributions;
- Impact of percentage of censoring: as the percentage of censoring decreases, the MLE method becomes more robust to variation to skewness;
- Impact of sample size: there is no magical sample size that guarantees the superiority of the MLE method over the others. In fact, our simulations show that the appropriate sample size strongly depends on the type of the distribution and its skewness.

Recall that the true distribution of left-censored environmental data and their characteristics (such as skewness) are unknown. Therefore, it is crucial to have an estimator that is robust to variations in skewness as well as model misspecification. This paper investigated the impact of model misspecification on the performance of rROS, GROS, and MLE covering a wide range of data skewness. Simulation results in this study show that the MLE under gamma assumption, rROS, and GROS are viable alternatives to accommodate a variety of right-skewed distributions, regardless of the percentage of censoring.

CHAPTER 4

ON THE USE OF THE SUBSTITUTION METHOD IN LEFT-CENSORED ENVIRONMENTAL DATA

Niloofar Shoari[a,b], Jean-Sébastien Dubé[a], Shoja'eddin Chenouri[b]

[a] Department of Construction Engineering, École de technologie supérieure, Université du Québec, 1100, rue Notre-Dame Ouest, Montréal, Québec, Canada, H3C 1K3.
[b] Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1.

This article has been published in the journal of
Human and Ecological Risk Assessment in September, 2015.

4.1 Abstract

In risk assessment and environmental monitoring studies, concentration measurements frequently fall below detection limits (DL) of measuring instruments, resulting in left-censored data. The principal approaches for handling censored data include the substitution-based method, maximum likelihood estimation, robust regression on order statistics, and Kaplan-Meier. In practice, censored data are substituted with an arbitrary value prior to use of traditional statistical methods. Although some studies have evaluated the substitution performance in estimating population characteristics, they have focused mainly on normally and lognormally distributed data that contain a single DL. We employ Monte Carlo simulations to assess the impact of substitution when estimating population parameters based on censored data containing multiple DLs. We also consider different distributional assumptions including lognormal, Weibull, and gamma. We show that the reliability of the estimates after substitution is highly sensitive to distributional characteristics such as mean, standard deviation, skewness, and also data characteristics such as censoring percentage. The results highlight that although the performance of the substitution-based method improves as the censoring percentage decreases, its performance still depends on the population's distributional characteristics. Practical implications that follow from our findings indicate

that caution must be taken in using the substitution method when analyzing censored environmental data.

4.2 Introduction

Analytical results of environmental samples often contain non-quantitative concentration measurements that are below the detection limits (DL) of measuring instruments. Left-censored concentrations, reported as less than DL, complicate any traditional statistical analysis. The common practice to circumvent the issues due to left-censoring is to substitute censored values with an arbitrary number (e.g., 0, DL/2, $DL/\sqrt{2}$, or DL) and to analyze data with traditional methods such that the substituted values are assumed to be actual observed data (e.g., Zhao & Frey, 2003; McCarthy, O'Brien, Charrier & Hafner, 2009; Wu et al., 2011; Struciński et al., 2015). In this regard, some studies recommend the use of the maximum likelihood estimation method (MLE) or robust regression on order statistics (rROS), which are based on a distributional assumption, or the non-parametric Kaplan-Meier (KM) approach. For example, Hewett & Ganser (2007) concluded that MLE is an appropriate method to estimate the mean and 95th percentile of left-censored occupational health data. In contrast, Antweiler & Taylor (2008) used KM to estimate the mean, standard deviation, and different quantiles of left-censored data with <70% censoring.

Helsel & Cohn (1988) used artificial censored data sets to assess the performance of several estimators and concluded that substituting left-censored values with an arbitrary constant produces estimates with large bias and mean square error (MSE). El-Shaarawi & Esterby (1992) provided a tool to quantify the bias due to substitution when the mean and standard deviation of data as well as the proportion of censored values are available. However, they acknowledged that, in practice, the magnitude and direction of the bias are not quantifiable because distributional parameters of data (i.e., mean and standard deviation) are unknown. In other studies, including Hornung & Reed (1990), Farnham et al. (2002), Hewett & Ganser (2007), Antweiler & Taylor (2008), European Food Safety Authority (2010), and Leith et al. (2010), the substitution-based method performed reasonably well under certain simulation

circumstances although its use was avoided due to lack of a theoretical basis. When the percentage of censoring is small (e.g. <15%), USEPA (2006) and Eastoe et al. (2006) agreed that substitution gives results comparable to both parametric and non-parametric methods. However, Helsel (2006) questioned whether there was a censoring percentage below which reliable results could be obtained.

Previous studies have focused mainly on evaluating the performance of the substitution-based method for normally and lognormally distributed data sets that contain a single DL. However, it often occurs, especially in contaminated site assessment for instance, that environmental samples are analyzed by different laboratories or in different time periods and, therefore, the resulting data sets often contain multiple DLs. In addition, no study has established that all environmental data follow a specific distribution. Thus, this study considers censored data sets characterized by multiple DLs and a variety of right-skewed distributions (i.e., lognormal, Weibull and gamma distributions with different levels of skewness). In this paper, inherent problems associated with arbitrary substitution are explored and compared to parametric and non-parametric methods. Among other investigations, this study also determines the efficacy of the substitution-based method versus the MLE, rROS, and KM methods for estimating the mean and standard deviation of distributions based on data with a small percentage of censoring.

4.3 Alternative methods for handling left-censored data

The most commonly used methods to handle left-censored data are MLE, rROS, and KM. The MLE method utilizes a likelihood function to estimate the distributional characteristics or attributes. The likelihood function provides the likelihood of observed data, under any given member of an assumed family of distributions. In this method, a distributional parameter θ (e.g., the mean or the standard deviation) is estimated by maximizing the likelihood function with respect to this parameter. Let $y_1, y_2, \dots, y_n$ be n independent and identically distributed observations from a population with the probability density function $f(x|\theta)$ and cumulative distribution function $F(x|\theta)$. Also, let $DL_1, DL_2, \dots, DL_n$ denote

detection limits or censoring points. The observed concentration data set consists of pairs (x_i, δ_i) where $x_i = \max(y_i, DL_i)$ and $\delta_i = 1$ if $y_i \geq DL_i$ and $\delta_i = 0$ otherwise. The likelihood function based on the observed data is given by

$$\prod_{i=1}^{n} f(x_i; \theta)^{\delta_i} F(x_i; \theta)^{1-\delta_i} \tag{4.1}$$

The method of robust ROS (rROS) is based on the assumption that the data generating distribution is either normal or lognormal. As discussed in Helsel (2012), the ROS method considers the scatter plot of the ordered uncensored data (in the case of normal distribution) or the ordered logarithm of uncensored data (in the case of lognormal distribution) against the quantiles of the standard normal distribution. If the distributional assumption (either normal or lognormal) is correct, the regression line fitted to this scatter plot is approximately linear. The intercept and the slope of the regression line are estimates of the mean and standard deviation of the underlying distribution, respectively. In a robust version of ROS, these estimates are then employed to predict censored observations. These predicted values are combined with the observed values resulting in a complete data set for which traditional estimation methods (e.g., the simple average of the observations and their standard deviation) can be used.

KM is a non-parametric method that does not rely on a distributional assumption to estimate the population characteristics. The KM method estimates the cumulative distribution function of contaminant concentration, $F(x) = P(X \leq x)$, without assuming any specific form for *F*. Further discussions on the use of the KM method with left-censored data can be found in Gillespie et al. (2010) and Helsel (2012).

4.4 Methodology

An extensive Monte Carlo simulation study was conducted to compare the performance of the MLE, rROS, and KM methods versus substituting the censored values with a constant prior to use of familiar methods for estimating the mean and standard deviation of contaminants distribution. Simulated data were generated under different right-skewed distributions for given μ and σ. To have censored data sets with multiple DLs, we imposed 10% and 50% censoring to the generated data. In total, for each censoring percent 180 simulation scenarios were studied in this paper. The estimated mean and standard deviation of the corresponding data generating distributions were compared for different methods of estimation and also to the true values. The simulation procedure is as follows.

Step 1: Generate data from lognormal, Weibull, and gamma distributions with sample size n=60 observations for all combinations of μ = 1, 2,…,10, and σ =0.5, 1.2, 1.9, 2.6, 3.3, 4. The wide range of μ and σ results in data sets with a coefficient of variation (CV=σ/μ) ranging between 0.05 and 4. As an example, Figure 4.1 shows different shapes of lognormal distribution for $\mu = 1$ and σ=0.5, 1.2, 1.9, 2.6, 3.3, 4. This Figure clearly illustrates that, for a given μ, as σ increases, the value of the CV increases and consequently, the distribution becomes more skewed. In this study, skewness was defined in terms of CV such that CV<1 refers to mildly skewed data, 1≤CV<2 to moderately skewed data, and CV≥2 to highly skewed data. A Similar convention was used in Singh et al. (2006) to represent the skewness.
Step 2: Use fictional DLs to accommodate left-censoring as follows. To obtain data sets with 50% censoring and multiple DLs, the 0.2, 0.4, 0.6, and 0.8 quantiles of the underlying data generating distribution were computed. 25% of the simulated data (from step 1) were censored at 0.2 quantile, 25% at 0.4 quantile, 25% at 0.6 quantile, and 25% at 0.8 quantile. Similarly, to obtain data sets with 10% censoring and multiple DLs, we computed 0.05, 0.10, and 0.15 quantiles of the distributions; 33% of the simulated data were censored at 0.05 quantile, 33% at 0.10 quantile ,and 33% at 0.15 quantile.

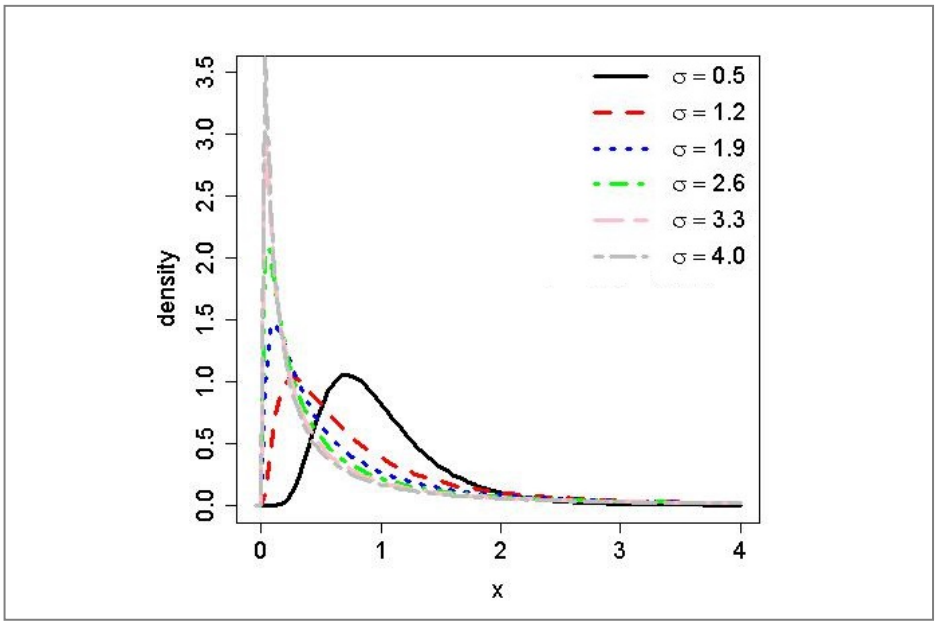


Figure 4.1 The shape of lognormal distribution for μ=1 and different σ values corresponding to CV=0.5, 1.2, 1.9, 2.6, 3.3, 4

Step 3: The mean and standard deviation of each population were estimated by

- Substituting censored values with DL/2 and using the traditional estimation methods for complete data sets;
- MLE under lognormal, gamma, and Weibull assumptions;
- rROS (under lognormal assumption);
- KM method.

Step 4: For each data generating distribution, repeat steps 1 to 3 N=1000 times and compute the Monte Carlo approximations of the MSEs of the estimators of the parameter θ (either μ or σ). To be precise, suppose $\hat{\theta}_i$ is an estimate of θ based on i^{th} simulated data set with certain combination of μ and σ. The approximated MSE of $\hat{\theta}$ is given by

$$\widehat{MSE}(\hat{\theta}) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_i - \theta)^2 \qquad (4.2)$$

where θ is the true value depending on the underlying combination of μ and σ. Note that similar to the MSE, we have the following Monte Carlo bias-variance decomposition.

$$\widehat{\mathrm{MSE}}(\hat{\theta}) = \widehat{\mathrm{var}}(\hat{\theta}) + \widehat{\mathrm{bias}}(\hat{\theta}) \tag{4.3}$$

where

$$\widehat{\mathrm{var}}(\hat{\theta}) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_i - \bar{\hat{\theta}})^2 \tag{4.4}$$

$$\widehat{bias}(\hat{\theta}) = \left(\bar{\hat{\theta}} - \theta\right)^2 \tag{4.5}$$

$$\bar{\hat{\theta}} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\theta}_i \tag{4.6}$$

All simulations were implemented in the statistical software R and the code for simulation scenarios based on lognormal distribution is available in Appendix IV (Algorithm-A IV-1).

4.5 Results and discussions

4.5.1 Data from lognormal distribution

Figure 4.2 and Figure 4.3 illustrate the MSE values provided by the traditional statistical methods after substitution of the censored values together with those provided by the MLE, rROS, and KM methods. These figures represent simulation scenarios based on the data sets generated from lognormal distributions with μ = 1, 2,…,10 and σ=0.5, 1.9, 3.3, with 50% censoring. The plots for all other scenarios of σ are available in Figure-A IV-1 and Figure-A-IV-2 of Appendix IV. Note that, in Figure 4.2 and Figure 4.3, the y-axis is in log-scale, whereas the x-axis is in linear scale. In general, the substitution-based method does not consistently perform better or worse than other methods (i.e., MLE, rROS, and KM) across all simulation scenarios. For example, for the mean estimation, Figure 4.2b shows that the substitution-based method has comparable or smaller MSEs compared to other estimators as

long as the simulated data are generated from μ equal to 1 and 2 and σ=1.9. However, the substitution-based method provides larger MSEs for any combination of $\mu>2$ and σ=1.9. For the standard deviation estimation (Figure 4.3b), the performance of the substitution-based method is similar or better than other methods in scenarios where the simulated data are generated from μ=1,2,3,4 and σ=1.9. When the μ of the data generating distribution exceeds 5, the performance of the substitution-based method starts to deteriorate. The same observations can be made for any given σ in this study (see the plots in Appendix IV).

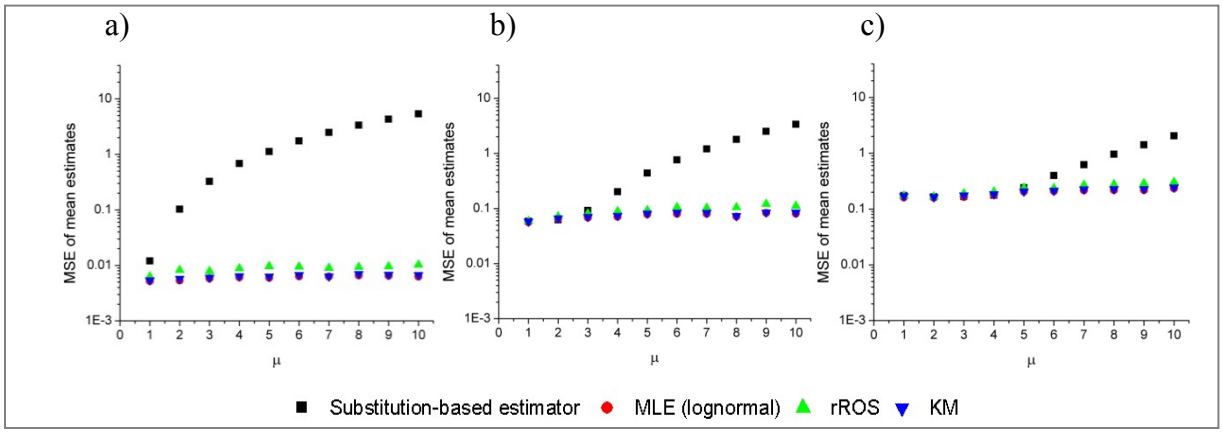


Figure 4.2 The MSEs of different methods in estimating the mean of lognormal distribution with μ=1,2,…,10 and a) σ=0.5, b) σ =1.9, c) σ=3.3

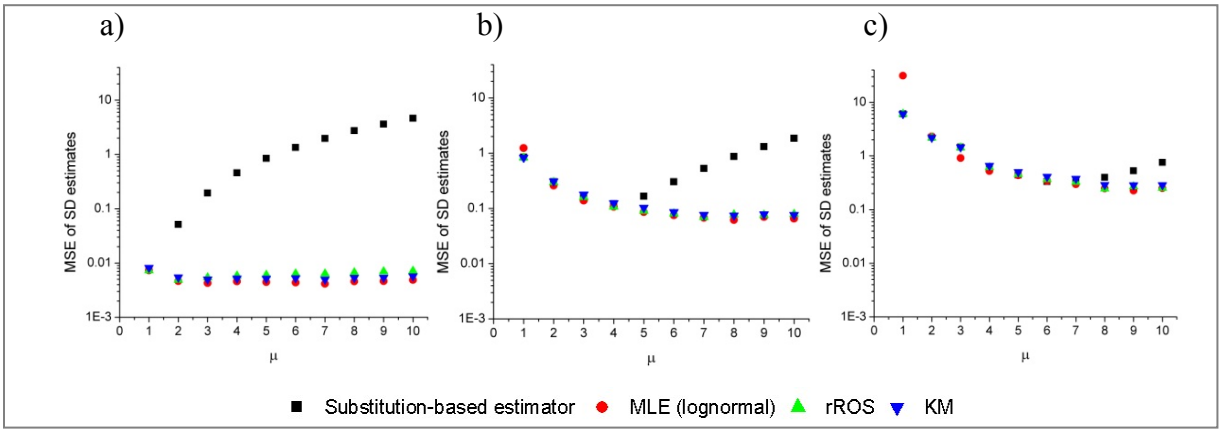


Figure 4.3 The MSEs of different methods in estimating the standard deviation of lognormal distribution with μ=1,2,…,10 and a) σ=0.5, b) σ =1.9, c) σ=3.3

As shown in Figure 4.2 and Figure 4.3, depending on the characteristics of the simulated data (i.e., mean and standard deviation of the data generating distributions), substituting the censored observations may or may not lead to good estimates. This result is probably due to misspecification of the shape of the original distribution of data after substituting the censored observations with a constant value. To further investigate the reason for this behavior of the substitution-based method, let us focus on the simulation scenarios of σ=1.9 with 50% censoring. Figure 4.4 shows the distribution of data after substitution with DL/2 superimposed on the distribution of uncensored data in the following three situations:

1) Substitution of the censored values results in estimates that are equivalent to those provided by MLE, rROS, and KM (μ=2, σ=1.9);

2) Substitution of the censored values leads to slight over/under estimation (μ=5, σ=1.9);

3) Substitution of the censored values clearly results in poor estimates (μ=10, σ=1.9).

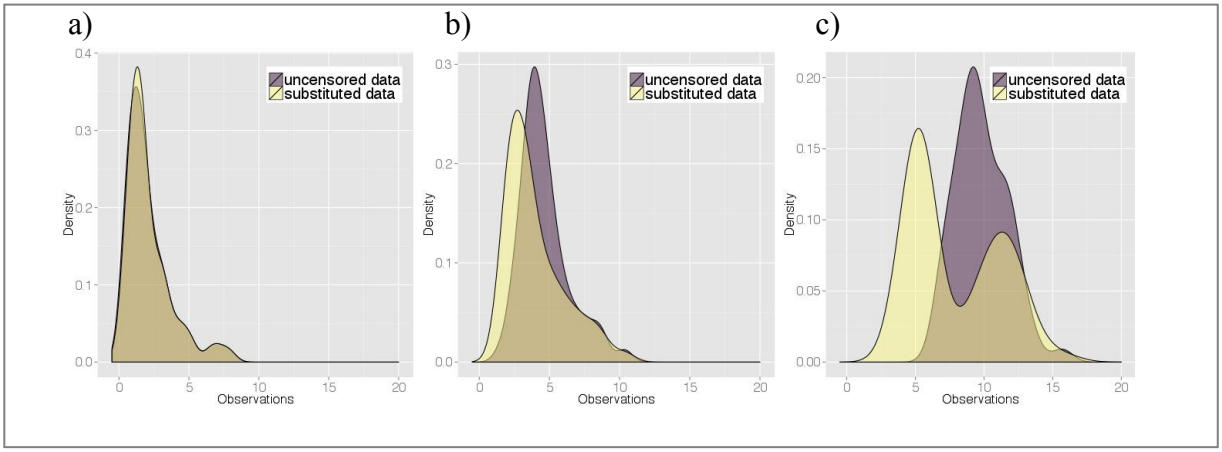


Figure 4.4 The distributions of original and substituted data generated from lognormal distributions with σ=1.9 and different μ values a) $\mu = 2$, b) $\mu = 5$, and c) $\mu = 10$

Noticeable is that, in the case of μ=2 (Figure 4.4a), the shape of the distribution of uncensored and substituted data is almost similar. However, when μ increases (Figure 4.4b and Figure 4.4c), substituting the censored values introduces a peak at the substituted value, leading to an incorrect characterization of the shape of the distribution. In fact, as observed in Figure 4.4c, substitution of censored values generates a bimodal distribution, which is far from the shape of the original distribution. We also investigate whether the distribution

remains lognormal after the substitution of censored observations visually by the quantile-quantile (Q-Q) plots and formally by the Shapiro-Wilk test. The Q-Q plots for the substituted data for the three aforementioned scenarios (i.e., lognormal distribution, $\mu = 2,5,10$ and $\sigma = 1.9$, 50% censoring), are shown in Figure 4.5. All Q-Q plots in this Figure include clusters of horizontal points of substituted values, changing the initial distribution of data sets. However, the impact of substitution is more pronounced in some simulation scenarios. For example, the Q-Q plot of the substituted data simulated from μ=2 (Figure 4.5a) appears roughly linear, indicating that the data set after substitution of the censored observation may be still lognormal. As μ increases, substantial deviation from linearity indicates that the substituted data no longer follow the lognormal distribution (Figure 4.5b and Figure 4.5c). A Shapiro-Wilk test provides p-values smaller than $\alpha = 0.05$, rejecting the normality of log-transformed data in all three scenarios.

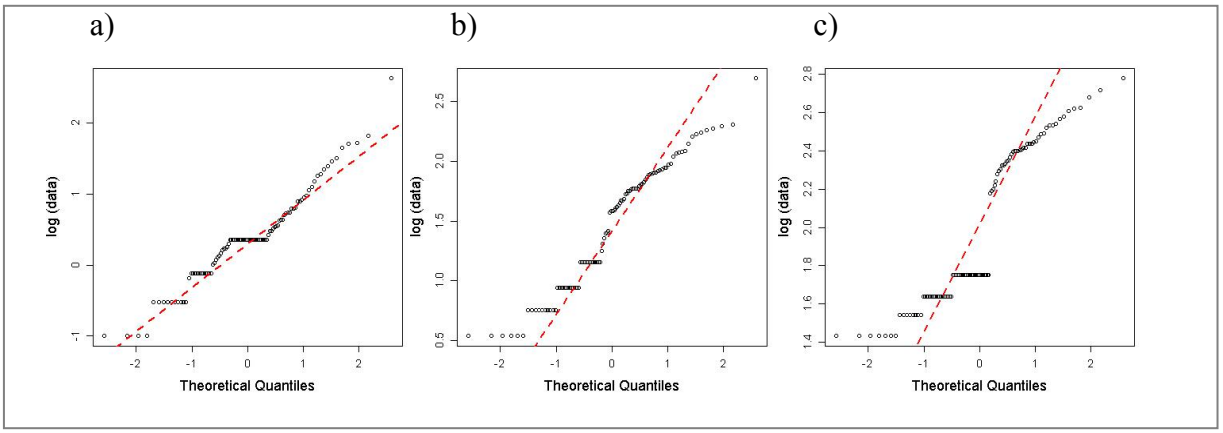


Figure 4.5 The Q-Q plots of substituted data generated from lognormal distribution with σ=1.9 and different μ values a) $\mu = 2$, b) $\mu = 5$, and c) $\mu = 10$

To provide a better demonstration of the shortcomings of the substitution-based method in estimating the mean and standard deviation compared to the alternative methods, Figure 4.6 illustrates MSEs of the substitution-based method for different combinations of μ and σ. The following can be inferred from Figure 4.6:

1. Related to the mean estimation, Figure 4.6a shows that, for a given σ, substitution produces larger MSEs as μ increases. Moreover, for any μ>4, substitution produces larger MSEs as σ decreases;

2. Obtaining good estimates of the standard deviation (i.e., estimates with small MSEs) is largely influenced by the underlying σ of data (Figure 4.6b). For example, when the underlying σ of the simulated data is 0.5, the MSE values increase as μ increases. On the other hand, when the underlying σ of the simulated data is 1.9, the MSEs initially decrease (up to μ =4) and then start increasing as μ becomes larger. This implies that the performance of traditional methods after substitution of the censored values is difficult to predict before knowing the distributional characteristics of the data.

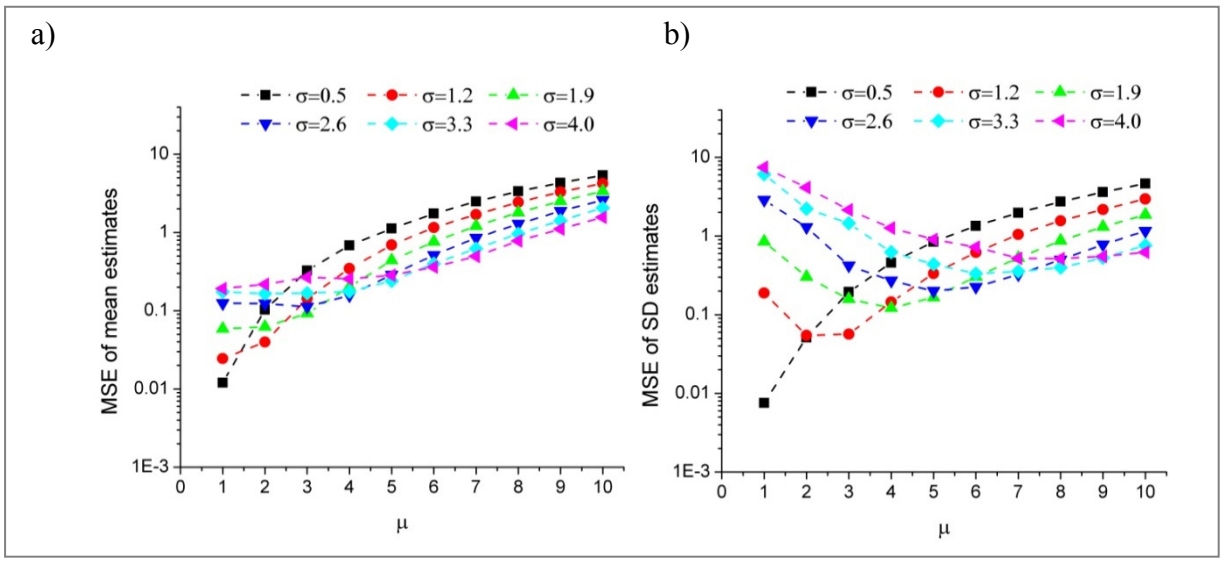


Figure 4.6 The MSEs of the substitution method in estimating a) the mean and b) standard deviation for different combinations of μ and σ of lognormal distribution

Figure 4.7a and Figure 4.7b illustrate the MSEs produced by MLE in estimating the mean and standard deviation of lognormal distributions with different combinations of the parameters μ and σ. For conciseness, only the results obtained from the MLE method are depicted since plots produced by the rROS and KM methods are similar. Plots relative to the rROS and KM methods are illustrated in Figure-A IV-3 and Figure-A IV-4 of Appendix IV. Comparison of Figure 4.7a and Figure 4.7b implies that MSEs of MLEs of the mean are approximately constant over different values of μ, whereas this behavior is not clearly observed for estimating the standard deviation. In fact, large MSEs are obtained at the left-end of the curves in Figure 4.7b, corresponding to moderate and highly skewed distributions (CV>1). This behavior is not surprising as Singh et al. (2006) and Shoari, Dubé & Chenouri

(2015) agreed that the performance of estimators in the case of moderately to highly skewed data sets differs from that of the mildly skewed data sets.

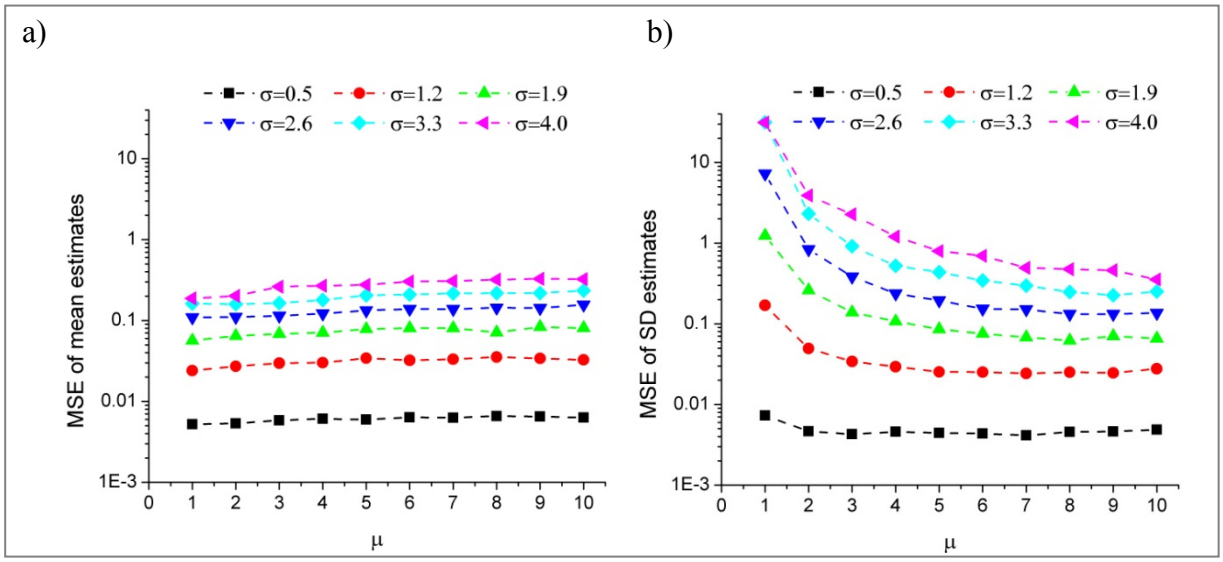


Figure 4.7 The MSEs of the MLE method in estimating a) the mean and b) standard deviation for different combinations of μ and σ of lognormal distribution

4.5.2 Data from Weibull and gamma distribution

In simulation scenarios based on the data sets generated from Weibull and gamma distributions, substituting censored values with a constant does not consistently lead to better or worse estimates than those provided by MLE, rROS, and KM. This result is in agreement with observations made for lognormal data sets discussed in Section "Data from lognormal distribution". The related Figures are available in Appendix IV (Figure-A IV-5 through Figure-A IV-8). Moreover, simulation results confirm that the performance of substitution-based estimators of the mean and standard deviation depends upon distributional parameters no matter whether the underlying distribution is Weibull or gamma (Figure 4.8 and Figure 4.9).

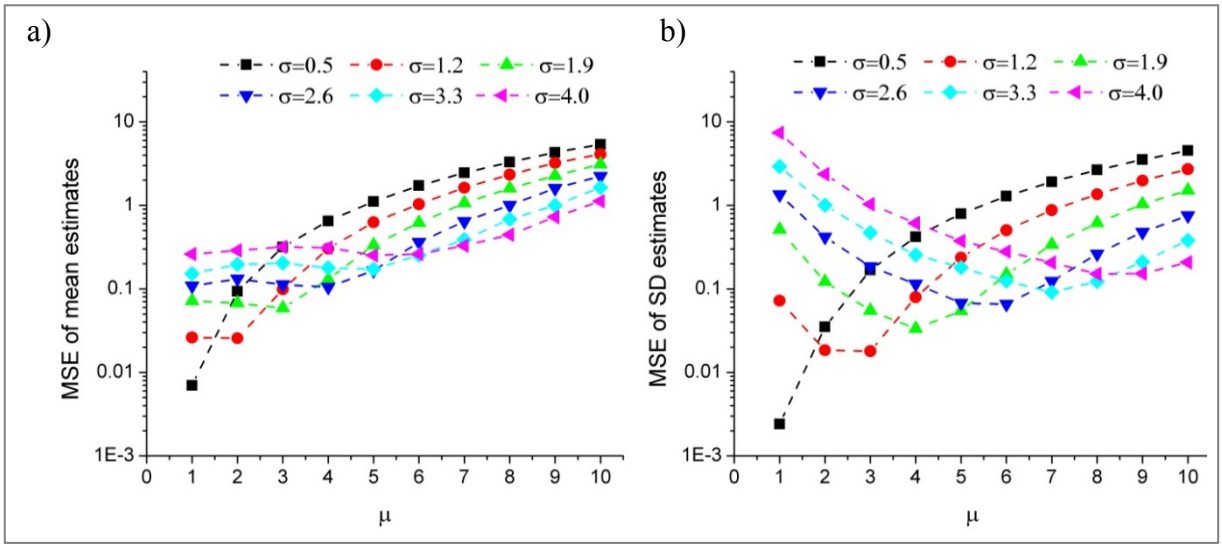


Figure 4.8 The MSEs of the substitution method in estimating a) the mean and b) standard deviation for different combinations of μ and σ of Weibull distribution

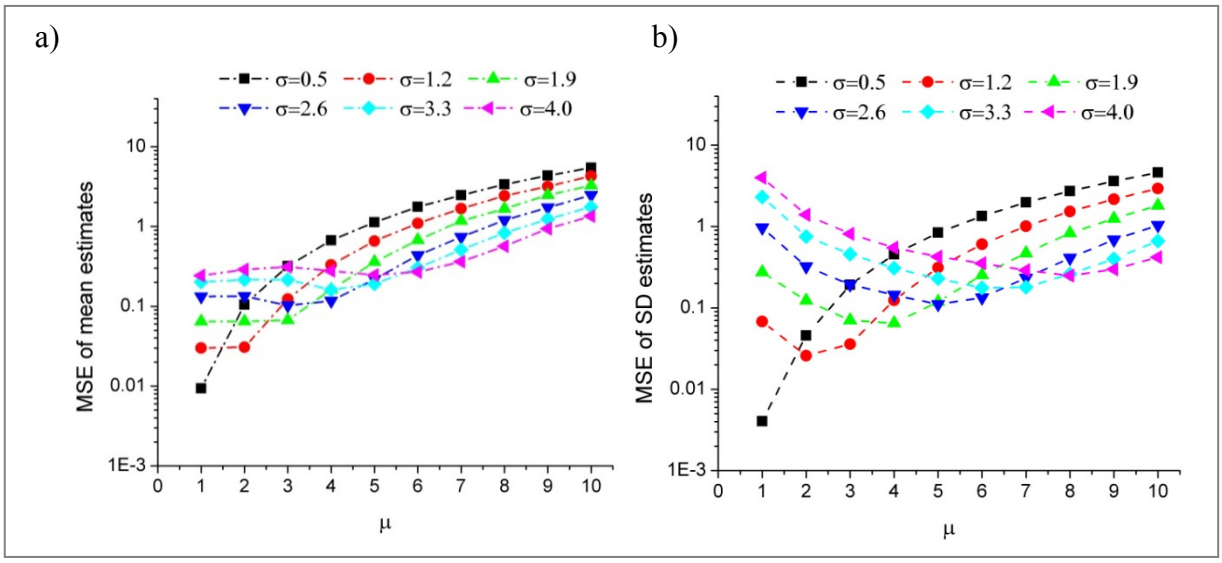


Figure 4.9 The MSEs of the substitution method in estimating a) the mean and b) standard deviation for different combinations of μ and σ of gamma distribution

4.6 Why not substituted even for small censoring percent?

To investigate the adequacy of substitution for small percentages of censoring, the simulation experiment was repeated for scenarios with only 10% censoring. As expected, simulation scenarios with 10% censoring generally result in smaller MSEs than those with 50%

censoring. However, this decrease is not systematic and its magnitude depends on the CV of the underlying distribution. Table 4.1 reports the average percent reduction in MSE for different distributions and CVs. As shown in Table 4.1, reducing the percentage of censoring to 10% substantially improves the performance of traditional estimators of the mean and standard deviation only in CV<0.5 simulation scenarios. When CV exceeds 0.5, MSEs either decrease slightly or, surprisingly, increase in some cases. Despite the recommendation of some environmental guidelines to use substitution for handling data sets with small amounts of censoring (e.g., USEPA, 2006; 2009), the simulation results reported herein show the inadequacy of substitution even for data with small percentages of censoring. Note that the inadequacy of substitution is also reported in other studies, notably El-Shaarawi & Esterby (1992), Singh & Nocerino (2002), and Helsel (2006).

Table 4.1 Averaged percent reduction in the MSE of the substitution-based method when the censoring percentage is reduced to 10%

	Estimation of the mean			**Estimation of the standard deviation**		
CV	**Lognormal**	**Weibull**	**Gamma**	**Lognormal**	**Weibull**	**Gamma**
CV<0.5	88%	86%	87%	62%	55%	62%
0.5≤CV<2	19%	12%	13%	6%	14%	10%
CV≥2	-10%	-2%	7%	-13%	-7%	4%

4.7 Summary and conclusions

This study investigated the performance of the substitution-based estimators of the mean and standard deviation of a distribution based on left-censored observations. Monte Carlo simulation results revealed that the performance of the substitution-based method depends on the intrinsic distributional characteristics of the lognormal, Weibull, and gamma distributions. This finding is in accordance with El-Shaarawi & Esterby (1992). They analytically demonstrated that the performance of substitution in estimating the parameters of

normal and lognormal distributions based on censored data with only one DL depends on distributional characteristics. This paper extended the conclusions to other right-skewed distributions such as Weibull and gamma distributions with different levels of skewness. In addition, we considered more realistic situations in which multiple DLs exist as these thresholds change among different laboratories/measuring instruments or over time.

Generally, the substitution method resulted in less reliable estimates than those obtained from the alternative methods. Only for certain pairs of μ and σ, substitution provided reliable estimates. However, it has to be stressed that in real environmental studies the intrinsic characteristics of populations under study are unknown. Therefore, it cannot be determined a priori if a given population is characterized by these specific μ and σ for which substitution would provide more reliable estimates than the alternative methods. The alternative methods discussed in this paper were found to be less sensitive to distributional parameters. In particular, the performance of these methods in estimating the mean was independent from the magnitude of the mean, for any σ. However, these methods produced larger MSEs for estimating the standard deviation when distributions were moderate to highly skewed (i.e., CV>1 herein).

Despite recommendations of some environmental guidelines on using the substitution method for estimating distributional parameters of contaminant concentration data with low percentage censoring, simulation results in this paper showed the inadequacy of this estimator even for these situations. We recommend practitioners adopt one of the alternative methods when analyzing data with censored observations, and avoid substituting censored data with arbitrary constants.

CHAPTER 5

AN INVESTIGATION OF THE IMPACT OF LEFT-CENSORED SOIL CONTAMINATION DATA ON THE UNCERTAINTY OF DESCRIPTIVE STATISTICAL PARAMETERS

Niloofar Shoari[a,b], Jean-Sébastien Dubé[a]

[a] Department of Construction Engineering, École de technologie supérieure, Université du Québec, 1100, rue Notre-Dame Ouest, Montréal, Québec, Canada, H3C 1K3.
[b] Department of Statistics, Texas A&M University, 3143 TAMU, College Station, Texas, USA, 77840.

This article has been published in the journal of
Environmental Toxicology and Chemistry in March, 2016.

5.1 Abstract

Left-censored concentration data are frequently encountered because measuring instruments cannot detect concentrations below instruments detection limit (DL). For statistical analysis of left-censored data, environmental literature mainly refers to the following methods: maximum likelihood estimation (MLE), regression on order statistics using lognormal and gamma assumption (rROS and GROS, respectively), and Kaplan-Meier. A number of simulation studies examined the performance of these methods in terms of bias and/or mean square error. However, no matter which method is adopted, some uncertainty is introduced into outcomes since all is known about a left-censored observation is that the concentration falls between 0 and the DL. Data used here come from analysis of soil samples collected for a site characterization in Montreal, Canada. Employing non-parametric bootstrap, we quantify the uncertainty and bias in the mean and standard deviation estimates obtained by the MLE (under lognormal, Weibull, and gamma distributions), rROS, GROS, and KM methods. First, we demonstrate that the highest uncertainty is associated with MLEs under lognormality and Weibull assumptions while a gamma assumption leads to estimates with less uncertainty. Second, we show that although an increase in sample size improves the uncertainty, it reduces the bias only in the rROS, GROS, and KM methods. Finally,

comparing percentage uncertainty in the mean of contaminant data, we illustrate that adopting an inappropriate estimator results in large uncertainties.

5.2 Introduction

Collecting representative soil samples is a crucial step to gain an unbiased and precise knowledge on levels of contamination of the soil population under study. Strategies required to ensure that a representative sample is obtained focus on minimizing the uncertainties associated with sampling protocol, sample preparation and chemical analysis; some examples can be found in Gerlach, Dobb, Raab & Nocerino (2002); Nocerino, Schumacher & Dary (2005); Boudreault, Dubé, Sona & Hardy (2012) and Dubé et al. (2015). Once a representative sample is obtained, the estimation of statistical parameters (e.g., the mean and standard deviation) of the resultant concentration measurements actually represents the contamination levels in the soil population under study. This procedure is usually straightforward when the precision of measuring instruments is sufficient to detect the presence of the contaminant from the background noise. However, we frequently encounter left-censored observations that are concentration measurements falling between 0 and the detection limit of measuring instruments. The lack of knowledge on the true concentration of left-censored observations leads to inferences with some amount of uncertainty.

The common practice to treat left-censored data has been to substitute censored values with an arbitrary constant (e.g., half DL) and to use standard techniques to analyze data. However, numerous studies (e.g., Hewett & Ganser, 2007 and Gilliom & Helsel, 1986) expressed concern about the biased estimates obtained by the substitution approach. Shoari, Dubé and Chenouri (2016) discouraged the use of the substitution method as they demonstrated that the reliability of the estimates after substitution is highly sensitive to intrinsic characteristics of a population (such as mean, standard deviation, skewness), which cannot be known a priori. More recently, several alternative methods are available in the literature to address the problem of left-censoring. The most common ones to estimate distributional parameters of a population are: i) maximum likelihood estimation (MLE) under lognormal, Weibull, and

gamma distributional assumption, ii) robust regression on order statistics (rROS), iii) gamma regression on order statistics (GROS), and iv) Kaplan-Meier (KM). Estimates obtained by these estimators contain some amount of bias and uncertainty.

In studies based on Monte Carlo simulations, the bias and uncertainty are quantifiable because the true values of parameters are known. Some examples of Monte Carlo simulation studies for left-censored data can be found in Kroll & Stedinger (1996), Sinha, Lambert and Trumbull (2006), Hewett & Ganser (2007), European Food Safety Authority (2010), and Shoari et al. (2015), among others. Typically, the mean square error (MSE) is used as criteria to reflect both bias and uncertainty of the estimates in each simulation scenario. However, the issue arises when quantifying the bias and uncertainty of the estimates based on real concentration data is of interest because the true parameters are unknown. Bootstrapping is a data-based simulation that circumvents this issue. In bootstrapping, random samples are repeatedly drawn from an approximation distribution (based on the original dataset) and the statistics of interest are estimated in each sampling event. The resultant replications of the bootstrapped statistics are used as the basis for computing the approximated bias and the uncertainty in the sample estimate of the unknown parameter. Based on complete (without censoring) simulated data, Tong, Chang, Jin & Saminathan (2012) and Tong, Saminathan & Chang (2016) concluded that bootstrap provides reliable uncertainty estimates for data with small sample sizes for a variety of data distributions (normal, lognormal, uniform, Weibull, gamma, and beta). Frey & Zhao (2004) proved the reliability of the bootstrapping technique in estimating the uncertainty of the mean estimates obtained by MLE when the simulated data were left-censored. Some other examples regarding the application of bootstrapping are in Zhao & Frey (2006), and Babamoradi, van den Berg & Rinnan (2013).

Environmental exposure assessments are based on exposure models that combine contaminant concentration levels and exposure time and pathways to predict a population exposure to a certain contaminant. To represent concentration levels in exposure models, estimates of statistical parameters of contamination data (e.g., the mean value) serve as input. Uncertainty of input data (due to the presence of left-censored observations) contributes to

the uncertainty of output. Using the bootstrap method on concentration data from a characterization study, we aim at quantifying the bias and, more importantly, the uncertainty of the estimates of the mean and standard deviation provided by the aforementioned estimators. The concentration data used in the present study are contaminant concentration measurements of soil samples collected for the purpose of a site characterization study in Montreal, Canada. In the present study, we assume that the sampling uncertainty has been minimized and consequently, representative samples have been obtained. Under this assumption, we show that inadequate analysis of left-censored concentration data generates an additional source of uncertainty, which is reflected in the estimates of the concentration mean and standard deviation.

5.3 Case study

We consider concentration data sets obtained from soil samples collected for a site characterization study conducted in Montreal, Canada (Quéformat Ltée., 2004 and Groupe Qualitas inc., 2010). Data sets consist of concentration measurements of 15 inorganic and 53 organic contaminants in soil samples collected from 45 sampling locations. For each contaminant, the sample size varies between 13 and 62 observations and the censoring percentage ranges between 0% and 100%. In the present study, we consider only those contaminants with the censoring percentage of 2%-80%; Table 5.1 reports the sample size and the censoring percentage for each contaminant. In the results section, the contaminants discussed are numbered for ease of reference to Table 5.1.

Table 5.1 The sample size and censoring percentage for each contaminant

	Contaminant	**Sample size**	**Censoring %**
1	As	51	2%
2	Hg	13	15%
3	Phenanthrene	62	16%
4	Fluoranthene	62	18%
5	Pyrene	62	19%
6	Chrysene	62	23%
7	Benzo (b,j,k)fluoranthene	62	23%
8	Benzo (a) pyrene	61	25%
9	Benzo (a) anthracene	61	27%
10	Sn	51	29%
11	Mo	51	33%
12	Indeno (1,2,3-cd) pyrene	61	36%
13	Benzo (g,h,i) perylene	61	38%
14	Anthracene	62	39%
15	Benzo (a,h) anthracene	61	44%
16	1-Methyl naphthalene	62	45%
17	2-Methyl naphthalene	62	45%
18	Acenaphtene	62	50%
19	1,3-Dimethyl naphthalene	62	52%
20	Acenaphtylene	62	52%
21	Naphthalene	62	53%
22	Cd	51	67%
23	2,3,5- Trimethyl naphthalene	62	77%

5.4 Methodology

In the present study, we use the bootstrap technique to evaluate the quality of the mean and standard deviation estimates that are obtained from the methods of MLE (under lognormal, Weibull, and gamma assumption), rROS, GROS, and KM. Some general definitions for bootstrapping are presented in this section. The bootstrap method is based on the assumption that the sample is representative of the population under study, and that the observations are independently and identically distributed. Under these assumptions, bootstrapping enables us to show the substantial uncertainty associated with statistical inferences. In the present study, like previous studies (e.g., Frey & Zhao, 2004; Zhao & Frey, 2006), we assume that our concentration data set satisfies both conditions mentioned above. Our statistical inferences are therefore given based on identically distributed observations and representativeness assumptions.

Suppose $X = \{x_1, x_2, \dots, x_n\}$ is a random sample of size n drawn from a population with an unknown distribution f. And, let the distribution $\hat{f}$ be a parametric or non-parametric estimate of f. Essentially, bootstrapping consists of taking a large number of bootstrap samples $X_i^* = \{x_{i,1}^*, x_{i,2}^*, \dots, x_{i,n}^*\}$, $i = 1,2, \dots, B$ from the distribution $\hat{f}$. In the case $\hat{f}$ is defined non-parametrically (by an empirical distribution function), the bootstrap method is referred to as non-parametric bootstrapping. This involves taking independent samples drawn with replacement from the original data set B times. A parametric bootstrap is performed when $\hat{f}$ is estimated by fitting a parametric model to the data (using the maximum likelihood estimation method for instance), and bootstrap samples are simulated from the fitted model. More details and applications of bootstrapping can be found in Efron (1981), Efron & Tibshirani (1986), and Davison & Hinkley (1997). Due to lack of knowledge of a specific family of parametric models that describes the original data, we limit our attention to the non-parametric bootstrap method. Detailed steps of the adopted approach for left-censored data are described as follows.

Step 1: Organize the original data set as pairs of $(x_i, \delta_i), i = 1,2, \dots, n$, where x_i is the i^{th} observable concentration value and δ_i is a binary indicator function that defines whether the observable concentration is censored or not. The indicator function δ_i takes the value 1 if x_i is uncensored, and 0 otherwise.

Step 2: Construct bootstrap samples by taking random samples with replacement *n* times, where *n* is the sample size of the original data set. This is achieved by simultaneous sampling of both observable concentration and its corresponding indicator so that the bootstrap sample gets the form $X^* = \{(x_i^*, \delta_i^*)\}, i = 1,2, \dots, n.$

Step 3: For the bootstrap sample, calculate the statistic of interest, $\hat{\theta}^*$ by the MLE (under lognormal, Weibull, and gamma distribution), rROS, GROS, and KM methods. In the present study, the statistic of interest is the mean, $\bar{x}^*$, and standard deviation, s^*, of the bootstrap sample. Details regarding the computation of the aforementioned estimators can be found in Hewett & Ganser (2007), Helsel (2012), and Shoari et al. (2015)

Step 4: Repeat steps 2 and 3 a large number of times, say *B*=1000, so that we have a sequence of bootstrap estimates, $\hat{\theta}_b^*, b = 1,2, \dots, B$.

Step 5: Construct the approximated bias and the 95% confidence interval by using the equations described in the next section.

5.4.1 Bootstrap approximated bias and confidence interval

The approximated bias of $\hat{\theta}^*$ is given by

$$\varepsilon = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}) \tag{5.1}$$

where $\hat{\theta}$ is the estimated statistic from the original data. Among various estimators of $\hat{\theta}$, we resort to the rROS method because previous simulation studies showed the good performance of rROS in a wide range of simulation scenarios (Gilliom & Helsel, 1986; Hewett & Ganser, 2007; Shoari et al., 2015).

The uncertainty is determined in terms of the length of the bootstrap confidence intervals. Two methods used in the present study to calculate bootstrap confidence intervals are the percentile and the bias corrected and accelerated percentile (BCa) methods. In the percentile method, the $(1-2\alpha)100\%$ confidence interval is calculated as

$$[\hat{\theta}^{*(\alpha B)}, \hat{\theta}^{*(1-\alpha)B}] \tag{5.2}$$

where $\hat{\theta}^{*(\alpha B)}$ and $\hat{\theta}^{*(1-\alpha)B}$ are the α^{th} and $(1-\alpha)^{\text{th}}$ values of the ordered estimates of the statistic of interest, $\hat{\theta}_b^*, b = 1,2,\dots,B$. If the distribution of B $\hat{\theta}^*$s does not resemble the curve of a normal distribution, the confidence intervals of the percentile method may be biased. To adjust for this bias, Efron & Tibshirani (1994) suggested using the BCa confidence intervals as

$$[\hat{\theta}^*(q_L), \hat{\theta}^*(q_U)] \tag{5.3}$$

where $\theta^*(q_L)$ and $\theta^*(q_U)$ are the q_Lth and q_Uth value of ordered $\hat{\theta}_b^*$, b= $1,2,\dots,B$. The values of q_L and q_U are given as

$$q_L = \Phi(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a\left(z_0 + z_{\alpha/2}\right)}) \tag{5.4}$$

$$q_U = \Phi(z_0 + \frac{z_0 + z_{(1-\alpha/2)}}{1 - a\left(z_0 + z_{(1-\alpha/2)}\right)}) \tag{5.5}$$

where $z_{\alpha/2}$ and $z_{(1-\alpha/2)}$ are the α^{th} and $(1-\alpha/2)^{\text{th}}$ quantiles of the standard normal distribution, z_0 and a are bias-correction and acceleration factors, respectively, and are given as

$$a = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(-i)}^{*} - \theta_{(-)}^{*})^{3}}{6\left(\sum_{i=1}^{n}(\hat{\theta}_{(-i)}^{*} - \theta_{(-)}^{*})^{2}\right)^{\frac{3}{2}}} \tag{5.6}$$

$$z_0 = \Phi^{-1}(\#\{\hat{\theta}_b^{*} < \hat{\theta}\}/B) \tag{5.7}$$

where $\hat{\theta}_{(-i)}^{*}$ is the value of $\hat{\theta}^{*}$ when the ith observation is deleted from the original data and $\theta_{(-)}^{*}$ is given by $\theta_{(-)}^{*} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)}^{*}$. Moreover, $\Phi(.)$ is the cumulative standard normal distribution function. All calculations were implemented in R statistical software.

5.5 Results

5.5.1 Uncertainty and approximated bias of the estimates

The performance of the MLE (under lognormal, Weibull, and gamma distributional assumption), GROS, and KM methods was evaluated using the approximated bias and length of confidence intervals around the estimates. Since the distributions of the bootstrap estimates of the mean and standard deviation are skewed, we use the BCa method for an accurate calculation of confidence intervals. The results are presented in Table 5.2 and Table 5.3. Overall, while the lengths of bootstrap confidence intervals provided by the rROS, GROS, and KM estimators are comparable, the length of confidence intervals of MLEs is strongly dependent on the distributional assumption. Table 5.2 clearly shows that the highest amount of uncertainty (i.e., largest confidence intervals) is attributed to the MLEs under lognormality assumption followed by those under Weibull and gamma assumptions.

Regarding the bias, Table 5.3 shows that there is not a single estimator that universally outperforms the others. Noticeably, the MLE estimator under the lognormality assumption systematically overestimates bias. Moreover, the bias of the MLE estimator is sensitive to the distributional assumption, censoring percentage, and the type of the statistic one wishes to estimate while the rROS, GROS, and KM estimators have the same magnitude in bias. In the statistical literature, the bias of $\hat{\theta}^{*}$ can be accepted only if it tends to vanish as the sample

size increases (for example, see Liero & Zwanzig, 2011). This property is defined as asymptotic unbiasedness where the $\hat{\theta}^*$ approaches the true value θ, as the sample size n tends to ∞. In the following section, "impact of sample size on the uncertainty and approximated bias of the estimates", we show that the bias can be negligible when the sample size is sufficiently large.

Table 5.2 The length of bootstrap confidence intervals for the mean and standard deviation estimates obtained by different methods[a]

	MLE (lognormal)	**MLE (Weibull)**	**MLE (gamma)**	**KM**	**rROS (lognormal)**	**GROS (gamma)**
1	5.1(9.6)	4.2(3.5)	4.2(4.0)	4.2(3.1)	4.2(3.1)	4.2(3.1)
2	1.6(30.4)	0.8(1.7)	0.8(1.2)	0.8(0.9)	0.8(0.9)	0.8(0.9)
3	263.9(*)	36.1(187.9)	38.8(90.5)	38.8(85.0)	38.8(84.8)	38.8(84.8)
4	742.0(*)	43.8(275.0)	42.3(108.0)	42.3(107.1)	42.3(107.0)	42.3(106.9)
5	617.4(*)	38.9(254.8)	34.3(87.1)	34.4(80.3)	34.4(80.1)	34.4(80.1)
6	204.6(*)	20.9(156.6)	21.4(51.6)	21.5(50.3)	21.5(50.1)	21.5(50.1)
7	216.3(*)	20.8(110.3)	15.5(35.8)	15.5(36.2)	15.4(36.2)	15.5(36.2)
8	327.2(*)	20.8(190.3)	17.9(50.5)	17.9(44.1)	17.9(43.9)	17.9(43.9)
9	330.5(*)	25.0(184.1)	21.0(56.2)	21.0(50.8)	21.0(50.6)	21.0(50.6)
10	47.3(678.3)	27.8(59.9)	26.5(42.2)	26.0(34.8)	26.5(34.8)	26.7(34.6)
11	0.8(0.9)	0.8(0.5)	0.8(0.7)	0.7(0.5)	0.9(0.4)	1.0(0.5)
12	260.4(*)	10.7(140.8)	8.6(23.0)	8.6(21.3)	8.6(21.2)	8.6(21.2)
13	315.1(*)	13.7(111.8)	9.0(25.7)	9.0(21.9)	9.0(21.7)	9.0(21.7)
14	149.3(*)	11.0(98.0)	9.6(26.4)	9.6(24.4)	9.7(24.3)	9.7(24.3)
15	15.2(*)	3.4(27.6)	3.2(8.3)	3.3(7.8)	3.2(7.7)	3.2(7.7)
16	2.0(79.5)	1.0(4.0)	1.0(2.4)	1.0(2.2)	1.0(2.2)	1.0(2.2)
17	4.0(923.2)	1.4(8.8)	1.3(3.4)	1.3(3.0)	1.3(3.0)	1.3(3.0)
18	25.2(*)	4.2(43.8)	3.4(10.0)	3.4(8.4)	3.4(8.3)	3.4(8.3)
19	4.0(704.3)	1.2(6.7)	1.1(2.8)	1.1(2.3)	1.1(2.3)	1.1(2.3)
20	20.0(*)	2.4(42.3)	1.7(4.7)	1.7(4.3)	1.7(4.2)	1.7(4.2)
21	138.0(*)	6.3(85.3)	2.8(8.8)	2.8(6.7)	2.8(6.7)	2.8(6.7)
22	0.5(0.8)	0.5(0.5)	0.5(0.5)	0.3(0.4)	0.5(0.5)	0.6(0.5)
23	77.9(*)	0.8(74.7)	0.4(1.4)	0.4(1.1)	0.4(1.0)	0.4(1.0)

[a] Values in parentheses represent the length of confidence interval for the standard deviation estimates. The * represents values larger than 1000. The numbers in column 1 refer to the contaminants listed in Table 5.1

Table 5.3 Bias of the mean and standard deviation estimates obtained by different methods[a]

	MLE (lognormal)	**MLE (Weibull)**	**MLE (gamma)**	**KM**	**rROS (lognormal)**	**GROS (gamma)**
1	0.41(2.69)	0.06(-0.28)	0.03(-0.17)	0.06(-0.12)	0.04(-0.11)	0.02(-0.08)
2	0.07(0.87)	0.00(-0.12)	0.00(-0.18)	0.01(-0.11)	0.00(-0.11)	0.00(-0.10)
3	30.85(*)	-1.76(-14.47)	0.03(-30.61)	0.05(-9.64)	0.04(-9.73)	0.03(-9.73)
4	65.94(*)	-0.82(-7.12)	0.19(-36.93)	0.21(-11.48)	0.20(-11.60)	0.19(-11.60)
5	56.75(*)	-0.02(0.85)	0.09(-25.96)	0.10(-8.15)	0.09(-8.26)	0.08(-8.25)
6	23.56(*)	-0.54(-2.74)	-0.11(-16.86)	-0.09(-5.80)	-0.11(-5.88)	-0.12(-5.87)
7	34.45(*)	0.84(11.93)	-0.07(-6.58)	-0.05(-1.88)	-0.06(-1.94)	-0.07(-1.94)
8	20.76(*)	-0.53(-1.51)	-0.17(-14.36)	-0.15(-4.93)	-0.17(-5.01)	-0.18(-5.00)
9	35.51(*)	0.01(2.44)	-0.18(-16.91)	-0.16(-6.12)	-0.17(-6.22)	-0.18(-6.21)
10	7.47 (91.81)	0.03 (5.01)	-0.21 (-1.79)	1.12 (-2.03)	0.02 (-1.56)	-0.669 (-1.11)
11	0.05(0.04)	-0.03(0.00)	0.00(-0.02)	0.23(-0.23)	0.01(-0.03)	-0.13(0.12)
12	19.00 (*)	0.49 (5.70)	-0.10 (-5.62)	-0.069 (-2.19)	-0.08 (-2.25)	-0.11 (-2.24)
13	24.31(*)	0.81(8.98)	0.02(-5.16)	0.07(-1.70)	0.05(-1.77)	0.02(-1.76)
14	17.97(*)	0.42(4.57)	0.04(-7.14)	0.09(-2.35)	0.06(-2.43)	0.03(-2.42)
15	1.47(102.15)	-0.10(-0.66)	-0.04(-2.43)	0.01(-0.88)	-0.02(-0.91)	-0.05(-0.90)
16	0.20(5.40)	-0.01(-0.14)	0.01(-0.51)	0.05(-0.13)	0.02(-0.13)	0.00(-0.12)
17	0.27(12.09)	-0.05(-0.20)	0.00(-0.77)	0.04(-0.21)	0.00(-0.22)	-0.01(-0.21)
18	2.62(355.69)	0.06(1.31)	0.03(-2.17)	0.09(-0.68)	0.05(-0.71)	0.02(-0.70)
19	0.60(23.18)	0.04(0.40)	0.00(-0.46)	0.05(-0.19)	0.01(-0.20)	-0.01(-0.19)
20	0.77 (132.41)	-0.02 (0.76)	0.00 (-0.82)	0.04 (-0.21)	0.00 (-0.23)	-0.019 (-0.22)
21	4.73(*)	0.26(5.53)	0.00(-0.83)	0.06(-0.31)	0.01(-0.34)	-0.01(-0.34)
22	0.01(0.08)	-0.06(0.04)	-0.06(0.04)	0.35(-0.22)	0.02(-0.03)	-0.20(0.11)
23	0.59 (*)	0.06 (1.33)	-0.01 (-0.02)	0.09 (-0.06)	0.01 (-0.06)	-0.029 (-0.05)

[a] Values in parentheses represent the length of confidence interval for the standard deviation estimates. The * represents values larger than 1000. The numbers in column 1 refer to the contaminants listed in Table 5.1

5.5.2 Impact of sample size on the uncertainty and approximated bias of the estimates

To evaluate the asymptotic unbiasedness of different estimators, the non-parametric bootstrap method is repeated for all contaminant data except, instead of drawing samples with the same amount of observations as the original data, the sample size gradually increases. The confidence interval and bias are calculated for each sample size. As some examples, Figure 5.1 through 5.4 illustrate the bias and confidence interval versus sample size for four contaminants, acenaphtene, benzo(a)anthracene, naphthalene, and chrysene. The results for other contaminants are presented in Appendix V.

Figure 5.1 and Figure 5.2 show the length of bootstrap confidence intervals around the mean and standard deviation estimates as a function of increasing sample size. For a better interpretation, the y-axis is represented in logarithmic scale. Also, the scale of y-axes can be different for the purpose of illustrating both small and large values. For all contaminants, as the sample size increases, the bootstrap confidence intervals becomes smaller and therefore, less uncertainty is associated with the estimates. Even so, the maximum likelihood estimates obtained under lognormality assumption have larger confidence interval lengths compared with other estimators even for sample sizes as large as 620. After maximum likelihood estimates based on lognormality, the largest uncertainty is generally attributed to the maximum likelihood estimates obtained under Weibull assumption. However, unlike the lognormal MLE, for some contaminants (e.g., chrysene), an increase in sample size results in confidence interval lengths comparable to those obtained by other estimators. We notice similar uncertainties of the mean and standard deviation estimates obtained by MLE under gamma assumption, rROS, GROS, and KM.

Figure 5.3 and Figure 5.4 represent the approximated bias of different estimators of the mean and standard deviation as a function of increasing samples size for the four aforementioned contaminants. An increase in sample size does not substantially reduce the bias of the mean estimates although larger sample sizes impact favorably on reducing the bias of the standard deviation estimates. Note that the results of MLE under lognormal assumption are not

illustrated in Figure 5.3 and Figure 5.4 because, for all contaminants, no matter how much we increase the sample size, the MLE method under lognormal distribution provides the largest bias compared to other estimators. Related to the MLE method under Weibull and gamma assumptions, the approximated bias remains approximately unchanged as the sample size increases. In contrast, the rROS, GROS, and KM methods show to be asymptotically unbiased. Combining the results of bias and confidence interval length (Figure 5.1 through 5.4), we observe that the MLE method under Weibull and gamma assumptions estimates the wrong values with small amount of uncertainty regardless of the sample size. The KM, rROS, and GROS methods generally provide estimates with small amounts of bias and uncertainty and thus are recommended in the present study.

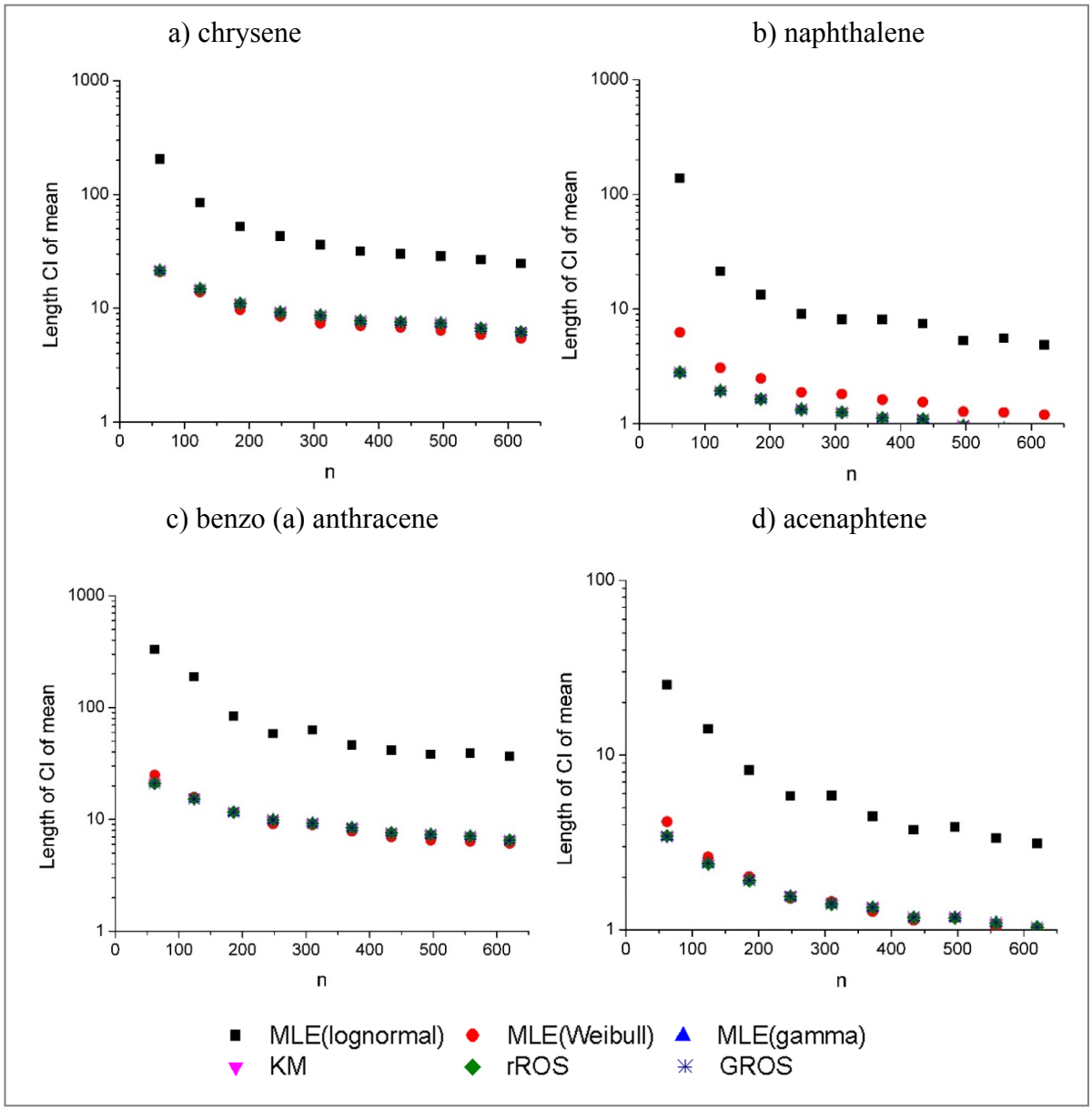


Figure 5.1 Bootstrap confidence interval lengths around the mean estimate of a) chrysene, b) naphthalene, c) benzo(a)anthracene, and d) acenaphtene concentration data

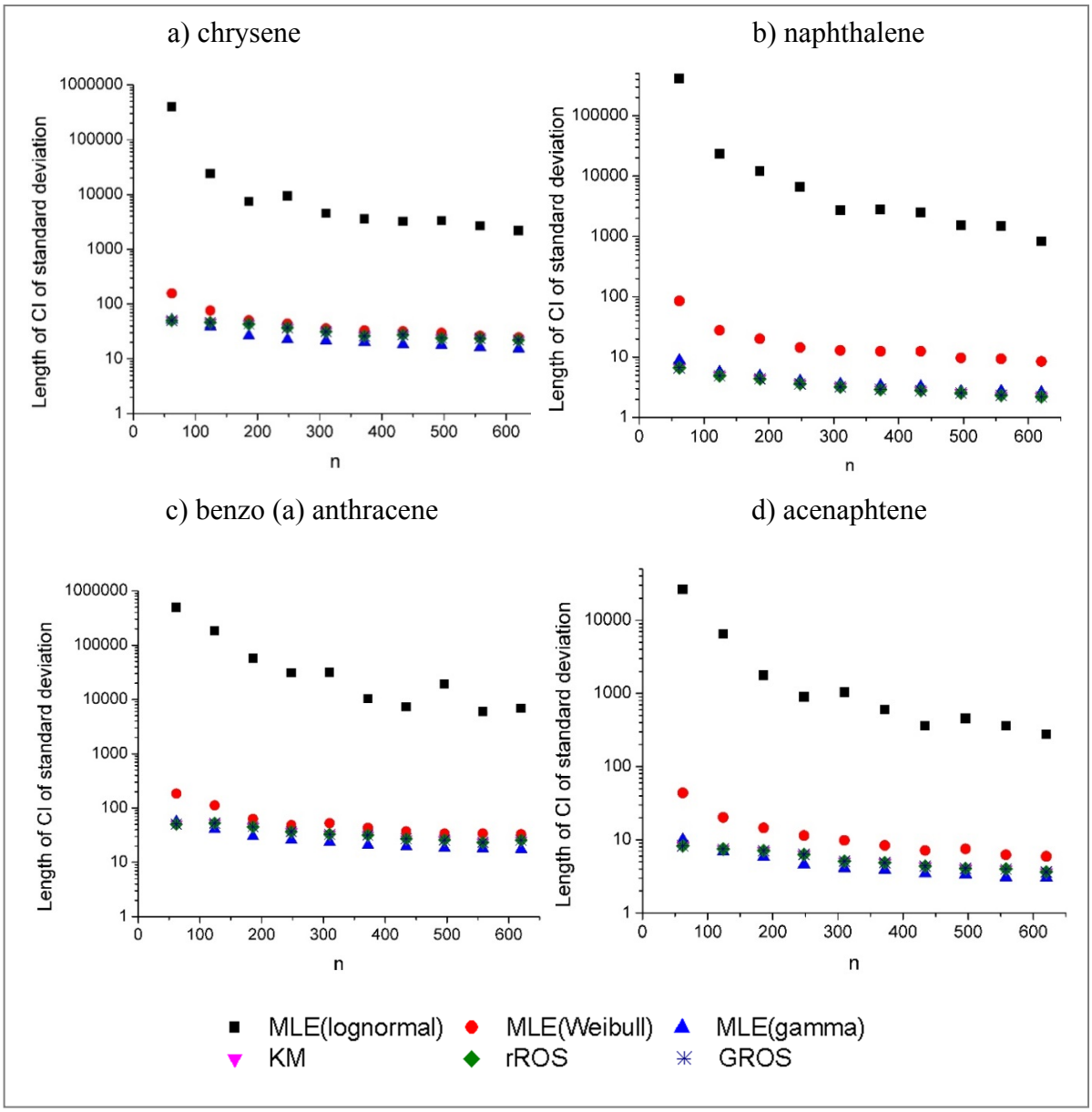


Figure 5.2 Bootstrap confidence interval lengths around the standard deviation estimate of a) chrysene, b) naphthalene, c) benzo(a)anthracene, and d) acenaphtene concentration data

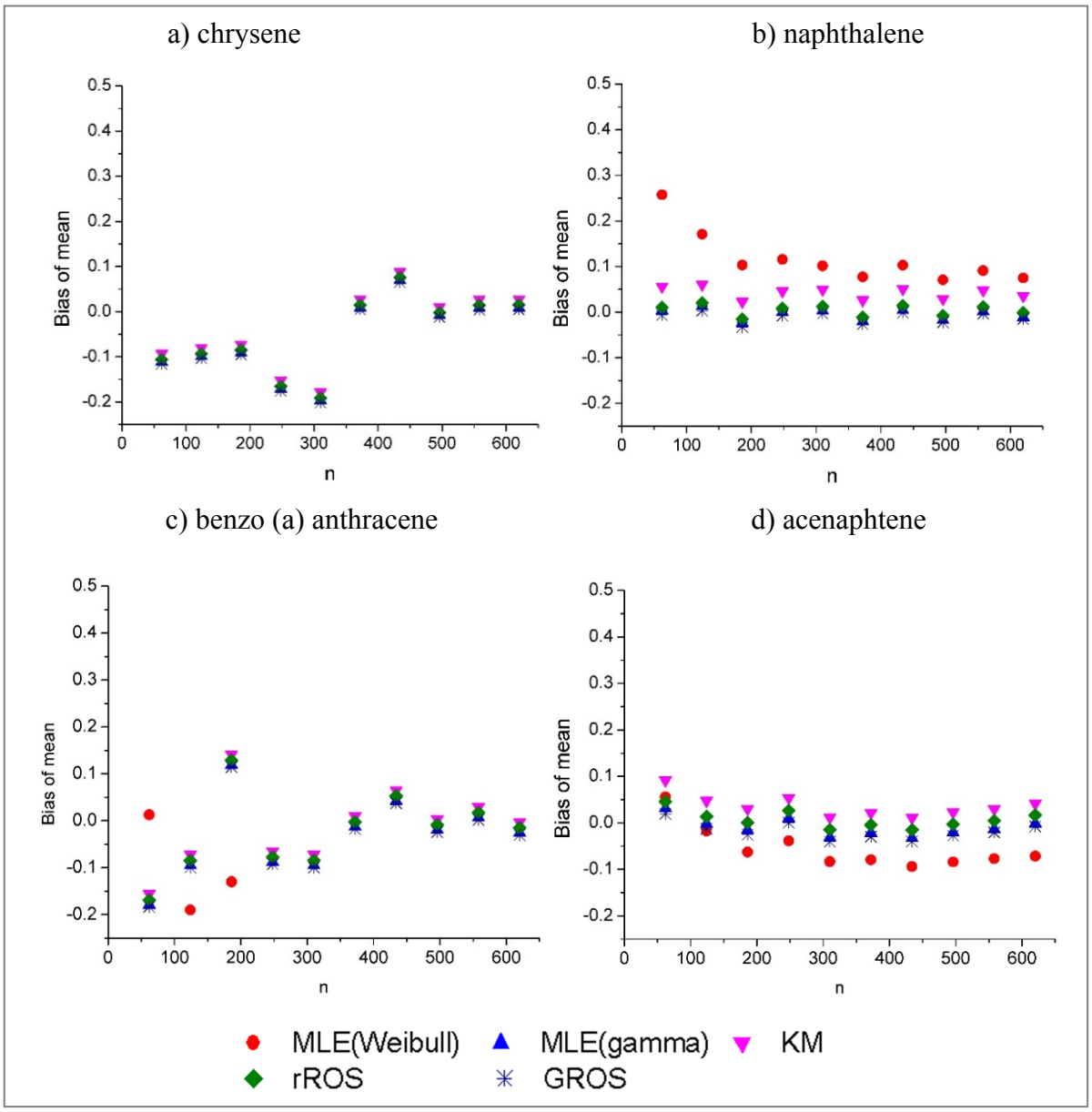


Figure 5.3 Approximated bias of the mean estimate of a) chrysene, b) naphthalene, c) benzo(a)anthracene, and d) acenaphtene concentration data

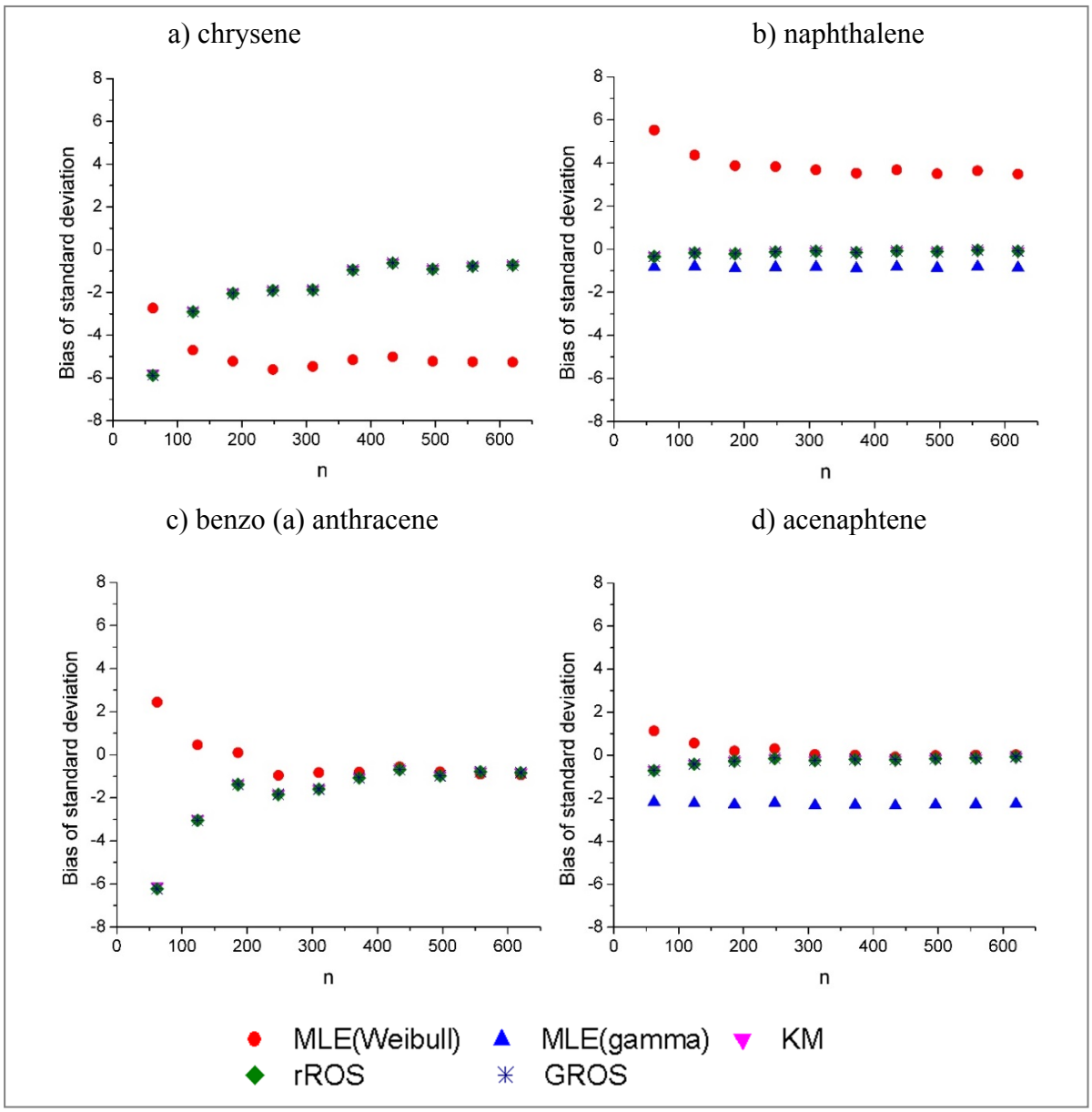


Figure 5.4 Approximated bias of the standard deviation estimate of a) chrysene, b) naphthalene, c) benzo(a)anthracene, and d) acenaphtene concentration data

5.6 Uncertainty estimation of the mean of concentration data

Risk-based decisions require some estimate of concentration that can be either the mean or 95th upper confidence level of the mean. In this section, we calculate the mean of the concentration data for the 23 contaminants under study using the MLE (under lognormal,

Weibull, and gamma distributional assumptions), rROS, GROS, and KM methods. We employ non-parametric bootstrapping followed by the percentile method to calculate the percentage uncertainty, $U\%$, around the mean estimates as

$$U\% = \frac{U}{\bar{x}} \tag{5.8}$$

where U is the absolute uncertainty expressed as half the 95% confidence interval, and $\bar{x}$ is the estimated mean. Table 5.4 displays the results of $\bar{x} \pm U\%$ when different estimators are used. It can be clearly observed that using the MLE method under lognormality assumption results in large estimates of the mean and percentage uncertainty. For example, for contaminant no.23 (2,3,5-trimethylnaphthalene), the mean and percentage uncertainty obtained by MLE under lognormality are 0.50 and 354%, respectively, which are remarkably larger than those estimated by other methods. The only exceptions are contaminants nos. 1 (As), 11 (Mo), and 22 (Cd), for which small amounts of uncertainty are achieved even when the lognormal MLE is used.

The reason for this exceptional observation may lie in the small values of coefficient of variation (CV), $CV = s/\bar{x}$, related to contaminants nos. 1, 11, and 22. Schmoyeri et al. (1996), Singh et al. (2006), and Shoari et al. (2015) noticed that when the *CV* is small, different statistical methods perform equally; however, when data distributions are characterized by large *CV* values, the estimators under the lognormality assumption can be misleading. We observe that the estimated *CV* for contaminant no. 1 is 0.5 and for contaminants nos. 11 and 22, it is 0.8; such small values of *CV* explain the reason for which comparable uncertainties are obtained. Note that we adopted the rROS method to estimate the *CV* values as Shoari et al. (2015) demonstrated the reasonable reliability of this method in estimating the mean and standard deviation. The *CV* for the remaining contaminants, where MLE (under lognormality) is distinguished because of the large uncertainty values, ranges between 1.51 and 3.86.

The other noticeable point is that the method of MLE assuming Weibull distribution results in large uncertainty values for some contaminants, making this estimator unreliable. The remaining estimators (i.e., MLE [under gamma assumption], rROS, GROS, and KM) provide comparable estimates of the mean and corresponding uncertainties.

Table 5.4 The estimate of the mean and its associated uncertainty (%) for contaminant data when different estimators are used[a]

	MLE (lognormal)	MLE (Weibull)	MLE (gamma)	KM	rROS (lognormal)	GROS (gamma)
1	9.77±23%	9.45±21%	9.42±21%	9.45±21%	9.43±21%	7.47±21%
2	0.45±128%	0.43±86%	0.43±81%	0.44±80%	0.43±80%	0.43±81%
3	40.32±141%	15.87±86%	18.39±85%	18.40±85%	18.39±85%	18.38±85%
4	71.92±167%	20.75±86%	22.54±86%	22.55±85%	22.54±85%	22.53±86%
5	60.25±170%	17.55±80%	18.35±77%	18.37±77%	18.36±77%	18.35±77%
6	28.56±195%	10.24±84%	11.16±91%	11.17±91%	11.16±91%	11.15±91%
7	38.20±147%	11.56±67%	11.00±60%	11.01±60%	11.00±60%	10.99±60%
8	24.75±176%	8.92±92%	9.73±87%	9.74±87%	9.73±87%	9.72±87%
9	36.70±182%	10.49±85%	10.86±83%	10.88±83%	10.87±83%	10.85±83%
10	39.31±60%	32.93±43%	32.89±41%	34.21±39%	33.08±41%	32.44±42%
11	2.78±15%	2.70±16%	2.74±15%	2.95±11%	2.73±16%	2.58±20%
12	19.02±244%	5.52±93%	5.20±82%	5.23±81%	5.22±81%	5.19±82%
13	21.50±208%	5.73±83%	5.26±78%	5.30±77%	5.28±77%	5.26±78%
14	17.33±201%	5.05±94%	4.95±86%	4.98±86%	4.97±86%	4.94±86%
15	2.65±158%	1.49±88%	1.63±87%	1.67±85%	1.64±86%	1.62±88%
16	0.80±100%	0.66±73%	0.70±69%	0.73±66%	0.70±68%	0.68±71%
17	0.98±124%	0.78±81%	0.86±75%	0.89±71%	0.86±75%	0.84±76%
18	3.27±200%	1.60±94%	1.68±84%	1.72±82%	1.69±83%	1.66±85%
19	1.20±121%	0.78±73%	0.77±68%	0.81±65%	0.77±67%	0.75±70%
20	1.38±176%	0.95±93%	1.02±76%	1.06±72%	1.02±75%	1.01±77%
21	4.02±272%	1.82±112%	1.75±78%	1.80±76%	1.76±78%	1.74±78%
22	0.94±25%	0.86±27%	0.86±27%	1.25±14%	0.93±29%	0.69±43%
23	0.50±354%	0.29±127%	0.25±82%	0.32±68%	0.26±77%	0.24±85%

[a] The numbers in column 1 refer to the contaminants listed in Table 5.1 The sample size and censoring percentage for each contaminant Table 5.1

5.7 Conclusions

Previous studies employed simulated left-censored data to investigate the performance of the MLE, rROS, GROS, and KM estimators. In the present study, we applied the non-parametric bootstrap technique to real concentration data and investigated the performance of the aforementioned estimators in terms of the uncertainty and approximated bias. Uncertainty was evaluated as the length of 95% confidence interval, and the approximated bias was calculated with respect to the rROS estimates. Among different estimators, bootstrap results indicated that the MLE method provided estimates with the highest amount of uncertainty, which is strongly impacted by the distributional assumption. For the present study, the MLE estimates obtained under the lognormality assumption were generally characterized by the highest uncertainties. Regarding the approximated bias, none of the estimators universally provided the least biased estimates.

We also investigated whether an increase in sample size could reduce the bias and uncertainty. The uncertainty of the estimates decreased as the sample size increased. Even so, the MLE estimates obtained under lognormality assumption were still characterized by the highest amounts of uncertainty. The bias provided by MLE relying on lognormal and Weibull assumptions occasionally improved as the sample size increased; however, the rROS, GROS, and KM estimators appeared to be asymptotically unbiased. Based on the bootstrap results, the present study concludes that the rROS, GROS, and KM methods provide estimates with small bias and uncertainty and thus are favored here. It is important to mention that the limitation of our adopted methodology is assumptions related to independently and identically distribute observations and representativeness.

CHAPTER 6

APPLICATION OF MIXED EFFECTS MODELS FOR CHARACTERIZING CONTAMINATED SITES

Niloofar Shoari[a], Jean-Sébastien Dubé[a]

[a] Department of Construction Engineering, École de technologie supérieure, Université du Québec, 1100, rue Notre-Dame Ouest, Montréal, Québec, Canada, H3C 1K3.

This article will be published in *Chemosphere* in January, 2017.

6.1 Abstract

In a typical data collection process for the purpose of characterizing contaminated sites, boreholes are usually drilled in different locations based on a sampling plan; and consequently, multiple samples are collected from each borehole. As a result, it is quite plausible that a certain degree of dependency or similarity exists among observations nested within a borehole. However, when classical regression models are employed, such dependencies are often ignored, resulting in biased estimates. In site characterization studies, further complication arises due to the presence of left-censored observations, those falling below the detection limit of measuring instruments. To overcome the above issues, this paper employs a mixed effects model that allows accounting for the within-borehole data dependency while accommodating left-censored concentrations. The benefits of the adopted methodology are explored by analyzing concentration data obtained from characterization study of a brownfield site located in Montreal, Canada. This paper illustrates that the estimated within-borehole correlation can be used to determine the optimal number of boreholes as well as the sample size to be collected from each borehole. Such correlation is underestimated when censored values are not accommodated in the model but substituted with a constant prior to data analysis. In addition, the adopted methodology provides an accurate insight into the vertical extent of contamination that can result in different compliance decisions when compared with classical approach.

6.2 Introduction

Characterization of a contaminated site often involves collecting samples along boreholes drilled over the site and analyzing them for contaminant concentrations. The resultant concentration data have two main features. First, data have a nested structure; that is, multiple concentration observations are obtained from the same borehole. Secondly, concentration data frequently contain left-censored observations that are measurements falling below the detection limit (DL) of analytical instruments. Even with technical advances in chemical analysis protocols and laboratory instrumentations, there remains a threshold below which contaminants concentrations is not distinguished from the background noise.

With respect to the first feature, it is quite plausible to speculate that observations obtained from the same borehole are correlated since they may share similar known or unknown attributes. With this data structure, contaminants concentration varies both within and between boreholes, these variance components should be taken into account. In environmental studies, however, the common practice overlooks the potential correlation between observations within a borehole, assuming that samples are collected from a single homogeneous population. Analysis of data sets characterized by a nested structure is best performed using mixed effects models, known also as multilevel models (Gbaguidi-Haore, Roussel, Reboux, Dalphinand & Piarroux, 2009; Hox, 2010). The peculiarity about these models is that they include both fixed effects and random effects. Modeling with the fixed effects allows examining the average relationship between a dependent variable (concentrations in this study) and predictor(s). This is equivalent to fitting a classical linear regression model. Inclusion of random effects, on the other hand, acknowledges the presence of some unobserved characteristics associated with each borehole and provides an estimation of the “between-borehole” contamination variability. Failure to recognize the nested structure of concentration data could result in misinterpretation in the analysis of data. In fact, fitting classical regression models to the data with dependent observations may result in an underestimation of standard errors and consequently, misleading conclusions about the

significance of the parameters whose effects are investigated (Pinheiro & Bates, 2006; Barr, Levy, Scheepers & Tily, 2013).

With respect to the second feature, left-censored observations are commonly substituted with constants (e.g., DL/2) so that complete data sets are "fabricated" (Helsel, 2006) prior to any analysis. A number of studies that accommodated censored data in the regression models include Slymen, de Peyster & Donohoe (1994), Liu et al. (1997), Gardner & Vogel (2005), and Dien, Hirai, Miyazaki & Sakai (2016); however, the random effect parameter was not incorporated in these studies. Some studies in the fields of biostatistics, epidemiology, ecology, and transportation have fit mixed effects models to data without censoring (e.g., Guo, 2005; Jordan, Schimleck, Clark, Hall & Daniels, 2007; Bogner et al., 2010; Warne et al., 2012; Heydari, Miranda-Moreno & Fu, 2014; Lee & Koutrakis, 2014; Giri, Nejadhashemi, Zhang & Woznicki, 2015; Wu et al., 2015; Chen, Qin, Zeng & Li, 2016) and less frequently, to data subject to censoring (e.g., Thiébaut & Jacqmin-Gadda, 2004; Jin et al., 2011; Vaida & Liu, 2012; Bakke, Ulvestad, Thomassen, Woldbæk & Ellingsen, 2014). To our knowledge, the use of mixed effects models in site characterization studies involving left censored concentrations is rare if nonexistent.

The general objective is to advance the use of mixed effects models for environmental data with left-censoring and to highlight how these models help developing better management practices for site characterization and remediation. The benefits of the adopted methodology are studied by fitting mixed effects models to censored concentration data of soil samples collected for the purpose of characterizing a brownfield site in Montreal (Canada). While accommodating censored values, mixed effects models estimate different variance components of contamination, i.e., within- and between-borehole variances. In addition, the relationship between contamination level and depth and the type of materials from which samples were collected is explored. Finally, practical implications for compliance with environmental standards and sample size determination are provided.

6.3 Site and data description

The site under study is a brownfield site that covers an area of 830,000 m^2 in Montreal, Canada. This site mainly consists of layers of backfill and waste material, crushed stone and natural soil. Several site characterization studies were conducted on this brownfield site between 1998 and 2009. Soil samples were collected at different depths from 242 boreholes dispersed over the site. Generally, one to four soil samples were analyzed for contaminant concentrations (14 inorganic compounds and 23 polycyclic aromatic hydrocarbons (PAH)). Concentration data contained some observations below the detection limits resulting in left-censored data. For a more practical insight into the vertical profile of the contamination, four depth categories on the basis of the dominant material were defined as:

- Depth I (from 0m to 1m) consists of crushed stones, backfill material with some portions of waste;
- Depth II (from 1m to 2m) consists of backfill and waste material;
- Depth III (from 2m to 3m) consists of waste;
- Depth IV (>3m) consists of natural soil.

6.4 Methodology

Employing simple linear regression and linear mixed effects models, this paper examined the extent of contamination at different depths and in different materials from which samples were collected. With the assumption that all observations including those coming from the same borehole are independent, simple linear regression model that includes only the fixed effects is fitted by

$$y_{ij} = \beta_j + \varepsilon_{ij} \quad i = 1, \dots, M, j = 1, \dots, N \tag{6.1}$$

where y_{ij} represents concentration measurement for borehole i in material (or at depth) j; the β_j represents fixed effects or the mean concentration in material (or at depth) j, and the ε_{ij} is

an independent error term that is normally distributed with mean zero and variance σ^2, $N(0, \sigma^2)$. Also, *M* is the number of boreholes and *N* is the number of materials- or depth categories- defined earlier.

The nested structure of characterization data violates the assumption of independence that underlies the simple linear regression technique. Under such circumstances, this paper adopts the methodology used by Vaida & Liu (2012) for left-censored HIV-1 viral load data in which a linear mixed effects model is defined as

$$y_{ij} = \beta_j + b_i + \varepsilon_{ij} \quad i = 1, \dots, M, j = 1, \dots, N \tag{6.2}$$
$$b_i \sim N(0, \sigma_b^2), \varepsilon_{ij} \sim N(0, \sigma^2)$$

As in equation (6.1), y_{ij} represents concentration measurements, β_j represents the fixed effects and ε_{ij} is the random error. The parameter b_i stands for random effects (here, borehole effects) with a mean of zero and variance of σ_b^2. Considering the case where some concentration measurements are left-censored, the observed concentrations y_{ij} are presented as pairs of (q_{ij}, δ_{ij}), where q_{ij} represents the observed value and δ_{ij} is the censoring indicator such that

$$y_{ij} = \begin{cases} q_{ij} & If\ \delta_{ij} = 0 \\ < DL & If\ \delta_{ij} = 1 \end{cases} \tag{6.3}$$

The models used in this study were built using the package "lmec" (Vaida & Liu, 2009) in R that estimates the following quantities through the maximum likelihood or restricted maximum likelihood method.

- Fixed effects parameter β_j, which represents the mean contaminant concentration, for each depth category or material type, and its corresponding variance;
- Random effects parameter or borehole effect b_i, which represents the deviation of the mean contaminant concentration at each borehole from the mean contaminant

concentration of the population under study; the variance for b_i, σ_b^2, provides an estimate of the "between-borehole" variability;

- Random error ε_{ij} whose variance represents the "within-borehole" variability, σ^2.

For details regarding the theory and computational methods of mixed effects models, interested readers are referred to Pinheiro & Bates (2006), Wu (2009), and West, Welch & Galecki (2014). To improve the normality of error terms, concentration observations are log-transformed as in Bogner et al. (2010); Janssen (2012), and Vaida & Liu (2012), among others. Therefore, the fixed effects represent the mean concentration of a given contaminant in each material or at each depth in log-scale, i.e., the geometric mean (GM). Due to difficulties in interpreting the GM estimates, these were back-transformed into original scale (details are provided in section 6.6). To choose between models with and without random effects (i.e., simple linear regression versus linear mixed-effects regression), the Akaike Information Criteria (AIC) was used (Burnham & Anderson, 2003). The model with the lowest AIC value provides a superior fit.

In addition to the characterization of the vertical distribution of contamination (in terms of the depth category and material), the mixed effects models provide insight regarding the data dependency through the intra-borehole correlation coefficient (IBC). The IBC is a measure describing the similarity of concentration observations nested in the same borehole and can be computed from

$$IBC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} \tag{6.4}$$

Values of IBC close to 0 indicate that observations are perfectly independent of each other and a simple regression analysis is sufficient. As IBC approaches 1, within-borehole dependency increases. To evaluate the role of censored values, mixed effects models were fit to data after left-censored concentrations were substituted with DL/2. This paper discusses

some examples that show how substitution of censored values provides biased IBC values, which would potentially alter the outcomes of site characterization.

6.5 Results

Table 6.1 and Table 6.2 report the results of the application of both simple linear regression and mixed effects models to cadmium, copper, lead, benzo(a)pyrene, and naphthalene concentration data that were identified as contamination indicators in a previous study conducted on this site by Dessau (2009). These tables display the estimates of variance components associated with random error and random effects error, which are indicated by the within-borehole (σ^2) and between-borehole variance (σ_b^2), respectively. Mixed effects model produces a much smaller estimate of the random error variance than the simple linear model because some variations are captured by the between-borehole variance. The estimates of IBC indicate that some level of correlation exists between the concentrations measured within a borehole. For a more comprehensive illustration of the range of IBC in our study, Figure 6.1 shows IBC for 14 inorganic and 23 PAH contaminants when material type is considered as fixed effects. While the IBC values in 79% of the contaminants are larger than 0.3, we observe substantially large IBC values for Cd, Hg, Se, and 7,12-dimethylbenz(a)anthracene. Due to data dependencies, the simple linear model should be abandoned as its underlying assumption (observations independency) is likely to be violated (Heck & Thomas, 2015).

Simultaneously, mixed effects models assess the relationship between contaminants concentration and material type or depth from which soil samples were collected. Table 6.1 and Table 6.2 show that the regression parameters (estimates of GM) are significant at α =5%. The only exception occurs in cadmium concentration estimated in waste material (Table 6.1). The analyses indicate that both mixed effects and simple regression models provide comparable estimates of GM. However, model comparison between the mixed effects and simple linear regression models show that the mixed effects models provide an improved model fit (smaller AIC values) over simple linear regression. It's worth mentioning

that mixed effects models have the advantage of accommodating missing data. In fact, in the data sets used in this study, while some materials, or depth categories, contain duplicate measurements, some others have no concentration measurements (missing data). If the model is correctly specified, the missing values do not bias the inference regarding the β_j estimates. This is not the case if the nested structure of characterization data is ignored, and inference related to each material, or depth category, is based only on available concentration measurements.

Table 6.1 Linear regression versus mixed effects models when the material type is considered as fixed effects[a]

	Mixed effects model	**Simple linear regression**
Copper n=428 ,Censoring=18%		
Within borehole variance σ^2	0.97	1.18
Between borehole variance σ_b^2	0.21	-
IBC	0.18	-
GM (waste)	5.03 [4.71,5.34]	5.08 [4.77,5.39]
GM (crushed stones)	2.95 [2.58,3.33]	2.92 [2.55,3.30]
GM (backfill)	3.95 [3.81,4.08]	3.93 [3.80,4.07]
GM (natural soil)	3.29 [3.03,3.54]	3.33 [3.08,3.58]
AIC	1313	1318
Lead n=434, Censoring=30%		
Within borehole variance σ^2	2.18	3.09
Between borehole variance σ_b^2	0.93	-
IBC	0.30	-
GM (waste)	5.24 [4.73,5.75]	5.44 [4.95,5.94]
GM (crushed stones)	2.25 [1.63,2.88]	2.12 [1.48,2.76]
GM (backfill)	3.36 [3.13,3.60]	3.35 [3.13,3.57]
GM (natural soil)	2.03[1.59,2.46]	1.92 [1.50,2.35]
AIC	1712	1732

(Continued)

	Mixed effects model	Simple linear regression
Cadmium n=423, Censoring=67%		
Within borehole variance σ^2	0.59	2.00
Between borehole variance σ_b^2	1.90	-
IBC	0.76	-
GM (waste)	0.03 [-0.42, 0.48]	0.99 [0.51, 1.48]
GM (crushed stones)	-1.62 [-2.28, -0.96]	-1.69 [-2.42, -0.96]
GM (backfill)	-0.95 [-1.21,-0.70]	-0.85 [-1.05, -0.65]
GM (natural soil)	-1.58 [-2.01,-1.15]	-1.87 [-2.36, -1.38]
AIC	1407	1504
Benzo(a)pyrene n=517,Censoring=51%		
Within borehole variance σ^2	2.81	4.37
Between borehole variance σ_b^2	1.63	-
IBC	0.37	-
GM (waste)	-1.47 [-2.01,-0.92]	-1.60 [-2.15,-1.05]
GM (crushed stones)	-4.35 [-5.44,-3.26]	-4.42 [-5.60,-3.24]
GM (backfill)	-1.85 [-2.14,-1.56]	-1.86 [-2.12,-1.59]
GM (natural soil)	-3.52 [-3.97,-3.06]	-3.43 [-3.84,-3.02]
AIC	2203	2240
Naphthalene n=516, Censoring= 57%		
Within borehole variance σ^2	2.95	4.57
Between borehole variance σ_b^2	1.68	-
IBC	0.36	-
GM (waste)	-1.72 [-2.28, -1.16]	-1.74 [-2.31, -1.18]
GM (crushed stones)	-4.04 [-5.15,-2.93]	-4.29 [-5.46,-3.11]
GM (backfill)	-2.44 [-2.74,-2.13]	-2.40 [-2.68,-2.12]
GM (natural soil)	-3.77[-4.24,-3.30]	-3.57 [-4.01,-3.14]
AIC	2221	2258

Note: GM=geometric mean, IBC=intra-borehole correlation.
[a] values in parenthesis refer to 95% upper and lower confidence levels

Table 6.2 Linear versus mixed effects models when the depth category is considered as fixed effects[a]

	Mixed effects model	**Simple linear regression**
Copper n=428 ,Censoring=18%		
Within borehole variance σ^2	1.15	1.46
Between borehole variance σ_b^2	0.31	-
IBC	0.21	-
GM(Depth I)	3.66 [3.41, 3.92]	3.69 [3.43,3.94]
GM(Depth II)	3.96 [3.74, 4.17]	3.93 [3.72,4.15]
GM(Depth III)	4.07 [3.84, 4.30]	4.08 [3.85,4.31]
GM(Depth IV)	3.73 [3.49, 3.96]	3.73 [3.51,3.96]
AIC	1402	1411
Lead n=434, Censoring=30%		
Within borehole variance σ^2	2.58	4.20
Between borehole variance σ_b^2	1.59	-
IBC	0.38	-
GM(Depth I)	3.28 [2.86,3.71]	3.28 [2.84,3.72]
GM(Depth II)	3.39 [3.01,3.76]	3.32 [2.94,3.70]
GM(Depth III)	3.44 [3.05,3.83]	3.50 [3.11,3.90]
GM(Depth IV)	2.74 [2.33,3.15]	2.66 [2.25,3.07]
AIC	1824	1864
Cadmium n=423, Censoring=67%		
Within borehole variance σ^2	0.66	2.51
Between borehole variance σ_b^2	2.38	-
IBC	0.78	-
GM(Depth I)	-1.18 [-1.57, -0.79]	-1.14 [-1.56, -0.72]
GM(Depth II)	-1.01 [-1.35, -0.68]	-0.80 [-1.13, -0.46]
GM(Depth III)	-0.90 [-1.24, -0.55]	-0.56 [-0.90, -0.21]
GM(Depth IV)	-1.45 [-1.84, -1.06]	-1.54 [-1.94, -1.13]
AIC	1475	1600

(Continued)

	Mixed effects model	Simple linear regression
Benzo(a)pyrene n=517,Censoring=51%		
Within borehole variance σ^2	3.21	4.92
Between borehole variance σ_b^2	1.75	-
IBC	0.35	-
GM(Depth I)	-2.84 [-3.39, -2.29]	-2.71 [-3.28,-2.15]
GM(Depth II)	-2.10 [-2.54, -1.66]	-2.20 [-2.64,-1.76]
GM(Depth III)	-2.24 [-2.67, -1.81]	-2.33 [-2.76,-1.90]
GM(Depth IV)	-2.34 [-2.72, -1.97]	-2.40 [-2.74,-2.06]
AIC	2264	2297
Naphthalene n=516, Censoring= 57%		
Within borehole variance σ^2	3.25	4.86
Between borehole variance σ_b^2	1.66	-
IBC	0.34	-
GM(Depth I)	-3.13 [-3.71, -2.54]	-3.18 [-3.77,-2.59]
GM(Depth II)	-2.46 [-2.91, -2.01]	-2.52 [-2.96,-2.07]
GM(Depth III)	-2.65 [-3.09, -2.21]	-2.65 [-3.09,-2.22]
GM(Depth IV)	-2.85 [-3.24, -2.47]	-2.72 [-3.06,-2.37]
AIC	2258	2286

Note: GM=geometric mean, IBC=intra-borehole correlation.

[a] values in parenthesis refer to 95% upper and lower confidence levels

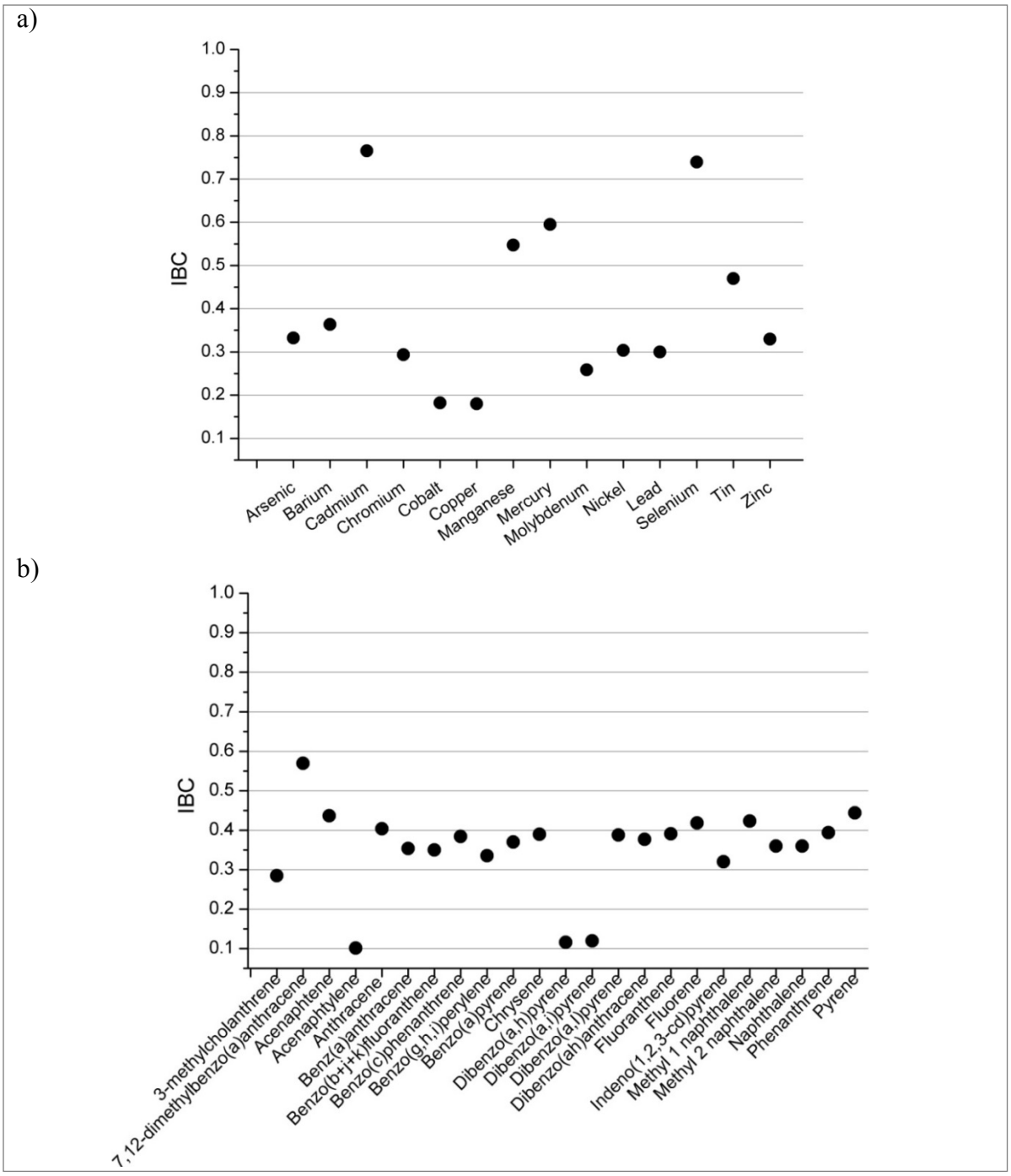


Figure 6.1 Intra-borehole correlation for a) inorganic compounds and b) PAH contaminants

The following discusses the results when mixed effects models were fitted to complete concentration data after left-censored observations were substituted with DL/2 (results are

reported in Table-A VI-1 of Appendix VI). Substitution of censored data results in underestimation of the fixed effects (or GM here) in cadmium, benzo(a)pyrene and naphthalene contaminant data for which the censoring percentage is relatively large, 67%, 51% and 57% respectively. In addition, smaller estimates of the within-borehole variance and more noticeably of the between borehole variance are obtained for these contaminants. For example, the σ^2 and σ_b^2 for naphthalene (reported in Table 6.1) are 2.95 and 1.68, respectively, whereas those after the substitution (reported in Appendix VI) are reduced to 1.45 and 0.50, respectively. Consequently, the estimate of IBC is reduced from 0.36 (in the case of accounting for censored data) to 0.26 (in the case of substituting censored data). As a result, substitution of censored data masks the true correlation between observations in the same boreholes as IBC values are generally underestimated. In the case of smaller censoring percentages, as for Cu and Pb, which have respectively 10% and 30% censoring, the impact of substitution is smaller.

6.6 Implications for site characterization

This section explores practical implications of mixed effects models in terms of compliance with environmental standards and sample size determination for site characterization studies.

6.6.1 Compliance with a soil regulatory standard

The 95% upper confidence level of the mean (95UCL) for contaminant data was calculated and then compared to the soil regulatory criteria reported in Schedule I of Land Protection and Rehabilitation Regulation (LPRR) published by "Ministère du Développement durable, de l'Environnement, de la Faune et des Parcs du Québec" (2003b). Although the estimates of fixed effects sufficiently explain the relationships between the material type– or depth category- and contamination level, estimates in log-scale are not very informative in environmental sciences. Therefore, the following equation was used to back-transform 95UCL of fixed effects estimates into the original scale, though some bias is inevitable (Gurka, Edwards, Muller & Kupper, 2006).

$$y_{ij}^{o} = \exp(UCL_{log} + 0.5(\sigma^2 + \sigma_b^2)) \quad (6.5)$$

where y_{ij}^{o} is the concentration on the original scale, UCL_{log} is the 95UCL of fixed effects estimates in log-scale, and other notations have been previously described. To highlight the benefits of employing mixed effects while accounting for left-censored data, the 95UCL was computed under the following modeling scenarios:

Scenario 1- Censored observations are substituted with DL/2. The fixed effects parameter and nested structure of concentration data are discarded in the model;
Scenario 2- Censored observations are substituted with DL/2. The fixed effects parameter of depth is introduced into the model, while nested structure of concentration data is ignored;
Scenario 3- Censored observations are substituted with DL/2. The mixed effects model is fitted to completed data fabricated from substituting left-censored observations (i.e., the random effect of borehole is included in the model described in scenario 2);
Scenario 4- Mixed effects model is fitted to data containing left-censored observations.

In the first scenario, where concentration observations are aggregated together regardless of the depth, material and the borehole, the usual formulas to calculate a global mean and its 95UCL of the complete data set are used. Figure 6.2a and Figure 6.2b illustrate examples of concentration variability among different material types and depth categories for Pb (n=434 observation and 30% censoring percent). As follows, contamination levels differ between materials and depth categories and thus could be represented by separate estimates of the 95UCL. This can be modeled by scenario 2 and the results of which are reported in the fourth column of Table 6.3. Even though the updated model gives information about the vertical distribution of contamination, it still does not account for variations between boreholes. In fact, boxplots of concentration data of Pb for boreholes (Figure 6.2c) indicate the importance of accounting for the boreholes effects as large between-borehole concentration variability is observed. Figure 6.2c underlines the need for including random effects (borehole effect) into the model (scenario 3). The 95UCL provided by fitting the mixed effects model to complete concentration data are reported in the fifth column of Table 6.3. In the presence of left-

censored observations, mixed effects models are particularly attractive because it is conceptually straightforward to incorporate censored measurements in likelihood inference for mixed effects models; results of fitting mixed effects models to scenario 4 data are provided in the sixth column of Table 6.3.

Including fixed effects of depth (as modeled in scenarios 2, 3, and 4) reveals that the vertical soil profile differs in 95UCL concentration, with depth categories II and III having the highest contamination levels. For example, for Cu, depth categories II and III exceed the regulatory criterion reported in Schedule I of the LPRR (i.e., 100 mg Cu/kg) and thus should be targeted for remedial actions. Of interest is that these depth categories (i.e., from 1 to 3m) are characterized by larger amounts of waste material. However, if only fixed effects were considered (scenario 2), all depths would have been categorized as contaminated and targeted for remedial actions. Including borehole effects improves the quality of the 95UCL estimates as small AIC values (reported in section 6.5) indicate that mixed effects models provide a better fit to contaminant data.

With respect to the role of left-censored data in the model, Table 6.3 shows that substitution of left-censored observations may also lead to incorrect compliance decisions. For example, for benzo(a)pyrene, considering left-censored observations in mixed effects models indicates non-compliance with regulatory criterion listed in LPRR (1 mg B(a)P/kg) in depth categories II, III and IV. However, when censored observations are substituted with a constant, the same conclusion cannot be reached as the 95UCL values are underestimated and are all smaller than the criterion.

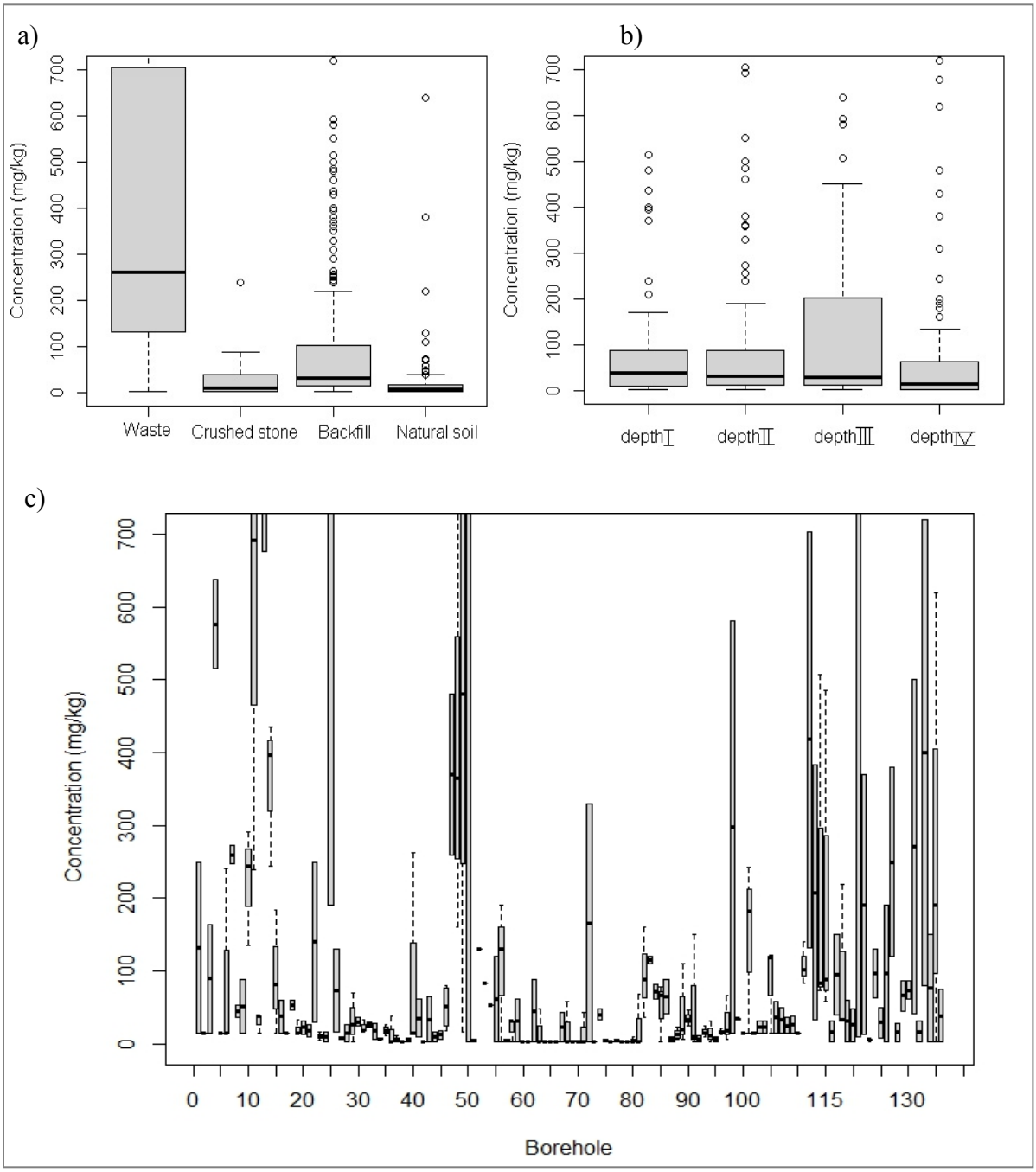


Figure 6.2 a) Boxplots of Pb concentrations for different materials and b) depth categories; c) boxplots of Pb concentrations for different boreholes

Table 6.3 Comparison of the 95UCL of the mean concentration (mg/kg) at each depth category using conventional, simple linear and mixed effects models

Contaminant	Depth category	Conventional methods		Mixed-effect model	
		95UCL of the global mean	Simple linear regression	Completed data after substitution	Censored data
		Scenario 1	Scenario 2	Scenario 3	Scenario 4
Copper	DepthI	111.09	102.63	79.09	80.91
	DepthII		127.41	105.74	108.42
	DepthIII		149.18	119.18	121.30
	DepthIV		105.06	84.30	86.29
Lead	DepthI	185.65	211.03	150.08	214.68
	DepthII		213.41	166.13	237.66
	DepthIII		267.03	177.82	251.37
	DepthIV		125.74	98.69	124.93
Cadmium	DepthI	1.25	1.29	1.39	2.08
	DepthII		1.41	1.43	2.30
	DepthIII		1.66	1.56	2.62
	DepthIV		1.05	1.20	1.57
Benzo(a)pyrene	DepthI	0.64	0.59	0.54	0.70
	DepthII		0.79	0.86	1.47
	DepthIII		0.70	0.75	1.27
	DepthIV		0.68	0.74	1.15
Naphthalene	DepthI	0.43	0.39	0.30	0.51
	DepthII		0.52	0.41	1.00
	DepthIII		0.48	0.37	0.83
	DepthIV		0.47	0.37	0.67

6.6.2 Sample size determination

Another important implication of mixed effects models is the possibility of determining optimal sample sizes at each level of the nesting hierarchy. In other words, mixed effects models provide researchers tools to decide how many boreholes and how many samples from each borehole are adequate. The optimal sample size is that required to minimize the standard error of the estimated IBC. As reported by Donner (1986), the standard error of the estimated IBC is calculated by

$$SE\ (IBC) = (1 - IBC)(1 + (N - 1)IBC)\sqrt{\frac{2}{N(N-1)(M-1)}} \qquad (6.6)$$

where M is the number of clusters (i.e., boreholes) and N is the number of individuals in a cluster (i.e., concentration observations obtained from a borehole). It is important to note that the formula discussed in this section is adopted from studies in the fields of epidemiology and social sciences (e.g., Scherbaum and Ferreter, 2009; van Breukelen and Candel, 2012) and adapted to the context of site characterization. Once an educated guess about the IBC can be made, from preliminary data for instance, the SE(IBC) can be plotted as a function of N and M values. Such a plot is illustrated in Figure 6.3. The results indicate that there is a relatively high level of dependence between observations within a borehole and assume that IBC is 0.39 (this is the average estimate of IBC for inorganic and PAHs). As can be seen in Figure 6.3, the estimated standard error decreases rapidly up to N=6 observations per borehole, after which negligible reduction in standard error is obtained (approximately 5%). Moreover, increasing the number of boreholes has a more substantial impact on decreasing the standard error than increasing the number of observations per borehole.

To illustrate the impact of IBC on sample size, Figure 6.4 depicts SE(IBC) as a function of IBC and the number of observations per borehole while the number of boreholes is fixed (M=150). This figure clearly shows that if the correlation between observations in a borehole increases (i.e., larger IBC), each concentration observation provides little unique information.

Therefore, taking several concentration measurements from a given borehole becomes more redundant as IBC increases. As a result, drilling more boreholes would then be more informative than collecting and analyzing more soil samples from a given borehole.

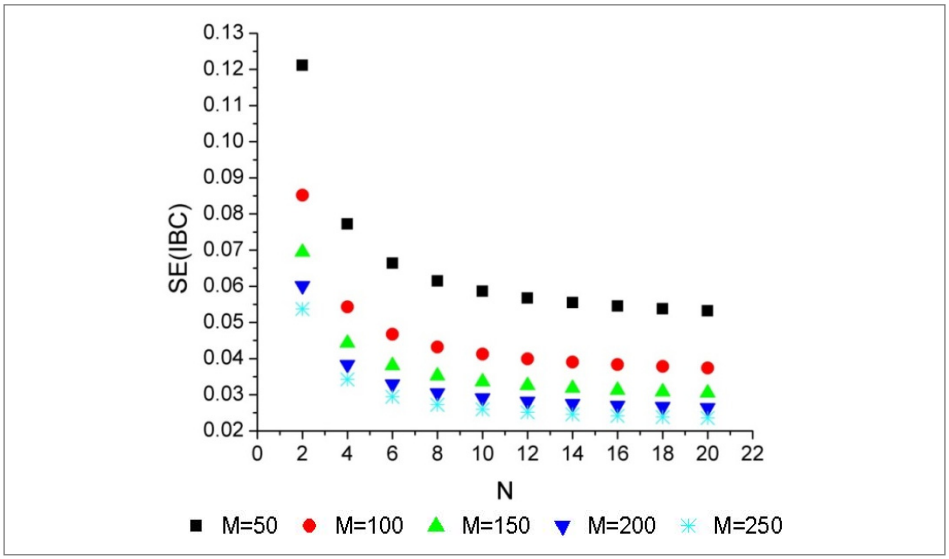


Figure 6.3 Standard error of IBC versus number of observations per borehole (N) for IBC=0.39 and different number of boreholes (M)

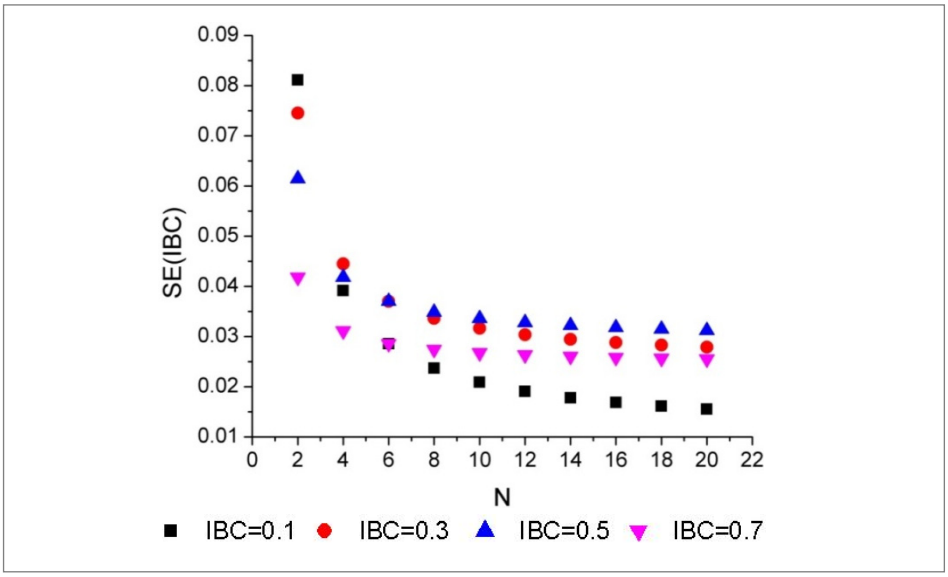


Figure 6.4 Standard error of IBC versus number of observations per borehole (N) for M=150 boreholes and different IBC

6.7 Conclusions

Using real data examples, this paper advances the application of mixed effects models in analyzing left-censored concentration data in the context of environmental site characterization. The adopted methodology was used to examine dependency in concentrations nested within boreholes and to estimate the between- and within-borehole variance. Management practices in terms of sample size determination have been ensued from the estimated variance components. In particular, this paper discussed how mixed effects models can help determining the optimal number of boreholes as well as concentration observations sampled from each borehole. In fact, when concentrations within the same borehole are highly correlated, taking and analyzing more samples from the same borehole is not as informative as drilling additional boreholes. Among other findings, the results showed that mixed effects models provided different vertical profiles of contamination as a function of depth and material type, compared to those obtained from classical models. It was also discussed how the substitution of censored observations can induce errors in the decision-making process regarding compliance with regulatory criteria. The analyses indicated that the substitution of left-censored concentrations with DL/2 would result in underestimated values of between-borehole variance. This impact was larger when the censoring percent was more than 50%.

CHAPTER 7 CONCLUSIONS AND RECOMMENDATIONS

This dissertation explored the quantitative impact of the below DL concentration (left-censored) data within the context of characterization of contaminated sites. This study sought to identify the statistical methods that can adequately analyze data with left-censored values. Concentration data resulting from characterization of two sites were used to illustrate how failing to account for left-censored values in analysis might affect the outcomes of a characterization study. The conclusions of this dissertation are organized according to the specific objectives of the study as defined in Section 1.1.

Estimation of descriptive statistics

The problem associated with the substitution of left-censored observations with arbitrary values, which is the most common way of handling censored concentration data, was first studied. Through an extensive simulation exercise we showed that the substitution approach results in biased estimates of descriptive statistics, which can potentially impact other statistical inference procedures that rely on these estimates (comparison of two or more soil populations, for example). Although substitution did not drastically affect the results in a number of simulation scenarios, these particular scenarios are hard to identify in real data as we do not have any knowledge about the underlying structure of data. For this reason, we do not recommend the substitution of censored observations even for data sets with small censoring percent.

Parametric and non-parametric alternative estimation techniques, which are based on survival analysis methods, should be preferred rather than the substitution approach. Our simulations showed that the performance of estimation techniques depends on various factors such as sample size, censoring percent, data skewness. More importantly, it depends on a combination of the aforementioned factors. A clear illustration of this finding was observed in the case of highly skewed data where the MLE method, with the assumption that concentration data are lognormally distributed, produced inflated estimates. The same technique however resulted in better estimates when the sample size increased. The

importance of this finding is that environmental studies often justify the choice of the MLE method based on the lognormality assumption by referring to previous studies such as Shumway et al. (2002), Hewett & Ganser (2007), and Helsel (2012) among many others, ignoring the fact that this technique may overestimate the descriptive statistics if data are highly skewed. As a result, a given soil may erroneously be identified as contaminated, while in reality it is not. The robustness of parametric methods (i.e., MLE, rROS, and GROS) against departure from the assumed distribution is another issue that has not been scrutinized in environmental studies. Simulation results demonstrated that the MLE method based on gamma assumption, rROS, and GROS provide estimates with the smallest MSE even when the underlying distribution of data does not match the assumed distribution. The non-parametric KM method also proved to be a reliable estimator when it was applied to censored data with less than 50% censoring.

We also employed the bootstrapping technique to concentration data from a characterization study and quantified the uncertainty associated with the mean and standard deviation values estimated by each of the alternative parametric and non-parametric estimators. The conclusions derived from bootstrapping of real data were consistent with those obtained from simulations. As a matter of fact, the MLE method under the lognormality assumption provided estimates with the highest uncertainty. In contrast, the MLE method based on gamma distribution, rROS, GROS, and KM generally produced estimates with small uncertainty.

Depending on the censored data percent and the data skewness, we suggest adopting one of the following approaches, as outlined in Figure 7.1.

- When there is less than 50% censoring and data exhibits low skewness, performance of different estimators becomes comparable. In the case of highly skewed data, the methods of KM, rROS, GROS, and MLE based on gamma assumption are recommended. Note that when data are censored at a single DL, the method of KM is not suggested as the mean estimate would be equal to that obtained after substitution with DL.

- When >50% censoring is present and data is low skewed, the MLE, rROS, and GROS methods provide good estimates of descriptive statistics. However, when data skewness is in doubt, the MLE method under lognormality is not recommended; instead, a gamma distributional assumption is preferred.

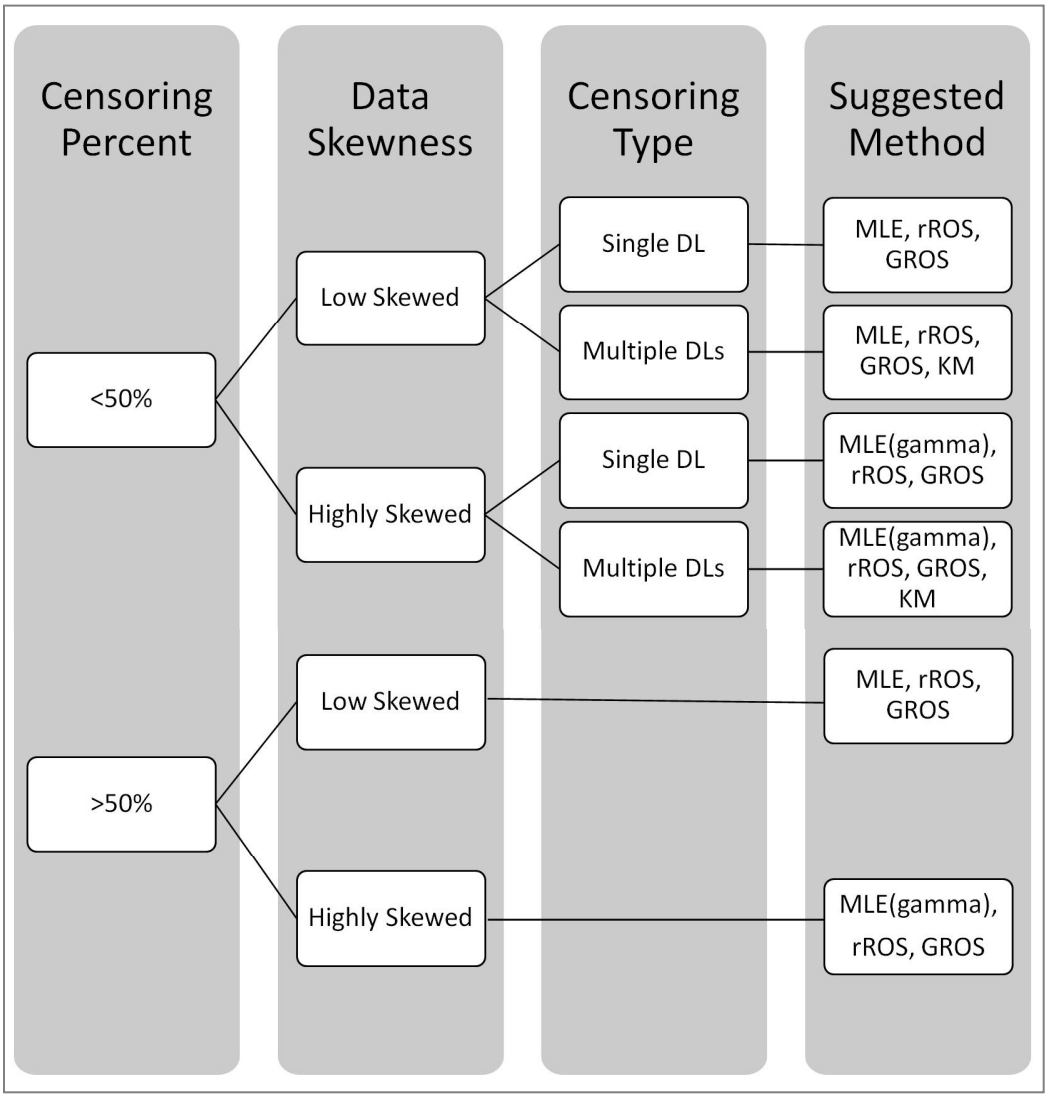


Figure 7.1 Recommended methods for estimating descriptive statistics of left-censored data

Accounting for dependency in left-censored concentration data

In a site characterization study, multiple soil samples are taken from a borehole. This sampling method might induce dependencies in concentration observations that are nested within the same borehole. Additional complications are missing data (some soil samples do not undergo chemical analysis, for example) and censored observations. Typically, censored values are substituted and data dependencies are ignored; these actions can result in misleading inferences. Accommodating censored values, we used mixed effects models to account for data dependencies and to estimate between-borehole contamination variability. The estimated variability served to determine the optimal number of boreholes as well as the number of soil samples collected from each borehole. To our knowledge, this is the first instance of employing mixed effects models for left-censored concentration data originating from contaminated sites characterization. This dissertation also illustrated the inadequacy ensued from the substitution of censored values as this approach erroneously underestimated the contamination variability. In addition, the adopted methodology provided a useful insight to the vertical extent of contamination. It was noted that the highest contamination levels were found in soil layers at 1-3 *m* of depth, which interestingly corresponded to the layers that contained the highest amount of waste material.

Overall, the substitution of censored observations is a flawed way of dealing with concentration data and thus should be avoided. Alternatively, the methods of MLE (assuming gamma distribution), rROS, GROS, and KM can provide more reliable estimates. Environmental policies should inform about the consequences of substitution and strongly encourage practitioners and researchers to employ alternative estimation techniques. Censored observations impact not only estimation of descriptive statistics, but also development of statistical models. In light of this study, we suggest that mixed effects models be considered as a statistical tool in characterization of contaminated sites due to their ability in quantifying the between-borehole contamination variability while accommodating censored data. The proposed methodology can be used to improve current sampling

strategies that only estimate the overall number of samples without specifying the number of required boreholes.

Recommendations for future work

The research presented in this thesis has raised several lines of research which should be pursued.

- The literature on the distribution of environmental data claims that concentration data can be modeled by lognormal distribution. Environmental publications that rely on lognormality often justify their assumption by (i) referring to a study conducted by Ott (1990), who demonstrated that concentration data were well approximated by a lognormal distribution, assuming that data were the result of many independent random dilution, or (ii) citing previous papers that employed the lognormality assumption (the previous papers generally rely on still-earlier papers without actually using the data to support their assumption). This is probably due to the complexity of implementation of distributional checking procedures that are tailored for left-censored data. In view of the fact that studies based on the lognormality assumption, including this research, did not necessarily report consistent results, identifying the actual distribution of concentration data in the presence of nondetects becomes quite relevant. We propose to employ a category of goodness-of-fit tests that compare the empirical cumulative distribution function (obtained in a non-parametric way) with its parametric counterpart. Since the shape of the distribution of environmental data is right-skewed, lognormal, Weibull, and gamma distributions are plausible candidates to model contaminants concentration data sets. Some simulations toward verifying the proposed goodness-of-fit test have been already done and are reported in Appendix II;
- If field results obtained from phase II and III of site characterization meet or exceed generic remediation guidelines, a remediation strategy and/or a risk assessment strategy is required. Developing the remediation objectives can be performed through a guideline approach that can be adopted from published environmental guidelines. Alternatively, we can employ a risk-based assessment that includes developing site-specific remediation

objectives based on human health and/or ecological risk assessment. While this research mainly focused on exploring the impact of censored data within the guideline approach concept, it would be interesting to see how censored concentration data affect the outcomes of a risk-based site characterization;

- Mixed effects models discussed in chapter 6 were employed to analyze concentration data of a contaminant. However, a site characterization study usually involves measuring concentration of more than one contaminant at each sampling location, and sometimes these contaminants are also correlated. Another future direction can focus on applying multivariate mixed effects models to censored concentration data in order to capture patterns of contamination across a site;
- It is quite interesting to extend the methodological framework to account for spatial dependencies between concentrations while accommodating the censored ones. This should allow a better understanding of the extent of contamination at any location in the site.

ORIGINALITY OF WORK

This thesis contributes to the literature related to statistical analyses performed in site characterization through unification of the field and enhancing our understanding of comparative aspects of available methodological frameworks using both simulated and real data. In particular, first, we employed a large number of data scenarios including the percentage of censoring, skewness, and sample size in our simulation exercise. Such a comprehensive approach, which has been missing in literature, allowed us to conduct a more detailed and informative investigation of various methods. Doing so, we were able to address major contradictory findings of previous studies that ignored non-standard data conditions occurring frequently in real data sets. Second, we examined the performance of available methods using a bootstrapping technique based on real data. Third, we proposed and successfully applied a statistical method that not only accounts for left-censored concentrations in contaminated soil samples, but it also accommodates interdependency in data generated from sampling procedures. Neglecting the dependence structure in data, as an inherent feature of standard methods, results in biased estimates. To our knowledge, this is the first study to examine and address issues relating to the aforementioned tasks in the context of environmental site characterization studies.

APPENDIX I

AN OVERVIEW OF STATISTICAL METHODS FOR LEFT-CENSORED DATA

Suppose one wishes to estimate the mean and standard deviation of a sample of n concentration observations, $X = \{x_1, x_2, \dots, x_n\}$, from which k are left-censored at DL. This section reviews key estimation methods for analysis of such data; the focus is mainly on those techniques that have been already discussed in the "literature review" chapter. The notations are generally adopted from Singh et al. (2006) and Helsel (2012).

Substitution method

The simplistic but most commonly practiced approach to estimate statistical parameters (e.g., mean and standard deviation) of left-censored data entails substituting censored observations with arbitrary constants, which are typically a fraction of DL such as DL itself, DL/2, DL/√2. For the sake of simplicity, this approach is referred to as the substitution method although it is not a statistical technique. Replacement of censored observations has the practical advantage of forming complete data sets that allow using standard data analysis methods. On the other hand, the obvious disadvantage of the substitution method is that data sets do not reflect sampling variability because all censored values are replaced with the same constant.

Trimmed mean and Standard deviation

Trimming consists of discarding 100p% of data in both lower and upper tails. Note that p must be chosen such that reasonable amount of observations remain after np observations from both tails are cut out. This technique is valid only if the underlying distribution of data is symmetric.

Winsorized mean and standard deviation

Gilbert (1987) proposed the winsorized mean and standard deviation estimates in the case that data distribution is symmetric. The winsorization procedure follows three steps:

Step 1: After ordering the data, censored observations are substituted with the next smallest uncensored observations.
Step 2: The same number of the largest observations is substituted with the next smallest value. For example, if three censored values (on the left tail of distribution) are substituted in step 1, three largest observations (on the right tail) should be substituted with the next smallest uncensored observations.
Step 3: The usual estimation techniques are applied to the modified data set to estimate the mean and standard deviation.

Maximum likelihood estimator (MLE)

The general idea of MLE is to determine the parameter(s) of an assumed distribution that most likely resulted in the sample data. Let X be a sample of n concentration observations that are thought to come from a lognormal distribution, $ln(x_i;\ \mu, \sigma)$. When all observations are detected, the likelihood of observing the sample data is the product of the probability density function (pdf) for each observation and thus is given by

$$L = \prod_{i=1}^{n} \frac{1}{\sigma 2\pi} \exp\left[\frac{-(\ln(x_i) - \mu)^2}{2\sigma^2}\right] \qquad \text{(A I-1)}$$

The maximum likelihood estimates ($\hat{\mu}$ and $\hat{\sigma}$) are those that maximize the function L. Taking the natural log of L and setting the partial derivative with respect to μ and σ to zero, $\hat{\mu}$ and $\hat{\sigma}$ can be found.

In the presence of k left-censored observations, specifying the likelihood of the observed data and maximizing the likelihood function becomes complicated. The likelihood function consists of a part related to uncensored data and another part based on censored observations. For the uncensored part, the likelihood function is constructed using the pdf for each uncensored value. For the censored part, however, each censored observation contributes to the likelihood function with the cumulative distribution function (cdf) evaluated at the DL, because we merely know that the value is less than the DL. The likelihood function is thus written as

$$L = \prod_{i=1}^{n} \left\{\frac{1}{\sigma 2\pi} \exp\left[\frac{-(\ln(x_i) - \mu)^2}{2\sigma^2}\right]\right\}^{\delta_i} . \left\{\Phi(\frac{\ln(DL) - \mu}{\sigma})\right\}^{1-\delta_i} \qquad \text{(A I-2)}$$

where Φ is the cdf of the standard normal distribution and δ_i indicates whether the observation is censored or not (If $\delta_i = 1$, the observation is detected and if $\delta_i = 0$, the observation is left-censored). The logarithm of the likelihood function is given as follows:

$$\ln(L) = -(n-k)\ln(2\pi\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n-k}(\ln(x_i) - \mu) + k\ ln\ \Phi(\frac{\ln(DL) - \mu}{\sigma}) \qquad \text{(A I-3)}$$

The likelihood function is maximized using the iterative methods such as Newton-Raphson algorithm since analytical solution of this log-likelihood does not exist. The MLEs are asymptotically unbiased, have the minimum variance, and are asymptotically normally distributed. However, these properties are valid as long as sample size is large enough (the rule of thumb is >30) and the underlying distribution of data is correctly identified.

Expectation-Maximization (EM) algorithm

Another way of maximizing the likelihood function is the iterative approach using the EM algorithm (Dempster, Laird & Rubin, 1977). It is an iterative sequence of estimating the censored observations from the current parameter estimates and then estimating the parameters from the actual and estimated observations. In the E-step, the conditional expectation of the complete data log-likelihood is computed. In the M-step, the parameters that maximize the complete data log-likelihood are estimated. The E-step and the M-step are alternately repeated until convergence is met. Following the notations in Singh & Nocerino (2002), let $(\hat{\mu}_j, \hat{\sigma}_j{}^2)$ be the estimates of (μ, σ^2) at j[th] iteration, then $(\hat{\mu}_{j+1},\ \hat{\sigma}^2_{j+1})$ are obtained as

$$\hat{\mu}_{j+1} = \frac{1}{n}\left[\sum_{i=k+1}^{n} x_i + \sum_{i=1}^{k} E_j(X_i | X_i \le DL)\right] \tag{A I-4}$$

$$\hat{\sigma}^2_{j+1} = \frac{1}{n-1}\left[\sum_{i=k+1}^{n} (x_i - \hat{\mu}_j)^2 + \sum_{i=1}^{k} E_j((X_i - \hat{\mu}_j)^2 | X_i \le DL)\right] \tag{A I-5}$$

where

$$E_j(X_i | X_i \le DL) = \hat{\mu}_j - \hat{\sigma}_j[\varphi(Z)/\Phi(Z)] \tag{A I-6}$$

$$\sum_{i=1}^{k} E_j((X_i - \hat{\mu}_j)^2 | X_i \le DL) = \hat{\sigma}_j{}^2(1 - Z[\varphi(Z)/\Phi(Z)]) \tag{A I-7}$$

$$Z = DL - \hat{\mu}_j/\hat{\sigma}_j \tag{A I-8}$$

The advantage of EM algorithm to Newton-Raphson optimization relies in its robustness to starting value, which can be the mean and standard deviation of the uncensored data.

Cohen's MLE

To estimate the mean and standard deviation of normal data censored at a single DL, Cohen (1959) developed a series of equations:

$$\hat{\mu} = \overline{x}_{un} - \lambda(g,h)(\overline{x}_{un} - DL) \tag{A I-9}$$

$$\hat{\sigma} = \sqrt{s^2_{un} + \lambda(g,h)(\overline{x}_{un} - DL)^2} \tag{A I-10}$$

where $\overline{x}_{un}$ and s^2_{un} are the estimates of mean and standard deviation based on uncensored part of the data. In these equation, h and g are defined as

$$h = \frac{k}{n} = \text{Censoring portion} \tag{A I-11}$$

$$g = \frac{s_{un}^2}{(\overline{x}_{un} - DL)^2} \tag{A I-12}$$

Cohen provided look-up tables of the function $\lambda(g,h)$ that are restricted to $g = 0.00\ (0.05)1$. Schneider & Weissfeld (1986) extended these tables to values of g up to 1.48. Haas & Scheff (1990) developed the following power series expansion of the function λ that fits the table values within a 6% relative error:

$$\begin{aligned} \ln \lambda(g,h) \simeq & \frac{0.182344 - 0.3756}{g+1} + 0.10017g + 0.78079y \\ & - 0.00581g^2 - 0.06642y^2 - 0.0234gy + 0.000174g^3 \\ & + 0.001663g^2y - 0.00086gy^2 - 0.00653y^3 \end{aligned} \tag{A I-13}$$

where $y = ln\frac{h}{1-h}$.

Bias corrected MLE

Schneider & Weissfeld (1986) provided computational formulas for the bias-corrected MLEs of the μ and σ based on type II censored and normally distributed data. It is assumed that these correction formulas can be approximately valid for type I censored data, which are typically the case in environmental studies. These formulas are

$$\hat{\mu}_u = \hat{\mu}_c - \frac{\hat{\sigma}_c B_u}{n+1} \tag{A I-14}$$

$$\hat{\sigma}_u = \hat{\sigma}_c - \frac{\hat{\sigma}_c B_\sigma}{n+1} \tag{A I-15}$$

where $\hat{\mu}_c$ and $\hat{\sigma}_c$ are the MLEs obtained by Cohen method, and B_u and B_σ are given as

$$B_u = -e^{2.692 - \frac{5.439(n-k)}{n+1}} \tag{A I-16}$$

$$B_\sigma = -\left(0.312 + \frac{0.859(n-k)}{n+1}\right)^{-2} \tag{A I-17}$$

Restricted MLE

Persson & Rootzen (1977) proposed the restricted MLE, which is similar to the maximum likelihood of Cohen but is simpler to compute. Let $y_i = x_i - DL$; $i = k+1, k+2, \ldots, n$ and $\theta = (DL - \mu)/\sigma$. The likelihood function can be simplified to

$$L = [\Phi(Z)]^k (2\pi\sigma^2)^{-(n-k)/2} \exp\left(-\left[\sum_{i=k+1}^{n} (y_i + Z\sigma)^2/2\sigma^2\right]\right) \qquad \text{(A I-18)}$$

where $\Phi(.)$ is the standard normal distribution function. The $n-k$ uncensored observations can be described by a binomial distribution as

$$\Pr(n-k=r) = \{n!\,\Phi(-\theta)\}^r \{\Phi(\theta)\}^{n-r} / \{r!\,(n-r)!\} \qquad \text{(A I-19)}$$

for $r = 0, \ldots, n$. The $\Phi(\theta)$ can be equivalently defined by $1-(n-k)/n$ for $0 < n-k < n$, thus $\theta^* = \Phi^{-1}\left(1 - \frac{n-k}{n}\right) = \lambda_{(n-k)/n}$, where $\lambda_{(n-k)/n}$ is the upper $(n-k)/n$th quantile of the standard normal distribution. Substituting $\theta^* = \lambda_{(n-k)/n}$ in the likelihood function and then maximizing it yields restricted MLEs as given below.

$$\hat{\sigma}_{rMLE} = \frac{1}{2}\left[\lambda_{(n-k)/n} \frac{1}{(n-k)} \sum_{i=k+1}^{n} y_i + \left\{\left(\lambda_{(n-k)/n} \frac{1}{(n-k)} \sum_{i=k+1}^{n} y_i\right)^2 + \frac{4}{(n-k)} \sum_{i=k+1}^{n} {y_i}^2\right\}\right] \qquad \text{(A I-20)}$$

$$\hat{\mu}_{rMLE} = DL - \lambda_{(n-k)/n} \hat{\sigma}_{rMLE} \qquad \text{(A I-21)}$$

The $\hat{\mu}_{rMLE}$ and $\hat{\sigma}_{rMLE}$ are biased and some correction factor are thus needed. In left-censored data, we have $E[\overline{x}] = \mu + \alpha\sigma$, and $E[s^2] \sim \sigma^2[1 + (\alpha\theta - \alpha^2)]$, where $= \frac{\varphi(\theta)}{1-\Phi(\theta)}$, and the bias corrected restricted MLEs are obtained by the following equations:

$$\hat{\sigma}^*{}_{rMLE} = \left[\frac{1}{(n-k)} \sum_{i=k+1}^{n} x_i^2 - \left(\frac{1}{(n-k)} \sum_{i=k+1}^{n} x_i^2\right)^2 - (\hat{\alpha}\lambda_{\frac{(n-k)}{n}} - \hat{\alpha}^2)\hat{\sigma}^2{}_{rMLE}\right]^{1/2} \qquad \text{(A I-22)}$$

$$\hat{\mu}^*{}_{rMLE} = \frac{1}{(n-k)} \sum_{i=k+1}^{n} x_i - \hat{\alpha}\hat{\sigma}_{rMLE} \qquad \text{(A I-23)}$$

where $\hat{\alpha} = \phi(\lambda_{(n-k)/n})n/(n-k)$. For $k = 0$ (i.e., censoring does not exist), the expressions for $\hat{\mu}^*_{rMLE}$ and $\hat{\sigma}^*_{rMLE}$ are simply $\sum x_i/n$ and $\sqrt{\sum \frac{x_i{}^2}{n} - (\frac{\sum x_i}{n})^2}$.

Robust MLE (rMLE)

The robust MLE (rMLE), proposed by Kroll & Stedinger (1996), is a hybrid of the MLE method with a regression on order statistics. Using the lognormality assumption, the mean and standard deviation in log-scale, $\hat{\mu}_{ln}$ and $\hat{\sigma}_{ln}$, are computed with the MLE method. These estimates are then employed to extrapolate censored values through

$$x_i = \exp(\hat{\mu}_{ln} + \hat{\sigma}_{ln}\Phi^{-1}(p_i)); \quad i = 1,2,\dots,k \tag{A I-24}$$

where $\Phi^{-1}(p_i)$ is the inverse cumulative normal distribution at the plotting position p_i. The plotting positions of k censored values are calculated as

$$p_i = \frac{k}{n}(\frac{i-3/8}{k+1/4}); \quad i = 1,2,\dots,k \tag{A I-25}$$

Since individual extrapolated values for censored observations are transformed back in the original scale by exponentiation, transformation bias in avoided. However, as mentioned in Singh et al. (2006), this estimator is unstable in data sets with high censoring percent.

Tobit regression

The Tobit regression model (Tobin, 1958) is characterized by a regression equation as

$$y_i^* = \alpha + x_i\beta + \varepsilon_i \tag{A I-26}$$

where y_i^* is the dependable variable, x_i is a vector of independent variable, α and β are vectors of regression parameters, and ε_i is the error term that is assumed to be independently and normally distributed with mean 0 and variance σ^2. In environmental studies, dependable variable is typically contaminant concentration in a medium (air, water, soil, etc.) while some measurements are nondetects. The observable concentration variable y_i is related to y_i^* according to

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > DL \\ DL & \text{otherwise} \end{cases} \tag{A I-27}$$

Let $\varphi(.)$ and $\Phi(.)$ denote the normal density and cumulative distribution functions, respectively, the likelihood function for the Tobit model is given by

$$l(\beta, \sigma^2|y, x) = \prod_{y_i > DL} \frac{1}{\sigma} \varphi\left(\frac{y_i - \alpha - x_i\beta}{\sigma}\right) . \prod_{y_i \leq DL} 1 - \Phi(\frac{\mathrm{DL} + \alpha + x_i\beta}{\sigma}) \quad \text{(A I-28)}$$

The estimates of α, β, and σ^2 are obtained by fitting a Tobit line to data by maximum likelihood estimation. The disadvantage of this method relies in its vulnerability to violation of the following assumptions: that residual are normally distributed and the variance is constant across the range of predicted values (i.e., homoscedastic errors). To approximate error normality, in some cases data transformations such as log transformation can be helpful.

Imputation methods

The general idea behind this methodology is to employ a parametric model to impute values for the below DL observations such that complete data are formed and standard statistical methods can be used. Some of the commonly used imputation techniques are discussed in the following. The main advantage of this methodology is that, once nondetects are replaced by imputed values, graphical representation of data is straightforward and any standard statistical method can be used. The disadvantage of these techniques, however, is that the imputed values strongly depend on the goodness of the initial parametric estimates (such as MLEs), and thus these methods are highly sensitive to data skewness and outliers. Moreover, as sample size increases, the number of censored observations increases, and this generally has adverse effects on the performance of the imputation techniques.

Robust Regression on Order Statistics

The fully parametric regression on order statistics (ROS) fits a linear regression to uncensored observations (in original or lognormal scale) against their normal quantiles. The intercept and slope of the regression line estimate the mean and standard deviation of the data (or log-transformed data), respectively. If observations are log transformed, the mean and standard deviations estimates are in logarithmic scale and should be back transformed into the original scale. This retransformation procedure introduces bias in the estimates. To avoid transformation bias, Helsel & Gilliom (1986) presented the robust ROS (rROS), in which censored observations are imputed based on a parametric model, then combined with the uncensored observations to compute the summary statistics of data as if no censoring had occurred. Using the notations in Helsel (2012), the rROS method is performed in four steps as follows:

Step1: Computation of plotting positions for both censored and uncensored observations: after ranking the data, the probability of exceeding j^{th} DL is calculated using the proportion of observations that are at or above that DL. The general formula can be written as

$$pe_j = pe_{j+1} + \frac{A_j}{A_j + B_j}\left[1 - pe_{j+1}\right] \quad \text{(A I-29)}$$

where A_j is the number of uncensored observations between the j^{th} and $j+1^{th}$ DL, and B_j is the number of observations, censored and uncensored, below the j^{th} DL. When j corresponds to the highest DL, in that case $pe_{j+1} = 0$ and $A_j + B_j = n$. The number of observations below the j^{th} DL is defined as

$$C_j = B_j - B_{j-1} - A_{j-1} \quad \text{(A I-30)}$$

To plot a probability plot, we need to calculate plotting positions for uncensored observations as given by

$$pd_i = (1 - pe_j) + \left[\frac{i}{A_j+1}\right].[pe_j - pe_{j+1}]; \quad \text{for } i = 1 \text{ to } A_j \quad \text{(A I-31)}$$

and for censored observation as given by

$$pc_i = \left[\frac{i}{C_j+1}\right].[1 - pe_j]; \quad \text{for } i = 1 \text{ to } C_j \quad \text{(A I-32)}$$

Step 2: Fitting a linear regression line: A regression line is fit to the probability plot, in which y-axis is uncensored data and x-axis is the normal quantiles of the uncensored plotting positions.
Step 3: Extrapolation of censored concentrations: Using the estimates of the mean and standard deviation obtained in step 2 (i.e., the intercept and slope of the regression line) together with the normal quantiles of the censored plotting positions, we can extrapolate values for censored data.
Step 4: Computation of summary statistics: The extrapolated values are combined with uncensored data and standard complete-data methods are used to estimate the summary statistics.

Singh et al. (2002) noted that gamma distribution can adequately fit right skewed environmental data sets. Based on this conclusion, Singh et al. (2006) suggested using the ROS technique that relies on gamma assumption (GROS). The procedure for computing the GROS estimates follows the same steps of the rROS method with two exceptions: a) the log-transformation of uncensored data is not required, and b) instead of normal quantiles, the gamma quantiles of plotting positions are employed. In the GROS method, the probability plot is constructed using the pairs (x_{0i}, x_i); $i = k+1, \ldots, n$, where x_i represents ranked uncensored data and x_{0i} is calculated by the following equation.

$$x_{0i} = z_{0i}\hat{\theta}/2; \quad i = 1,2, \ldots, n \quad \text{(A I-33)}$$

where $\hat{\theta}$ is the estimate of the scale parameter computed by the MLE method (under gamma assumption) and the quantiles z_{0i} are obtained by

$$\int_0^{z_{0i}} f\left(\chi^2_{2\hat{\boldsymbol{\kappa}}}\right) d\chi^2_{2\hat{\boldsymbol{\kappa}}} = (i - 0.5)/n;\ \ i = 1,2,\dots,n \text{ (In the case of single DL)} \quad \text{(A I-34)}$$

$$\int_0^{z_{0i}} f\left(\chi^2_{2\hat{\boldsymbol{\kappa}}}\right) d\chi^2_{2\hat{\boldsymbol{\kappa}}} = p_i;\quad i = 1,2,\dots,n \text{ (In the case of multiple DLs)} \quad \text{(A I-35)}$$

In these equations $\chi^2_{2\hat{\boldsymbol{\kappa}}}$ represents a chi-square random variable with $2\hat{k}$ degrees of freedom (*df*) and p_i is the plotting position can be computed the formulas for pd_i and pc_i as described earlier for the rROS method.

Lynn's method

This is a maximum likelihood-based imputation approach. The steps for the Lynn's method (Lynn, 2001) are described as follows:
Step 1: The maximum likelihood method is used to obtain preliminary estimates of the mean and standard deviation, i.e., $\hat{\mu}$ and $\hat{\sigma}^2$.
Step 2: The $\hat{\mu}$ and $\hat{\sigma}^2$ are used to obtain midlevel estimates of the mean and standard deviation (μ^* and σ^{2*}) by drawing from the following random variables

$$\sigma^{2*} \sim (n-1)\hat{\sigma}^2/\chi^2_{n-1} \quad \text{(A I-36)}$$

$$\mu^* \sim N(\hat{\mu}, \frac{\sigma^{2*}}{n}) \quad \text{(A I-37)}$$

where n is total number of observations (censored and not).
Step 3: Censored observations are substituted by the imputed nondetects that are random draws from the lower tail of $N(\mu^*, \sigma^{2*})$ with the restriction that they are smaller than DL.
Step 4: Combining the imputed values with uncensored ones, final sample estimates can be computed using standard techniques for complete data sets.

Succop's method

The Succop's method (Succop et al., 2004) is also a maximum likelihood-based imputation approach. Preliminary MLEs are served to construct the cumulative distribution function for each DL. The censored observations are then substituted with "the most probable value", which corresponds to half of the percentile at which a laboratory DL falls. For example, if a DL is found at the 10th percentile of a lognormal distribution with $\hat{\mu}$ and $\hat{\sigma}^2$, each censored observation is substituted with the concentration corresponding to the 5th percentile.

Lubin's method

The method proposed by Lubin et al. (2004) is as follows. Assuming lognormality of observations, a Tobit regression followed by bootstrapping is used to compute the sample's statistical parameters. Then, for each censored observation, an imputed value is generated by randomly drawing from a lognormal distribution whose parameters are already estimated with maximum likelihood. This procedure is repeated *M* times (typically between 3 and 5) such that *M* completed data sets and thus *M* estimates are obtained. The estimates based on imputed samples are combined or averaged in order to avoid the bias due to a specific imputed value.

Non-parametric Kaplan-Meier method

The Kaplan-Meier (KM) method was originally developed and used to estimate the survival curve of right-censored data in medical science and reliability analysis. The survival curve defines the probability that the failure time of an event (e.g., death after use of a medicine) goes beyond a given time x, that is $S(x) = P(X > x)$. To make the application feasible for left-censored data, Helsel (1990) suggest "flipping" the data to construct right-censored data sets. A fixed constant (a value larger than the maximum uncensored observation) is chosen and each observation is subtracted from this constant. After calculations, the estimated probabilities can be transformed back in to the original scale.

In the context of left-censored environmental data, the KM estimator estimates the cumulative distribution function, $F(x)$, which defines the probability that an observation is at, or below, a reported concentration. The cumulative distribution and survival function are complements of each other thus $\hat{F}(x) = 1 - \hat{S}(x)$. Following Singh et al. (2006) notations, we describe how the KM method estimates the mean and standard deviation of data based on $\hat{F}(x)$.

Let $x_1, x_2, \dots, x_n$ be n concentration observations, $y_1, y_2, \dots, y_p$ be p distinct uncensored observations, b_j denote the number of observations at and below each detected concentration, and d_j represent the number of uncensored concentrations equal to $y_j; j = 1,2, \dots, p$.
The estimated cumulative function is defined by

$$\hat{F}(x) = 1 \qquad x \geq y_p \qquad \text{(A I-38)}$$

$$\widehat{F}(x) = \prod_{j \ni y_j > x}^{p} \frac{b_j - d_j}{b_j} \qquad y_1 \leq x \leq y_{p-1} \qquad \text{(A I-39)}$$

$$\hat{F}(x) = F(y_1) \qquad x_1 \leq x \leq y_1 \qquad \text{(A I-40)}$$

$$\hat{F}(x) = 0 \qquad 0 \leq x \leq x_1 \qquad \text{(A I-41)}$$

The resultant $\hat{F}(x)$ is a step function that drops at each uncensored concentration and remains constant for censored values. The mean value is estimated as the area between $\hat{F}(x)$ and 1.0; mathematically it is

$$\hat{\mu} = \sum_{j=1}^{p} y_j [F(y_j) - F(y_{j-1})] \quad \text{(A I-42)}$$

An estimate of the standard error of the mean is given by

$$se(\hat{\mu}) = \frac{n-k}{n-k-1} \sum_{j=1}^{p-1} a_j^2 \frac{d_{j+1}}{b_{j+1}(b_{j+1} - d_{j+1})} \quad \text{(A I-43)}$$

where k is the number of uncensored data and a_j is defined as

$$a_j = \sum_{i=1}^{j} (y_{i+1} - y_i) F(y_i) \quad \text{for } j = 1,2, \dots, p-1 \quad \text{(A I-44)}$$

An estimate of the variance of censored data is computed from

$$\hat{\sigma}^2 = \sum_{j=1}^{p} (y_j - \hat{\mu})^2 [F(y_j) - F(y_{j-1})] \quad \text{(A I-45)}$$

The principle advantage of KM is that it does not involve data transformation to obtain normality or require any distributional assumption about the shape of data. However, being a non-parametric method, it relies exclusively on data and cannot use a model to estimate probable values for the below DL observations. When all nondetects are censored at the same DL (single-censored data), Helsel (2010b) does not recommend KM to estimate the mean because that estimate would be equal to the mean estimated after substituting data with DL.

A cautionary note on employing software platforms for KM computation

Two most commonly employed Software platforms for computing KM are Minitab and NADA package in R. As mentioned in the previous section, KM estimates the mean by integrating the area delimited by the $\hat{F}(x)$ curve. Since this curve is a step function, the integration is computed by summing the area of horizontal rectangles having the length equal to the value of observation and width equal to the corresponding cumulative probability. The problem arises when the smallest observation of data is left-censored. In such situation, Gillespie et al. (2010) discusses that the left end of $\hat{F}(x)$ curve is "hanging", making it impossible to calculate the area delimited by the plot (Figure-A I-1). Minitab and R programs address this problem differently. We illustrate their different approaches through an artificial data set, which was reported in supplementary materials of the Gillespie et al. (2010). This data set is <3, 4, 6, 8, <10, 12.

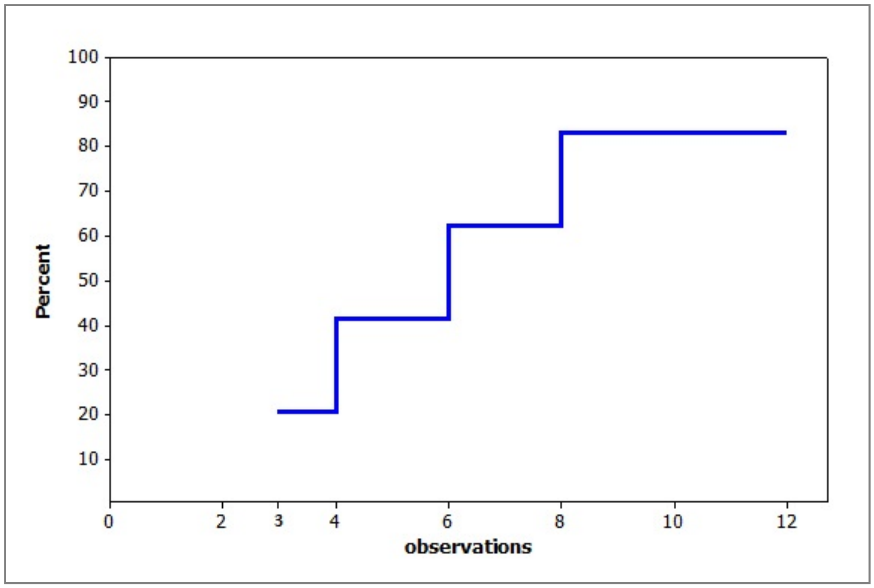


Figure-A I-1 The estimated cumulative distribution curve

Minitab approach: As shown in Figure-A I-1, the $\hat{F}(x)$ cannot be estimated for observations smaller than 3. Since there is no information on how the plot would proceed for observations <3, Minitab software assumes that the probability of having observations less than 3 is zero. Mathematically, it can be represented as $\hat{F}(x) = 0$ for $0 \leq x \leq x_1$. Under this assumption, the censored observation is actually considered as an uncensored one leading to an overestimation of the mean. The shaded area of the plot in Figure-A I-2 illustrates the Minitab approach in estimating the mean.

NADA package approach*:* NADA ignores the presence of the first censored value, i.e., x_1. Therefore, the probability of having observations less than 3 equals that of the smallest uncensored observation. Mathematically, it is $\hat{F}(x) = \hat{F}(y_p)$. NADA computes the area under the curve up to the first uncensored observation (see shaded area in Figure-A I-3).

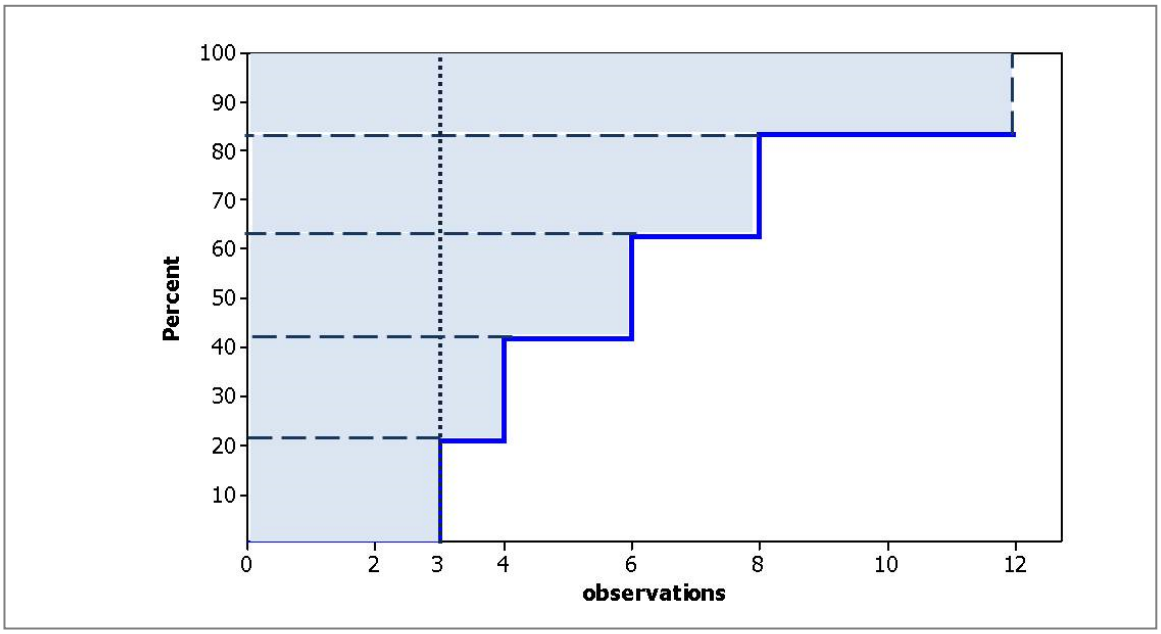


Figure-A I-2 The Minitab procedure for KM estimation of the mean;
the dotted line represents the position of the smallest observation

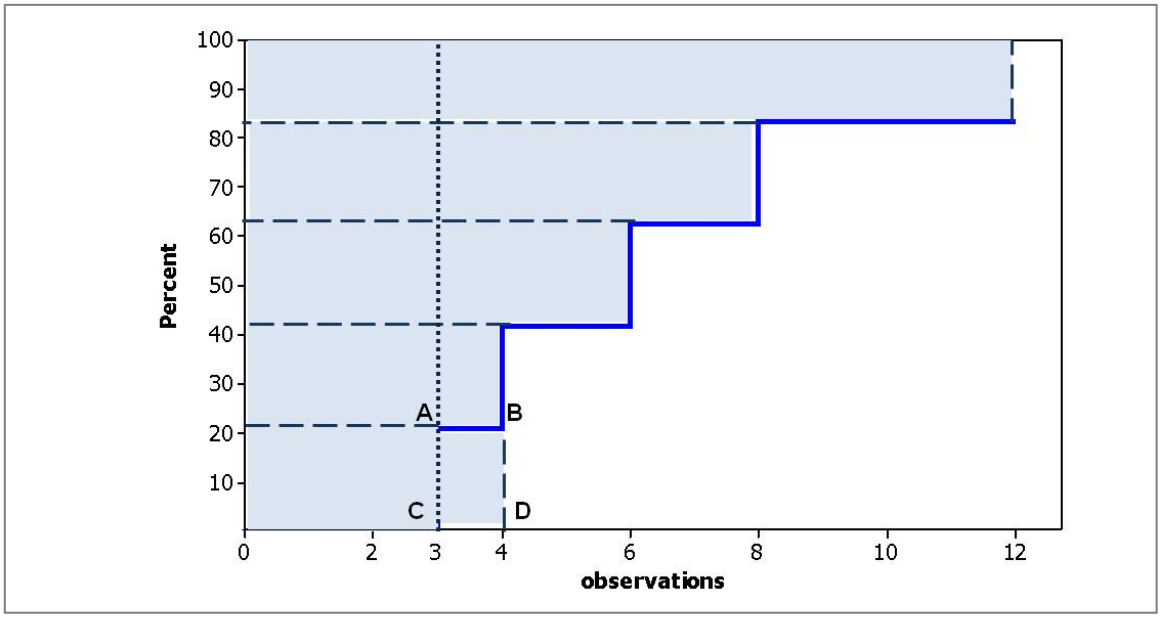


Figure-A I-3 The NADA package procedure for KM estimation of the mean;
the dotted line represents the position of the smallest observation

As shown in Figure-A I-3, mean estimation through NADA package has an overestimation respect to that estimated by the Minitab Software. This overestimation is equal to the area of the ABCD rectangle in Figure-A I-3. Under the following conditions, such overestimation has more evident impact on the estimates:

a) When the low end of data contains a succession of censored values that increases the length of ABCD rectangle. In other words, the distance between the smallest censored and the smallest uncensored values becomes larger;
b) When occurrence of having the smallest censored value(s) ahead of other observations is high and this increases the height of ABCD rectangle. That is, estimates of cumulative probabilities for these censored values are large.

APPENDIX II

ANDERSON-DARLING GOODNESS OF FIT TEST FOR LEFT-CENSORED ENVIRONMENTAL DATA

Despite a rich literature about distributional checking procedures tailored for left-censored data, these are generally overlooked in environmental studies due to the complexity of implementation. The most commonly studied categories of procedures are described below.

Graphical model selection procedures visually assess the appropriateness of a parametric model. As discussed earlier, probability plotting has been a common graphical procedure to assess the normality and lognormality of concentration data and to estimate their respective statistical parameters (e.g., Huybrechts et al., 2002 and Helsel, 2005).

Likelihood-based information procedure compares the likelihood of fitting a given distribution penalizing on the number of parameters in the model. The commonly used ones in this category are the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). For example, European Food Safety Authority (2010) used the AIC and BIC model selection criteria to assess the goodness of MLEs based on lognormal, Weibull, and gamma assumptions. They showed that concentration data of dinophysis toxins were generally well described by lognormal distribution. In another example of using the AIC, Singh, Bartolucci & Bae (2001) proposed that generalized log-logistic distribution was a better fit to environmental data when compared to its competitors such as lognormal, Weibull and gamma.

Empirical cumulative distribution function-based (ECDF) procedure compares the ECDF with its parametric counterpart. The most relevant tests in this category include Hollander and Proschan (HP), Anderson-Darling (AD), Cramer-von Mises (CvM), and Kolmogorov-Smirnov (KS). These tests are explained in more detail in D'Agostino & Stephens (1986). Tate & Freeman (2000) applied HP test to censored droughts duration data and found out that either Weibull or exponential were adequately fit. In a study conducted by Zhao & Frey (2003), the distribution of urban air toxic emissions (benzene, formaldehyde, chromium, and arsenic) was evaluated by the KS test; however, censored observations were substituted with DL/2. The AD goodness of fit test was used for left-censored financial data to check whether power-law distribution could adequately represent the data (Coronel-Brizio & Hernández-Montoya, 2010).

Goodness of fit based on the Anderson-Darling statistic

A goodness of fit test is a statistical hypothesis test to assess whether a random sample of $X = \{x_1, x_2, \dots, x_n\}$ comes from a specified distribution, $F(x|\theta)$ where θ is a vector of distributional parameters. The null and alternative hypotheses are:

$$H_0: X \in F(x|\theta)$$
$$H_a: X \notin F(x|\theta) \quad \text{(A II-1)}$$

The AD test statistic for testing the H_0 is based on measuring the distance between ECDF and cdf. Mathematically, the AD statistic takes the form

$$n\int_{-\infty}^{\infty} \frac{\hat{F}(x) - \hat{F}(x|\theta)]^2}{\hat{F}(x|\theta)[1-\hat{F}(x|\theta)]} d\hat{F}(x|\theta) \quad \text{(A II-2)}$$

where $\hat{F}(x)$ is the Kaplan-Meier estimate of ECDF and $\hat{F}(x|\theta)$ is the maximum likelihood estimate of cdf.

Let $x_{(1)} < x_{(2)} < \cdots < x_{(n-r)} < x_{(n-r+1)} < x_{(n-r+2)} < \cdots < x_n$ be an ordered data set in which the *n-r* smallest observations are left-censored and the remaining *r* observations are uncensored. For convenience, assuming that $Z_{(.)}$ is the cdf of an assumed distribution evaluated at $x_{(.)}$, we have $z_{(1)} < z_{(2)} < \cdots < z_{(n-r)} < z_{(n-r+1)} < z_{(n-r+2)} < \cdots < z_{(n)}$. D'Agostino and Stephens (1986) define the computing formula for the AD statistic for singly-censored data as

$$AD = -\frac{1}{n}\sum_{i=1}^{r+1}(2i-1)\{\ln[1-z_{(n-i+1)}] - \ln z_{(n-i+1)}\} - 2\sum_{i=1}^{r+1}\ln z_{(n-i+1)} - \frac{1}{n}[(r-n)^2 \ln z_{(n-r+1)} - r^2 \ln z_{(n-r+1)} + n^2(1-z_{(n-r+1)}] \quad \text{(A II-3)}$$

where $z_{(n+1-i)} = F(x_{(n-i+1)}|\theta)$ for *i=1,..,r+1* and $z_{(n-r+1)} = F(x_{(n-r+1)})|\theta)$.

If the AD statistic obtained exceeds a critical value at a significance level α the null hypothesis is rejected. The main difficulty with this procedure is that the critical values are sensitive to a number of factors such as the model being fitted; therefore, a single Table of critical values does not exist. In this situation, a parametric bootstrap approach is useful to overcome this problem. Using a bootstrapping technique enables us to characterize the asymptotic null distribution of the test statistic AD so that we can estimate either a critical value or p-value of the test.

Assessing the performance of AD statistic

We study the behavior of goodness of fit based on the AD statistic applied for left-censored data sets. We employ simulations to investigate the distribution of p-values as a criterion for assessing the reliability of the AD test statistic in distinguishing the correct parametric family. Under the H_0, p-values are expected to be uniformly distributed. Any strong deviation from this expected distribution indicates the inappropriateness of the statistical test.

Simulation scenarios

In the goodness of fit hypothesis testing discussed here, $F(x|\theta)$ is referred to lognormal, Weibull and gamma distributions as these are candidate distributions to model right-skewed environmental data. The cdf for lognormal distribution with mean μ_{ln} and variance σ_{ln}^2 in logarithmic scale is defined by

$$F(x|\mu_{ln}, \sigma_{ln}^2) = \Phi(\frac{logx - \mu_{ln}}{\sqrt{\sigma_{ln}^2}}) \quad \text{(A II-3)}$$

For Weibull distribution with shape (α) and scale (β) the cdf is given by

$$F(x|\alpha, \beta) = 1 - e^{-(x/\beta)^{\alpha}} \quad \text{(A II-4)}$$

The gamma distribution with shape α and rate β has cdf

$$F(x|\alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad \text{(A II-5)}$$

where $\Gamma(.)$ is the gamma function.

Random samples are generated from one of the above-mentioned distributions with μ=1 and σ takes any of values 0.5, 1.5, 3. Specifically, the generated samples are of size n=60 and 200. Moreover, we allocate a fictional censoring point at 30th, 50th and 70th quantile of the data generating distributions so that the observations below the computed censoring point are attributed as censored. In this way, censored data sets with 30%, 50%, and 70% censoring percentage are obtained. Assuming that the distribution $\hat{F}(x|\theta)$ is a good fit for data, equation A II-3 is used to calculate the AD statistic for the simulated data set. Consequently, we perform a parametric bootstrapping by drawing several bootstrap samples, X_b^*, *b=1,2,...,B*, from the fitted distribution $\hat{F}(x|\theta)$. Note that B is sufficiently large (1000 for example) and each bootstrap sample contains the same number of observations and the same censoring percentage as original data. For each bootstrap sample the AD statistic, AD_b^*, *b=1,2,...1000*, is computed in exactly the same way the AD was computed from the original data. The p-value is, approximately, the fraction of the number of times the AD_b^* is larger than the AD. For each combination of data generating distribution, μ, σ, n, and censoring percentage, N=1000 replications are simulated. In each replication, we compute the AD from the simulated data, AD^* for the bootstrap samples, and consequently the corresponding p-value. If the simulated data come from the same model stated in the H_0, the distribution of obtained p-values should be uniform over the interval [0,1]. In contrast, if the simulated data arise from any alternative distributions, p-values distributions tend to cluster toward zero. Detailed simulation procedure is as follow.

Step 1: Generate a random sample of n from one of these distributions: lognormal $X \sim Ln(\mu_{ln}, \sigma_{ln})$, Weibull $X \sim Weib(\alpha, \beta)$, and gamma $X \sim Gm(\alpha, \beta)$ with given μ and σ.

Step 2: Randomly censor data sets by imposing a single censoring point (i.e., DL) at 30th, 50th, and 70th quantiles of data generating distributions.
Step 3: Estimate distributional parameter, θ, of the simulated data with a consistent estimator, $\hat{\theta}$. In the case of lognormal distribution, for example, $\hat{\theta}$ represents the estimates of $\hat{\mu}_{ln}$ and $\hat{\sigma}_{ln}$.
Step 4: Obtain $z = F(x_{(n-i+1)}|\hat{\theta})$, for *i=1,2,...,r+1*.
Step 5: Evaluate the AD test statistic for the simulated data using the equation A II-3.
Step 6: Carry out a parametric bootstrap to estimate the p-value of the test. The bootstrap estimate of the p-value is computed as follows.

- Generate *B*=1000 bootstrap samples, X_b^*, *b=1,2,...,B*, of size *n* from the distribution in null hypothesis. Each bootstrap sample has the same sample size and censoring percentage as in step 1 and step 2;
- Estimate the distributional parameters of each bootstrap sample, $\hat{\theta}_b^*$, *b=1,2,...,B*;
- Compute the AD for each bootstrap sample, AD_b^*, *b=1,2,...,B*;
- The estimated p-value is computed as p= $\frac{\#AD_b^* \geq AD}{B}$.

Step 7: Repeat step 1-6 for *N*=1000 times.
Step 8: Calculate the average of p-values and plot the histogram of p-values.

Preliminary results

As mentioned earlier, the p-values are uniformly distributed if the distribution of the simulated data conforms to that examined in null distribution. In our case, this can be seen in Figure-A II-1.a, where the null hypothesis tests whether the data follow a lognormal distribution and the simulated data are indeed generated from a lognormal distribution. In contrast, when data are generated from an alternative model (Weibull or gamma), the p-values in Figure-A II-1.b and Figure-A II-1.c tend to cluster closer to zero, suggesting that the test more often rejects the null hypothesis.

The histograms of the p-values give information about the type I error and type II error of the goodness of fit test based on the AD statistic. For a nominal significance level α=0.05, the probability of rejecting the null hypothesis when it is in fact true is define as the type I error; while the probability of falsely accepting the null hypothesis is represented by the type II error. From the latter, the power of a test, that is the probability of correctly rejecting the null hypothesis, can be estimated. For example, when data are generated from the lognormal distribution with $\mu = 1$, $\sigma = 0.5$, and 30% censoring, the simulations report 0.06 of the p-values are actually less than the significance level (results highlighted in red in Figure-A II-1.a). The closeness of this proportion to the nominal value confirms that the used goodness of fit test performs well. When data are generated from alternative distributions, Figure-A II-1.b and Figure-A II-1.c show that the test correctly rejects the null hypothesis more often (red bars highlighted in red).

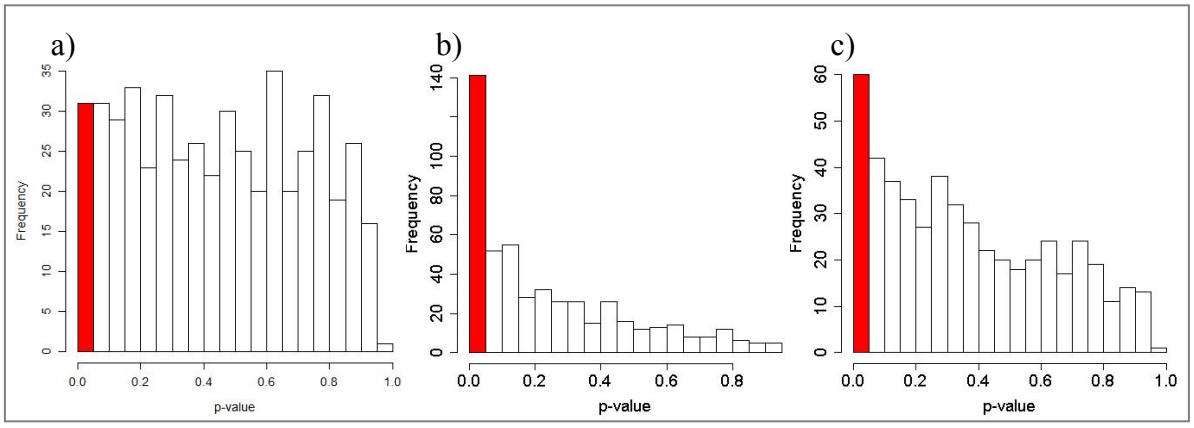


Figure-A II-1 Distribution of p-values when a) data come from the assumed distribution (lognormal), b) data do not come from the assumed distribution but from a Weibull c) data do not come from the assumed distribution but from a gamma

Table-A II-1 summarizes the results in terms of proportion of rejection under the null and alternative hypotheses (at 0.05 significance level), for different censoring percentage and sample sizes. The former reflects the type I error (the fourth column), while the latter represents the power of the test (the fifth and sixth columns). When the null hypothesis tests whether the data arise from a lognormal distribution, the most salient aspects of the simulation outcomes are as follows:

- When sample size is 60, the estimated type I error is slightly larger than the nominal level (α=0.05);
- When sample size increases to 200, the power of the AD goodness of fit increases. On the other hand, the power of test decreases with increasing percentage of censoring. Indeed, when more that 50% of the data is censored, then it is almost impossible to distinguish between lognormal, Weibull, and gamma distributions;
- For a fixed sample size, as σ increases (equivalently, as the degree of skewness increases), the power of the goodness of fit test in discriminating between lognormal and gamma distributions increases. However, this behavior is not observed when the interest lies in distinguishing between lognormal and Weibull distributions.

Based on the preliminary results, we found that the goodness of fit test based on AD statistic exhibit a good performance only for large data sets. Another limitation to this methodology is that it is restricted to singly censored data sets in which censored values are ranked before the uncensored observations.

Table-A II-1 Simulation results for the null hypothesis of the lognormal distribution when data are generated from a lognormal, a Weibull, or a gamma distribution

n	Censoring %	σ	X~lognormal	X~Weibull	X~Gamma
60	30%	0.5	0.06	0.28	0.12
60	30%	1.5	0.06	0.28	0.54
60	30%	3	0.07	0.28	0.93
60	50%	0.5	0.08	0.20	0.11
60	50%	1.5	0.08	0.16	0.32
60	50%	3	0.09	0.17	0.68
60	70%	0.5	0.07	0.10	0.09
60	70%	1.5	0.07	0.10	0.16
60	70%	3	0.09	0.10	0.35
200	30%	0.5	0.06	0.73	0.25
200	30%	1.5	0.06	0.73	0.99
200	30%	3	0.06	0.73	1.00
200	50%	0.5	0.06	0.40	0.16
200	50%	1.5	0.07	0.40	0.71
200	50%	3	0.06	0.42	1.00
200	70%	0.5	0.09	0.16	0.11
200	70%	1.5	0.07	0.18	0.37
200	70%	3	0.07	0.17	0.80

APPENDIX III

SUPPLEMENTARY MATERIAL OF ARTICLE 1

The skewness of random variable X is defined by

$$\gamma = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \tag{A III-1}$$

where μ is the mean and σ is the standard deviation.

By expanding the previous formula, we obtain

$$\begin{aligned}\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] &= \frac{\mathbb{E}[X^3] - 3\mathbb{E}[X]\mathbb{E}[X^2] + 2(\mathbb{E}[X])^3}{\sigma^3} \\ &= \frac{\mu^{(3)} - 3\mu^{(1)}\mu^{(2)} + 2(\mu^{(1)})^3}{\sigma^3}\end{aligned} \tag{A III-2}$$

where $\mu^{(1)}$, $\mu^{(2)}$, and $\mu^{(3)}$ are the first, second, and third moments of X. Using the moment generating function (MGF), we can find the above-mentioned moments. In fact, for $r = 1,2,3$, the r^{th} moment can be found by evaluating the r^{th} derivative of the MGF at zero.

Suppose that X follows the lognormal distribution with parameters μ_y and s_y (mean and standard deviation in log-scale), the moments are given by

$$\mu^{(1)} = \exp(\mu_y + 0.5{s^2}_y) \tag{A III-3}$$

$$\mu^{(2)} = \exp(2(\mu_y + {s^2}_y)) \tag{A III-4}$$

$$\mu^{(3)} = \exp(3\mu_y + \frac{9}{2}{s^2}_y) \tag{A III-5}$$

Suppose that X follows the Weibull distribution with scale parameter λ and shape parameter k, the moments are given by

$$\mu^{(1)} = \lambda\Gamma(1 + \frac{1}{k}) \tag{A III-6}$$

$$\mu^{(2)} = \lambda^2\Gamma(1 + \frac{2}{k}) \tag{A III-7}$$

$$\mu^{(3)} = \lambda^3\Gamma(1 + \frac{3}{k}) \tag{A III-8}$$

Suppose that X follows the gamma distribution with shape parameter α and rate parameter λ, the moments are given by

$$\mu^{(1)} = \frac{\alpha}{\lambda} \quad \text{(A III-9)}$$

$$\mu^{(2)} = \frac{\alpha^2 + \alpha}{\lambda^2} \quad \text{(A III-10)}$$

$$\mu^{(3)} = \frac{\alpha^3 + 3\alpha^2 + 2\alpha}{\lambda^3} \quad \text{(A III-11)}$$

Table-A III-1 The MSE of the mean and standard deviation produced by rROS, GROS and MLE under different distributional assumptions and model misspecification in scenarios with 30% censoring

		MSE of the mean estimates					MSE of the standard deviation estimates				
True dist.	**Parameters ($\mu = 1$)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**
Weibull	$\sigma = 0.5$	0.005	NA	0.004	0.005	0.005	0.006	NA	0.003	0.004	0.004
	$\sigma = 1.2$	0.062	NA	0.024	0.024	0.025	1.874	NA	0.048	0.065	0.063
	$\sigma = 1.9$	0.625	NA	0.062	0.062	0.062	168.92	NA	0.298	0.490	0.487
	$\sigma = 2.6$	3.700	NA	0.108	0.109	0.108	*	NA	1.022	1.556	1.554
	$\sigma = 3.3$	23.983	NA	0.167	0.167	0.166	*	NA	2.432	3.212	3.210
	$\sigma = 4$	62.824	NA	0.248	0.249	0.248	*	NA	4.494	6.513	6.511
Percentage of error		*	-	2%	4%	5%	*	-	3%	49%	44%
gamma	$\sigma = 0.5$	0.005	0.005	NA	0.005	NA	0.006	0.004	NA	0.004	NA
	$\sigma = 1.2$	0.089	0.025	NA	0.025	NA	3.545	0.073	NA	0.064	NA
	$\sigma = 1.9$	26.497	0.134	NA	0.065	NA	*	3.534	NA	0.298	NA
	$\sigma = 2.6$	*	4.917	NA	0.116	NA	*	*	NA	0.918	NA
	$\sigma = 3.3$	*	*	NA	0.185	NA	*	*	NA	2.055	NA
	$\sigma = 4$	*	*	NA	0.264	NA	*	*	NA	4.332	NA
Percentage of error		*	*	-	1%	-	*	*	-	15%	-
lognormal	$\sigma = 0.5$	NA	0.005	0.004	NA	0.006	NA	0.006	0.005	NA	0.01
	$\sigma = 1.2$	NA	0.021	0.022	NA	0.025	NA	0.106	0.110	NA	0.162
	$\sigma = 1.9$	NA	0.046	0.074	NA	0.075	NA	0.574	0.689	NA	1.829
	$\sigma = 2.6$	NA	0.060	0.101	NA	0.102	NA	1.584	1.826	NA	2.924
	$\sigma = 3.3$	NA	0.088	0.169	NA	0.169	NA	3.311	3.800	NA	5.819
	$\sigma = 4$	NA	0.101	0.190	NA	0.191	NA	5.698	6.465	NA	7.436
Percentage of error		-	3%	53%	-	61%	-	1%	11%	-	91%

(Continued)

		MSE of the mean estimates					MSE of the standard deviation estimates				
True dist.	**Parameters ($\mu = 1$)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**
Mixture Weibull	$\sigma = 1.00$	0.217	0.238	0.225	0.207	0.226	0.326	0.326	0.341	0.380	0.350
	$\sigma = 1.38$	0.249	0.302	0.301	0.282	0.336	0.151	0.228	0.235	0.218	0.187
	$\sigma = 1.88$	0.276	0.419	0.403	0.396	0.425	3.694	0.505	0.575	0.572	0.557
	$\sigma = 2.43$	0.332	0.541	0.501	0.500	0.512	91.813	1.174	1.394	1.560	1.550
	$\sigma = 3.00$	0.602	0.645	0.604	0.606	0.610	*	2.419	2.952	4.626	4.619
	$\sigma = 3.58$	1.088	0.778	0.729	0.732	0.732	*	4.239	5.221	7.769	7.763
Percentage of error		10%	29%	22%	19%	27%	*	9%	23%	47%	41%
Mixture gamma	$\sigma = 1.00$	0.021	0.019	0.019	0.019	0.026	0.078	0.009	0.014	0.007	0.008
	$\sigma = 1.38$	0.198	0.031	0.032	0.034	0.037	7.284	0.099	0.108	0.03	0.032
	$\sigma = 1.88$	20.112	0.113	0.060	0.062	0.061	*	3.286	0.415	0.176	0.193
	$\sigma = 2.43$	*	3.552	0.100	0.101	0.093	*	276.565	1.061	0.621	0.665
	$\sigma = 3.00$	*	116.169	0.150	0.151	0.136	*	*	2.075	1.470	1.546
	$\sigma = 3.58$	*	*	0.221	0.222	0.2	*	*	3.52	3.502	3.627
Percentage of error		*	*	6%	8%	12%	*	*	116%	3%	12%
Mixture lognormal	$\sigma = 1.00$	0.022	0.021	0.02	0.021	0.03	0.087	0.011	0.017	0.007	0.010
	$\sigma = 1.38$	0.067	0.031	0.032	0.033	0.035	1.189	0.058	0.055	0.101	0.105
	$\sigma = 1.88$	0.169	0.055	0.058	0.059	0.058	5.170	0.155	0.130	0.476	0.475
	$\sigma = 2.43$	0.404	0.088	0.102	0.104	0.099	20.387	0.478	0.498	1.566	1.562
	$\sigma = 3.00$	0.632	0.086	0.132	0.134	0.129	51.450	0.963	1.309	4.306	4.301
	$\sigma = 3.58$	1.016	0.135	0.309	0.311	0.303	115.994	2.016	3.304	12.468	12.464
Percentage of error		332%	3%	36%	38%	45%	*	12%	40%	241%	248%

NA : Not Applicable

Table-A III-2 The MSE of the mean and standard deviation produced by rROS, GROS and MLE under different distributional assumptions and model misspecification in scenarios with 70% censoring

		MSE of the mean estimates					MSE of the standard deviation estimates				
True dist.	**Parameters ($\mu = 1$)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**
Weibull	$\sigma = 0.5$	0.011	NA	0.01	0.014	0.014	0.007	NA	0.006	0.010	0.008
	$\sigma = 1.2$	0.041	NA	0.028	0.034	0.038	0.597	NA	0.065	0.077	0.074
	$\sigma = 1.9$	0.169	NA	0.060	0.068	0.060	18.54	NA	0.333	0.490	0.472
	$\sigma = 2.6$	0.746	NA	0.109	0.116	0.105	*	NA	1.047	1.574	1.546
	$\sigma = 3.3$	3.427	NA	0.175	0.181	0.171	*	NA	2.602	3.571	3.532
	$\sigma = 4$	13.431	NA	0.277	0.284	0.272	*	NA	4.930	7.530	7.493
Percentage of error		*	-	5%	22%	20%	*	-	10%	66%	50%
gamma	$\sigma = 0.5$	0.008	0.009	NA	0.010	NA	0.007	0.005	NA	0.008	NA
	$\sigma = 1.2$	0.049	0.030	NA	0.041	NA	0.610	0.076	NA	0.073	NA
	$\sigma = 1.9$	0.970	0.079	NA	0.074	NA	*	1.144	NA	0.319	NA
	$\sigma = 2.6$	*	0.478	NA	0.123	NA	*	57.946	NA	1.026	NA
	$\sigma = 3.3$	*	13.875	NA	0.194	NA	*	*	NA	2.122	NA
	$\sigma = 4$	*	*	NA	0.257	NA	*	*	NA	3.964	NA
Percentage of error		*	*	-	16%	-	*	*	-	14%	-
lognormal	$\sigma = 0.5$	NA	0.015	0.011	NA	0.028	NA	0.010	0.009	NA	0.020
	$\sigma = 1.2$	NA	0.029	0.030	NA	0.054	NA	0.114	0.107	NA	0.170
	$\sigma = 1.9$	NA	0.053	0.059	NA	0.074	NA	0.607	0.564	NA	0.879
	$\sigma = 2.6$	NA	0.081	0.113	NA	0.121	NA	1.997	1.789	NA	3.337
	$\sigma = 3.3$	NA	0.098	0.150	NA	0.153	NA	3.157	3.394	NA	5.367
	$\sigma = 4$	NA	0.147	0.228	NA	0.230	NA	6.528	6.015	NA	8.399
Percentage of error		-	17%	36%	-	95%	-	8%	1%	-	74%

(Continued)

		MSE of the mean estimates					MSE of the standard deviation estimates				
True dist.	**Parameters ($\mu = 1$)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**	**MLE (lognormal)**	**MLE (Weibull)**	**MLE (gamma)**	**rROS (lognormal)**	**GROS (gamma)**
Mixture Weibull	$\sigma = 1.00$	0.216	0.277	0.236	0.203	0.252	0.370	0.309	0.357	0.388	0.329
	$\sigma = 1.38$	0.275	0.341	0.332	0.266	0.392	0.220	0.210	0.225	0.216	0.156
	$\sigma = 1.88$	0.378	0.460	0.458	0.391	0.478	2.866	0.532	0.546	0.541	0.493
	$\sigma = 2.43$	0.458	0.557	0.543	0.493	0.530	140.236	1.560	1.380	1.754	1.712
	$\sigma = 3.00$	0.592	0.655	0.628	0.585	0.600	*	3.137	2.689	3.215	3.163
	$\sigma = 3.58$	1.346	0.738	0.723	0.694	0.689	*	7.090	4.937	7.013	6.965
Percentage of error		26%	31%	24%	10%	29%	*	20%	12%	27%	15%
Mixture gamma	$\sigma = 1.00$	0.06	0.050	0.05	0.075	0.074	0.032	0.008	0.012	0.024	0.012
	$\sigma = 1.38$	0.129	0.104	0.097	0.168	0.134	0.306	0.059	0.042	0.127	0.102
	$\sigma = 1.88$	0.420	0.103	0.100	0.169	0.090	184.978	0.268	0.156	0.295	0.291
	$\sigma = 2.43$	*	0.716	0.102	0.111	0.086	*	51.947	0.682	0.662	0.740
	$\sigma = 3.00$	*	14.776	0.152	0.158	0.126	*	*	1.427	1.653	1.763
	$\sigma = 3.58$	*	500.288	0.206	0.214	0.167	*	*	2.223	3.262	3.417
Percentage of error		*	*	12%	49%	14%	*	*	11%	93%	64%
Mixture lognormal	$\sigma = 1.00$	0.059	0.050	0.050	0.074	0.072	0.037	0.008	0.013	0.020	0.013
	$\sigma = 1.38$	0.093	0.070	0.069	0.119	0.096	0.157	0.082	0.061	0.169	0.158
	$\sigma = 1.88$	0.116	0.075	0.078	0.128	0.078	0.699	0.339	0.288	0.837	0.811
	$\sigma = 2.43$	0.139	0.092	0.122	0.178	0.112	1.510	0.968	1.021	3.299	3.221
	$\sigma = 3.00$	0.175	0.104	0.135	0.206	0.118	5.170	1.960	1.945	4.378	4.254
	$\sigma = 3.58$	0.205	0.110	0.158	0.214	0.141	8.432	3.347	3.472	6.719	6.492
Percentage of error		52%	0%	18%	79%	25%	172%	9%	12%	164%	140%

NA : Not Applicable

APPENDIX IV

SUPPLEMENTARY MATERIAL OF ARTICLE 2

Algorithm-A IV-1 R code for estimating the mean and standard deviation based on data generated from lognormal distribution with 50% censoring

```
library(NADA)
library(fitdistrplus)
library(xts)
library(mixdist)
library(car)
n <-60          #sample size
N <- 1000    #number of iterations
b <- rep(0,N)
mean.sub <- rep(0,N)
sd.sub <- rep(0,N)
mean.diff.sub <- rep(0,N)
sd.diff.sub <-rep(0,N)
mean.mle.L <- rep(0,N)
sd.mle.L <-rep(0,N)
mean.diff.mle.L <- rep(0,N)
sd.diff.mle.L <- rep(0,N)
mean.KM <-rep(0,N)
sd.KM <-rep(0,N)
mean.diff.KM <-rep(0,N)
sd.diff.KM<-rep(0,N)
mean.ROS <-rep(0,N)
sd.ROS <-rep(0,N)
mean.diff.ROS<-rep(0,N)
sd.diff.ROS<-rep(0,N)
mu <- 1
sigma <- c(0.5,1.2,1.9,2.6,3.3,4)
p <- c(0.2,0.4,0.6,0.8)
lm <- length(mu)
ls <- length(sigma)
results <- matrix(NA,lm*ls,11)
colnames(results)<-
c("mu","sigma","PC","sub.mean","mle.LOG.mean","KM.mean","ROS.mean","sub.sd",
"mle.LOG.sd","KM.sd","ROS.sd")
```

```
B.results <- matrix(NA,lm*ls,8)
    colnames(B.results)<-
c("B.sub.mean","B.mle.LOG.mean","B.KM.mean","B.ROS.mean",
                "B.sub.sd","B.mle.LOG.sd","B.KM.sd","bias.ROS.sd")

                for(i in 1:lm){
                for (j in 1:ls){

                s<-(log(1+(sigma[j]^2/mu[i]^2)))^0.5
                m <- log(mu[i])-((s^2)*0.5)
                CP <- qlnorm(p,m,s)             #### computed censoring points ####

                results2 <- matrix(NA,N,8)

                colnames(results2)<- c("subs mean","LOG MLE mean","KM mean","ROS
mean","subs sd","log MLE sd", "KM sd","ROS sd")

for (t in 1:N)              {
  k <- j+(ls*(i-1))
  y <- rlnorm(n,m,s)
  c<-sample(CP,n,replace=TRUE)
  my.data <- data.frame(y,c)
  my.data$obs<-pmax(y,c)
  my.data$cens <-ifelse(y<c,"TRUE","FALSE")
  a<-my.data$cens
  b[t]<-length(a[a=="TRUE"])
  mean(b)
  newdata<-my.data[,3:4]
  pc <- signif(mean(b)/n,digits=2)

  #### Substitution-based  method  ####

  data.sub <- ifelse(my.data$cens=="TRUE",my.data$c*0.5,my.data$obs)
  mean.sub[t] <- mean(data.sub)
  sd.sub [t] <- sd(data.sub)
  mean.diff.sub [t]<- mean.sub[t] -mu[i]
  sd.diff.sub[t] <- sd.sub[t]-sigma[j]


  #### MLE Lognormal ####
  MLE.L<-with(newdata,cenmle(newdata$obs,as.logical(newdata$cens)))
  mean.mle.L[t] <- mean(MLE.L)[[1]]
  sd.mle.L[t]<-sd(MLE.L)
  mean.diff.mle.L [t]  <- mean.mle.L[t] - mu[i]
```

```
  sd.diff.mle.L [t]  <- sd.mle.L[t] -sigma[j]
  #### Kaplan-Meier ####
  KM<-with(newdata,cenfit(newdata$obs,as.logical(newdata$cens)))
  mean.KM[t]<-mean(KM)[[1]]
  sd.KM[t]<-sd(KM)
  mean.diff.KM [t]<-mean.KM[t]-mu[i]
  sd.diff.KM[t]<-sd.KM[t]-sigma[j]
    #### rROS ####
  ROS<-with(newdata,cenros(newdata$obs,as.logical(newdata$cens)))
  mean.ROS [t] <- mean(ROS)
  sd.ROS [t]<- sd(ROS)
  mean.diff.ROS [t]<-mean.ROS[t]-mu[i]
  sd.diff.ROS[t]<-sd.ROS[t]-sigma[j]
 results2[t,] <-
c(mean.sub[t],mean.mle.L[t],mean.KM[t],mean.ROS[t],sd.sub[t],sd.mle.L[t],sd.KM[t],sd
.ROS[t])
}


bias.sub.mean <- mean(mean.diff.sub)
bias.sub.sd <- mean(sd.diff.sub)
var.sub.mean <- var(mean.diff.sub)
var.sub.sd <- var(sd.diff.sub)
MSE.sub.mean <- (bias.sub.mean^2)+var.sub.mean
MSE.sub.sd <- (bias.sub.sd ^2) +var.sub.sd


bias.LMLE.mean <- mean(mean.diff.mle.L)
bias.LMLE.sd <- mean(sd.diff.mle.L)
var.LMLE.mean <- var(mean.diff.mle.L)
var.LMLE.sd <- var(sd.diff.mle.L)
MSE.LMLE.mean <- (bias.LMLE.mean^2)+var.LMLE.mean
MSE.LMLE.sd <- (bias.LMLE.sd ^2) +var.LMLE.sd

bias.KM.mean<-mean(mean.diff.KM)
bias.KM.sd <-mean(sd.diff.KM)
var.KM.mean <- var(mean.diff.KM)
var.KM.sd<-var(sd.diff.KM)
MSE.KM.mean <- (bias.KM.mean^2)+var.KM.mean
MSE.KM.sd <- (bias.KM.sd ^2) +var.KM.sd
```

```
bias.ROS.mean<-mean(mean.diff.ROS)
bias.ROS.sd <-mean(sd.diff.ROS)
var.ROS.mean <- var(mean.diff.ROS)
var.ROS.sd<-var(sd.diff.ROS)
MSE.ROS.mean <- (bias.ROS.mean^2)+var.ROS.mean
MSE.ROS.sd <- (bias.ROS.sd ^2) +var.ROS.sd

B.results [k,] <-
c(bias.sub.mean,bias.LMLE.mean,bias.KM.mean,bias.ROS.mean,bias.sub.sd,bias.LMLE
.sd,bias.KM.sd,bias.ROS.sd)

results [k,] <-
c(mu[i],sigma[j],pc,MSE.sub.mean,MSE.LMLE.mean,MSE.KM.mean,MSE.ROS.mean,
MSE.sub.sd,MSE.LMLE.sd,MSE.KM.sd,MSE.ROS.sd)

                }
                }
```

The figures related to the supplementary material of article 2 (chapter 4) are illustrated below.

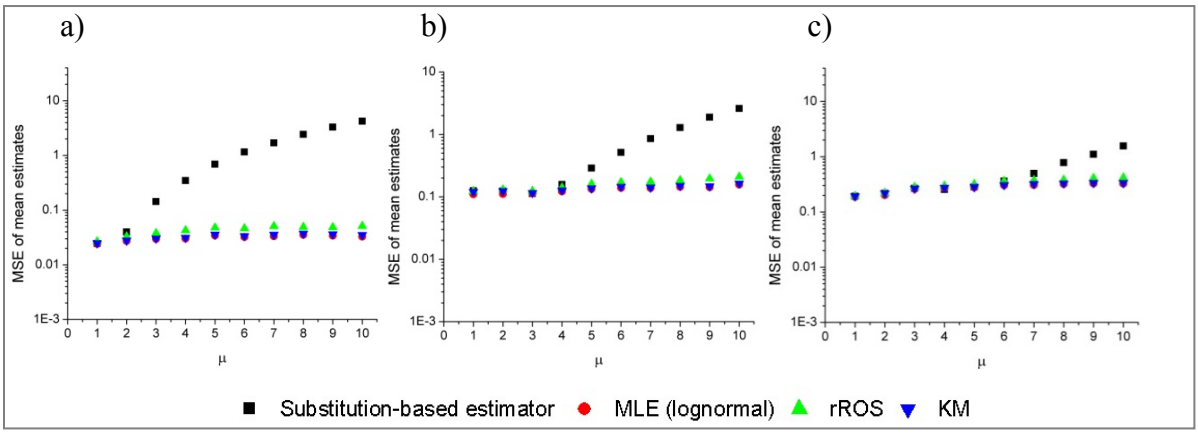


Figure-A IV-1 The MSEs of different methods in estimating the mean of lognormal distribution with μ=1,2,…,10 and a) σ=1.2, b) σ =2.6, c) σ=4

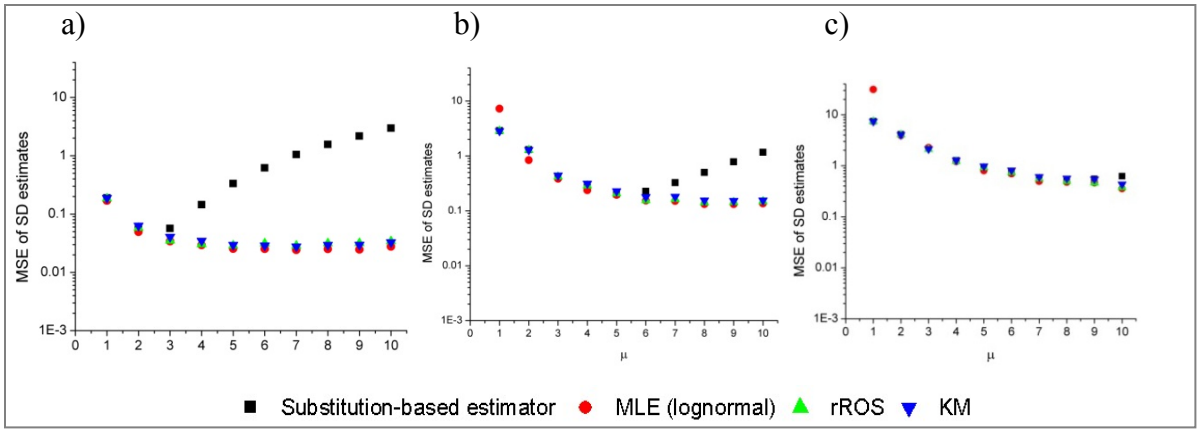


Figure-A IV-2 The MSEs of different methods in estimating the standard deviation of lognormal distribution with μ=1,2,…,10 and a) σ=1.2, b) σ =2.6, c) σ=4

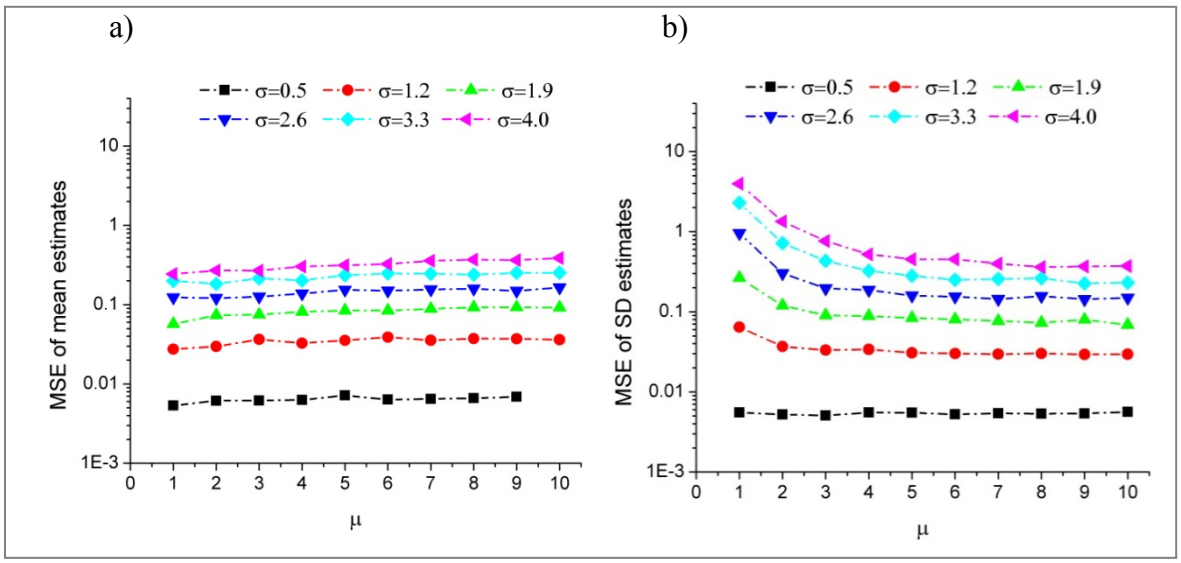


Figure-A IV-3 The MSEs of the rROS method in estimating a) the mean and b) standard deviation for different combinations of μ and σ of lognormal distribution

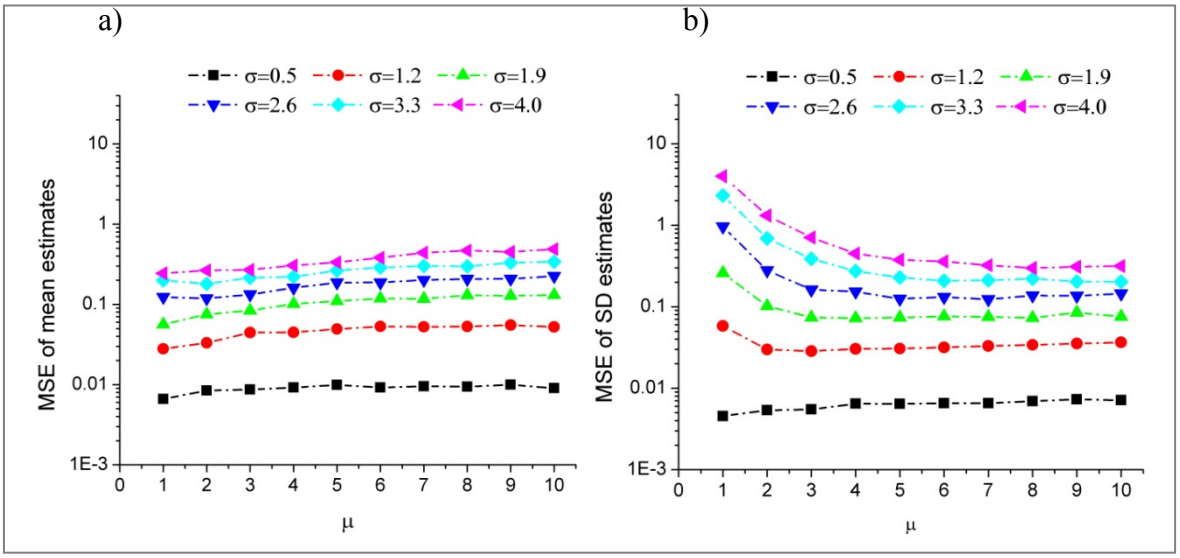


Figure-A IV-4 The MSEs of the KM method in estimating a) the mean and b) standard deviation for different combinations of μ and σ of lognormal distribution

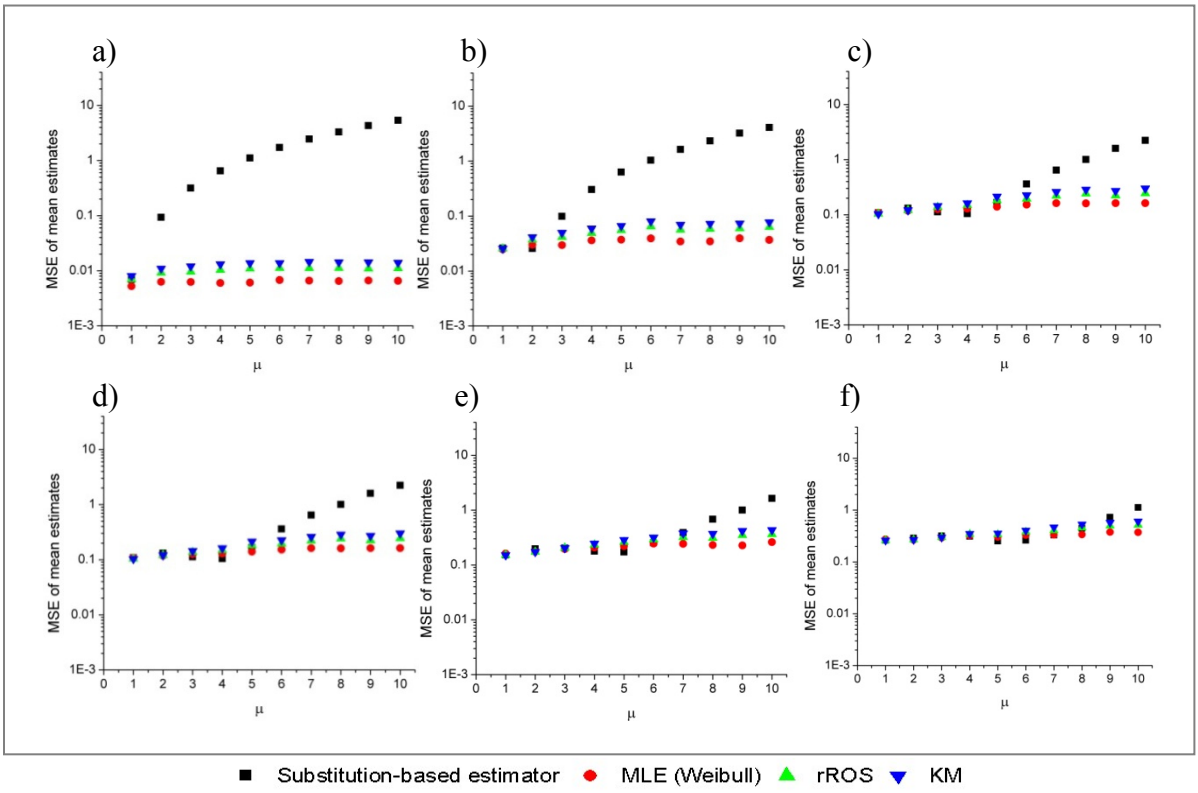


Figure-A IV-5 The MSEs of different methods in estimating the mean of Weibull distribution with μ=1,2,…,10 and a) σ=0.5, b) σ=1.2, c) σ=1.9, d) σ=2.6, e) σ=3.3, (f) σ= 4

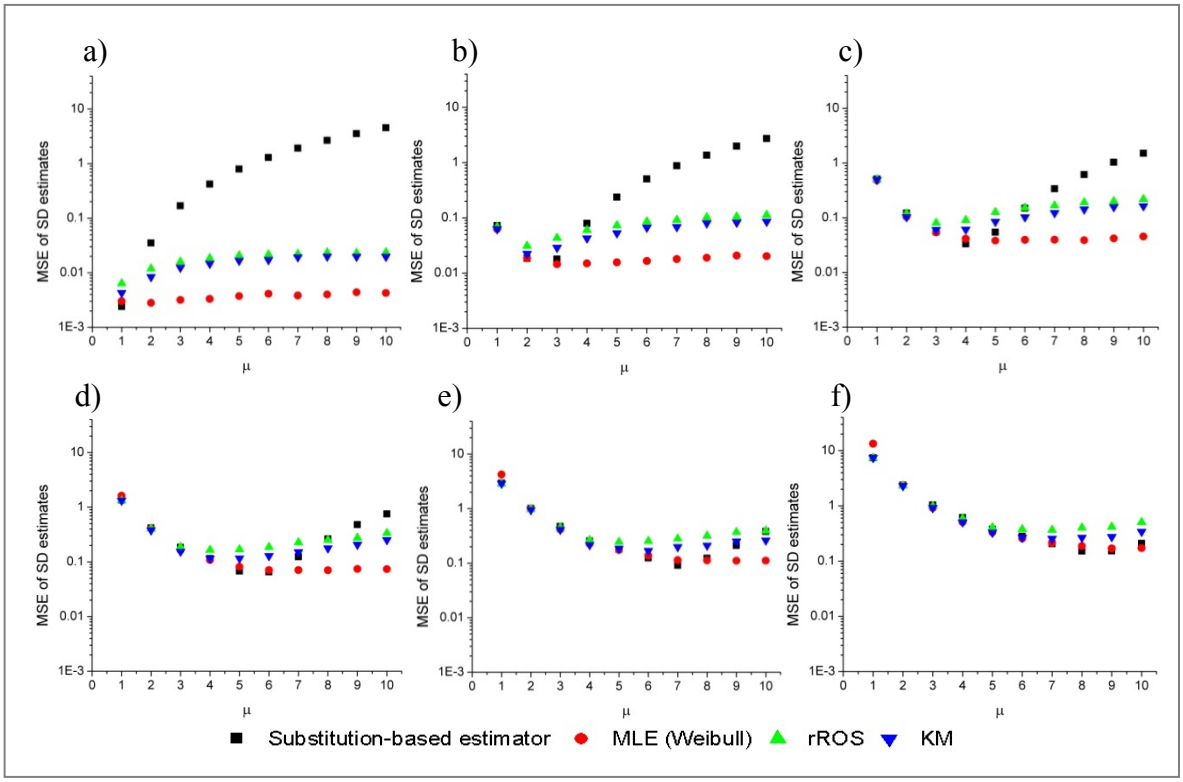


Figure-A IV-6 The MSEs of different methods in estimating the standard deviation of Weibull distribution with μ=1,2,…,10 and a) σ=0.5, b) σ=1.2, c) σ=1.9, d) σ=2.6, e) σ=3.3, and f) σ= 4

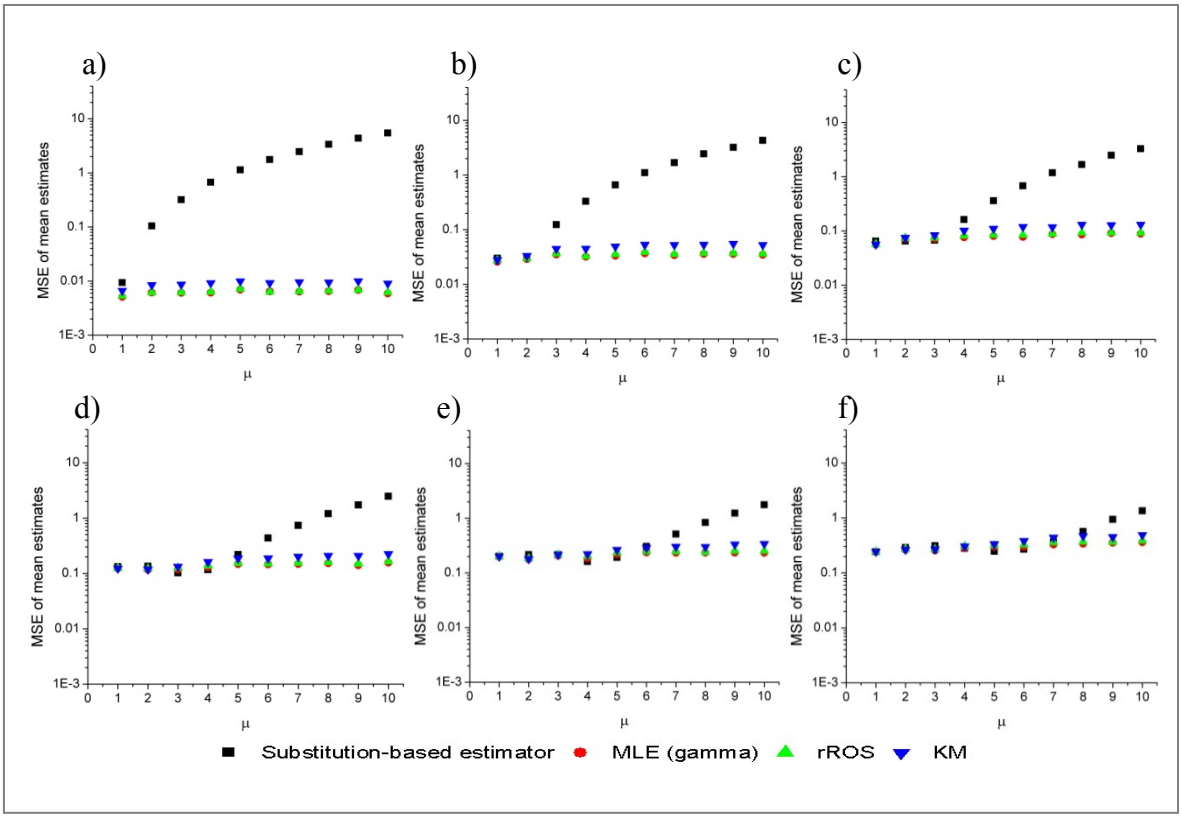


Figure-A IV-7 The MSEs of different methods in estimating the mean of gamma distribution with μ=1,2,…,10 and a) σ=0.5, b) σ=1.2, c) σ=1.9, d) σ=2.6, e) σ=3.3, and f) σ= 4

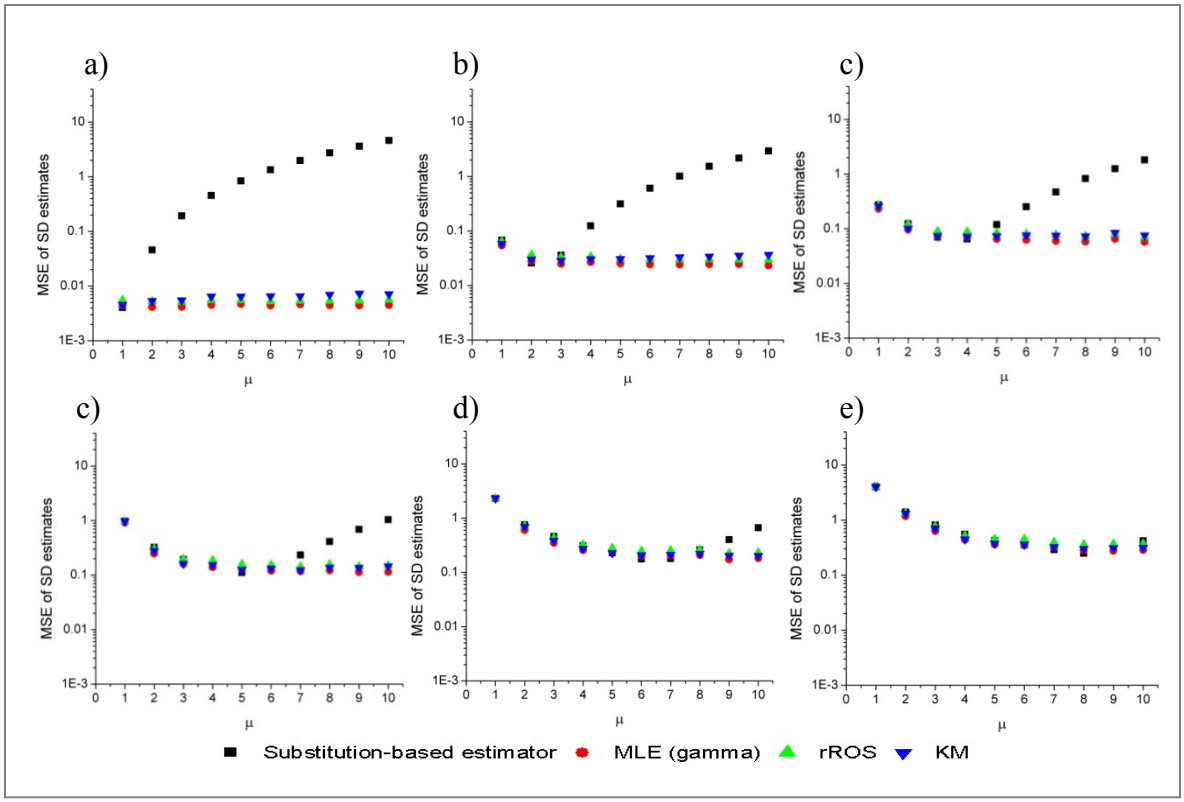
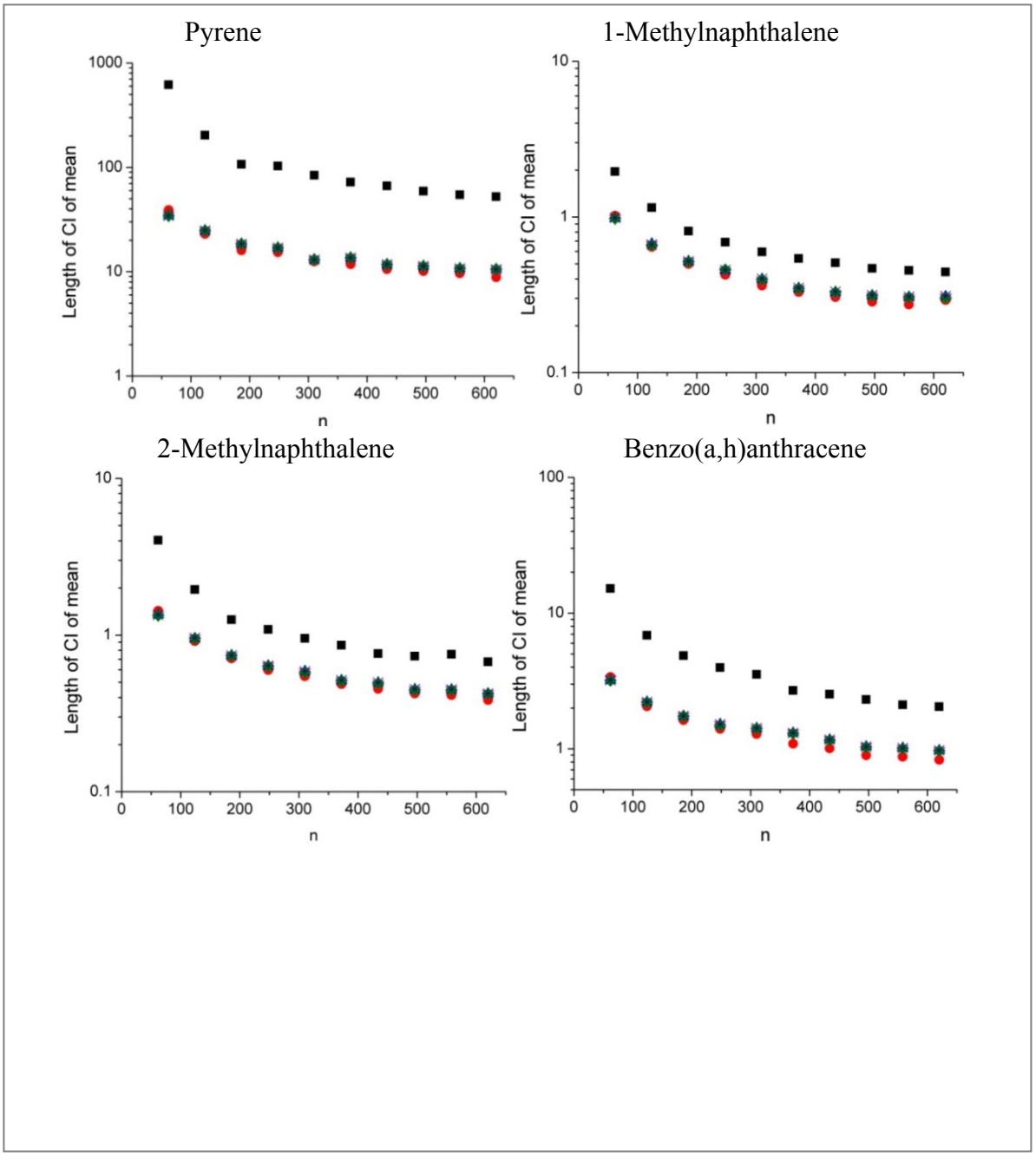


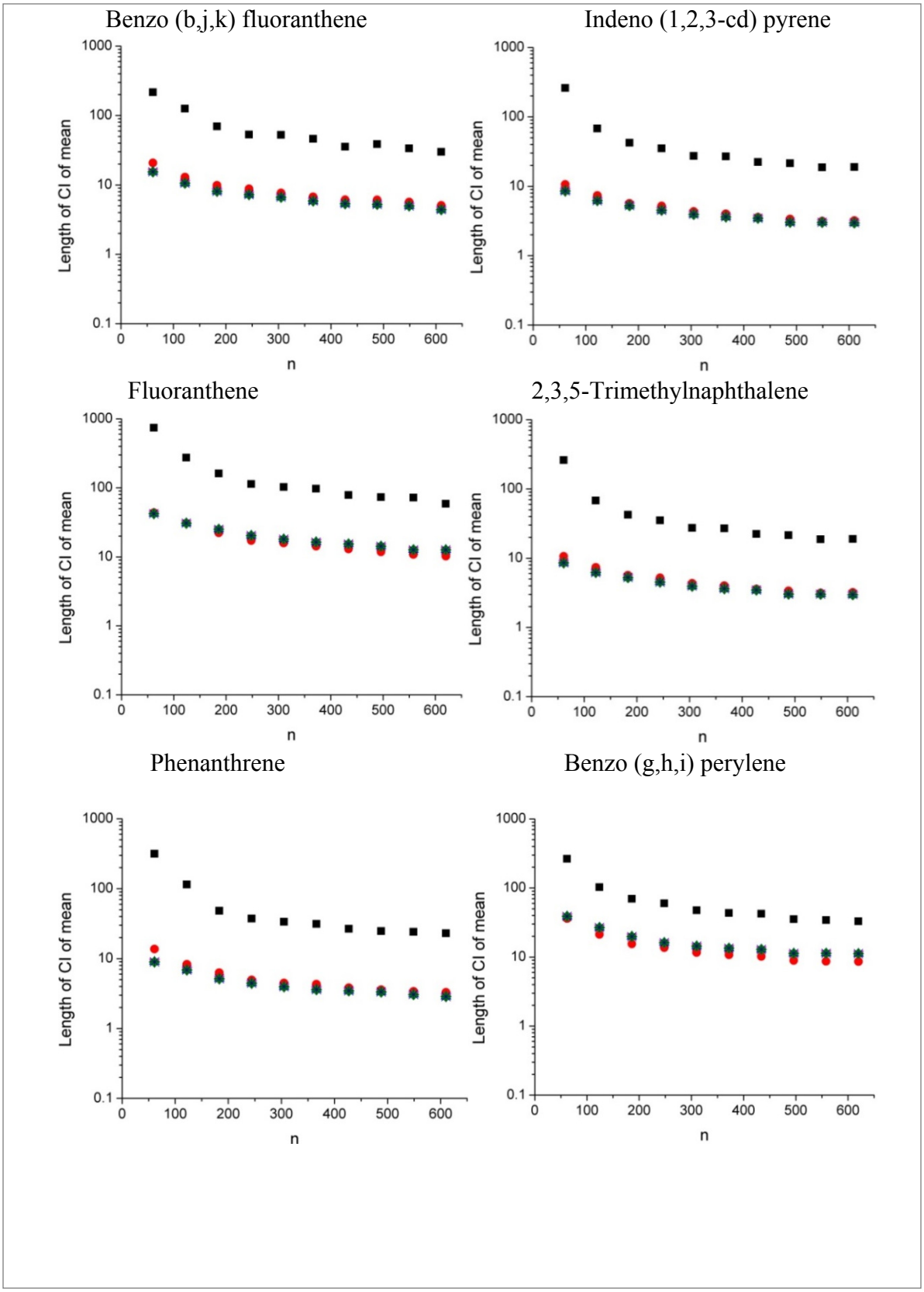
Figure-A IV-8 The MSEs of different methods in estimating the standard deviation of gamma distribution with μ=1,2,…,10 and a) σ=0.5, b) σ=1.2, c) σ=1.9, d) σ=2.6, e) σ=3.3, and f) σ= 4

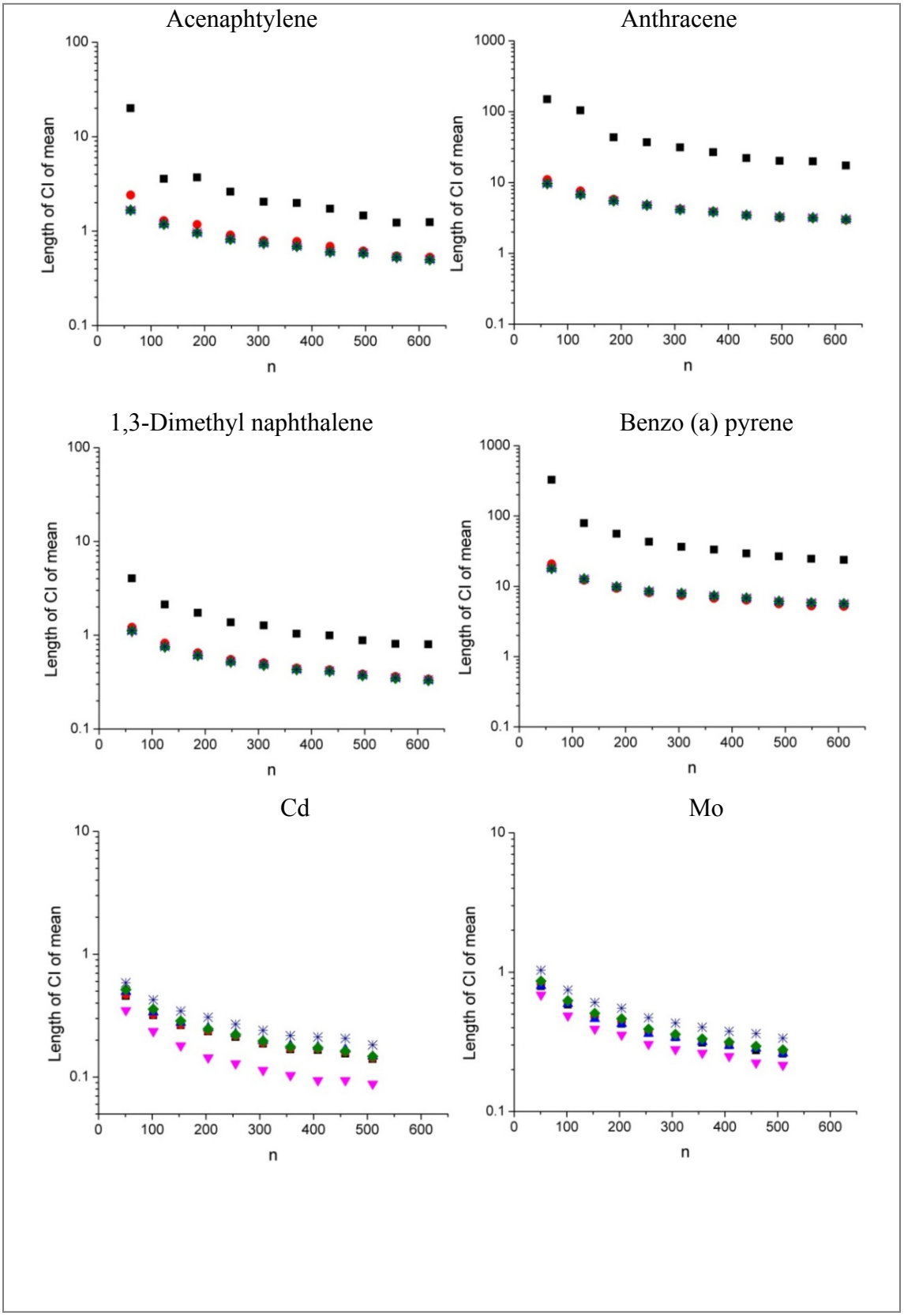
APPENDIX V

SUPPLEMENTARY MATERIAL OF ARTICLE 3

Figures related to the length of CI around the mean and standard deviation estimates are illustrated here.




Benzo (b,j,k) fluoranthene
Indeno (1,2,3-cd) pyrene
Fluoranthene
2,3,5-Trimethylnaphthalene
Phenanthrene
Benzo (g,h,i) perylene
Length of CI of mean
n
1000
100
10
1
0.1
0
100
200
300
400
500
600


Acenaphtylene
Anthracene
1,3-Dimethyl naphthalene
Benzo (a) pyrene
Cd
Mo
Length of CI of mean
n

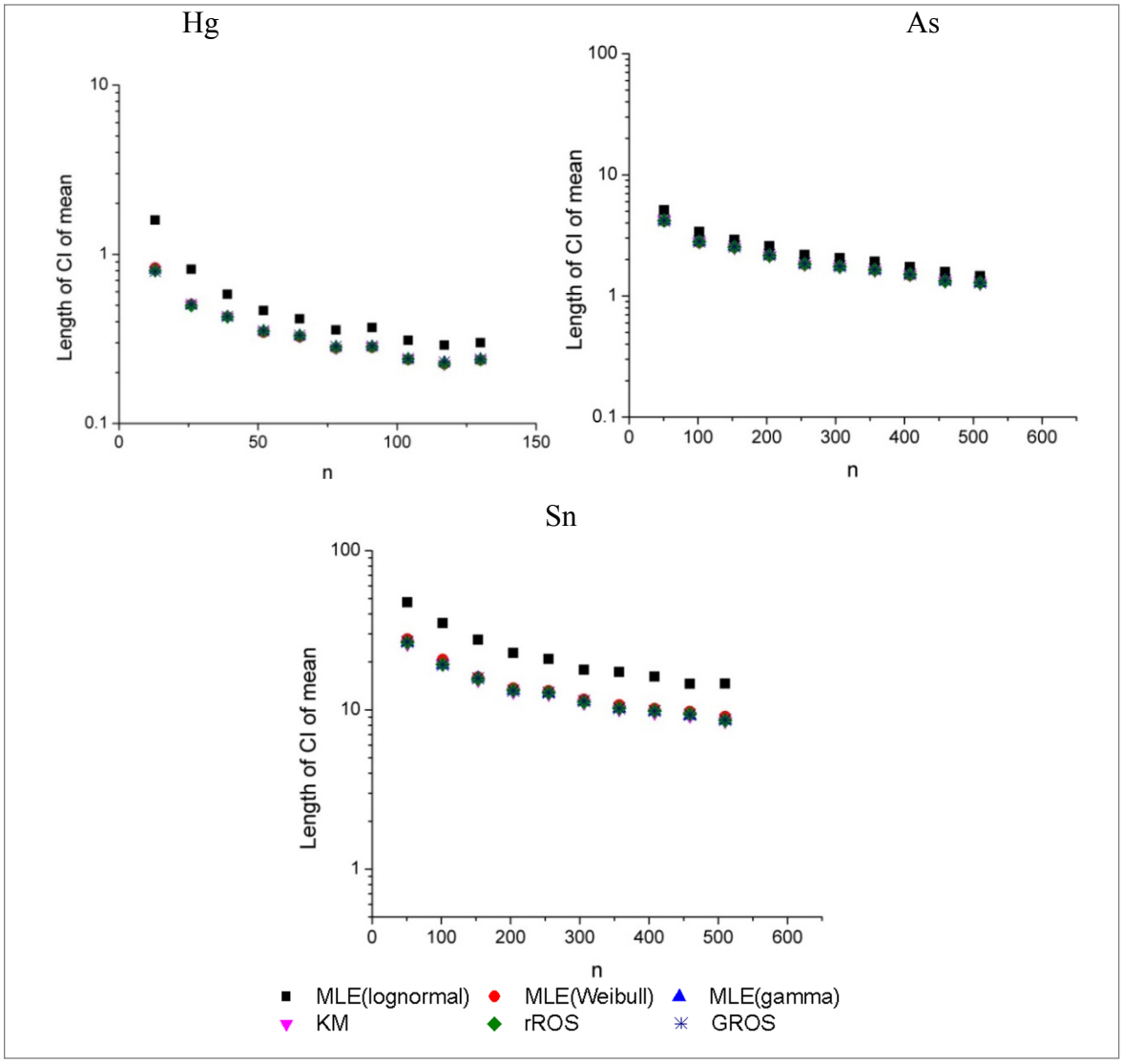
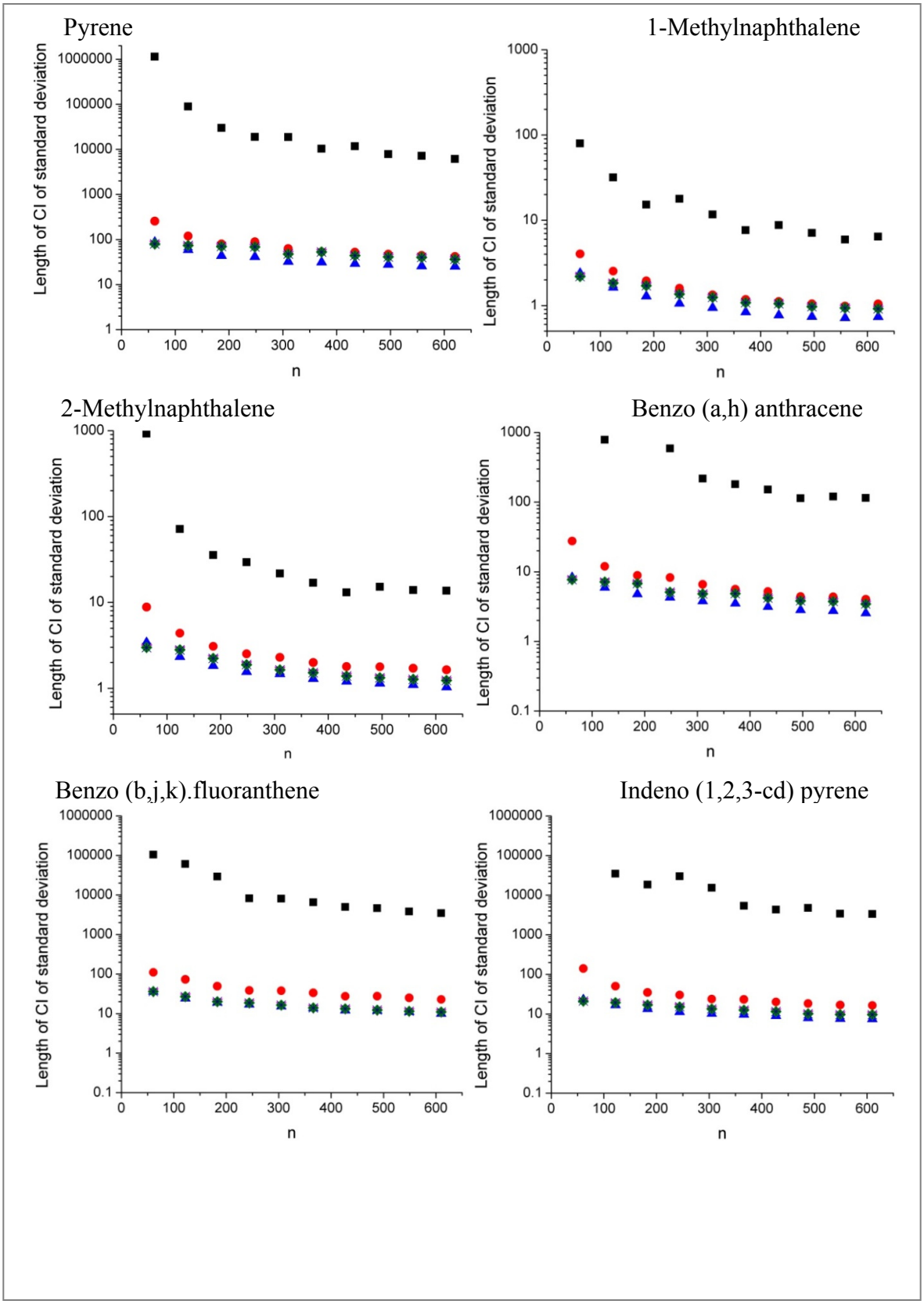
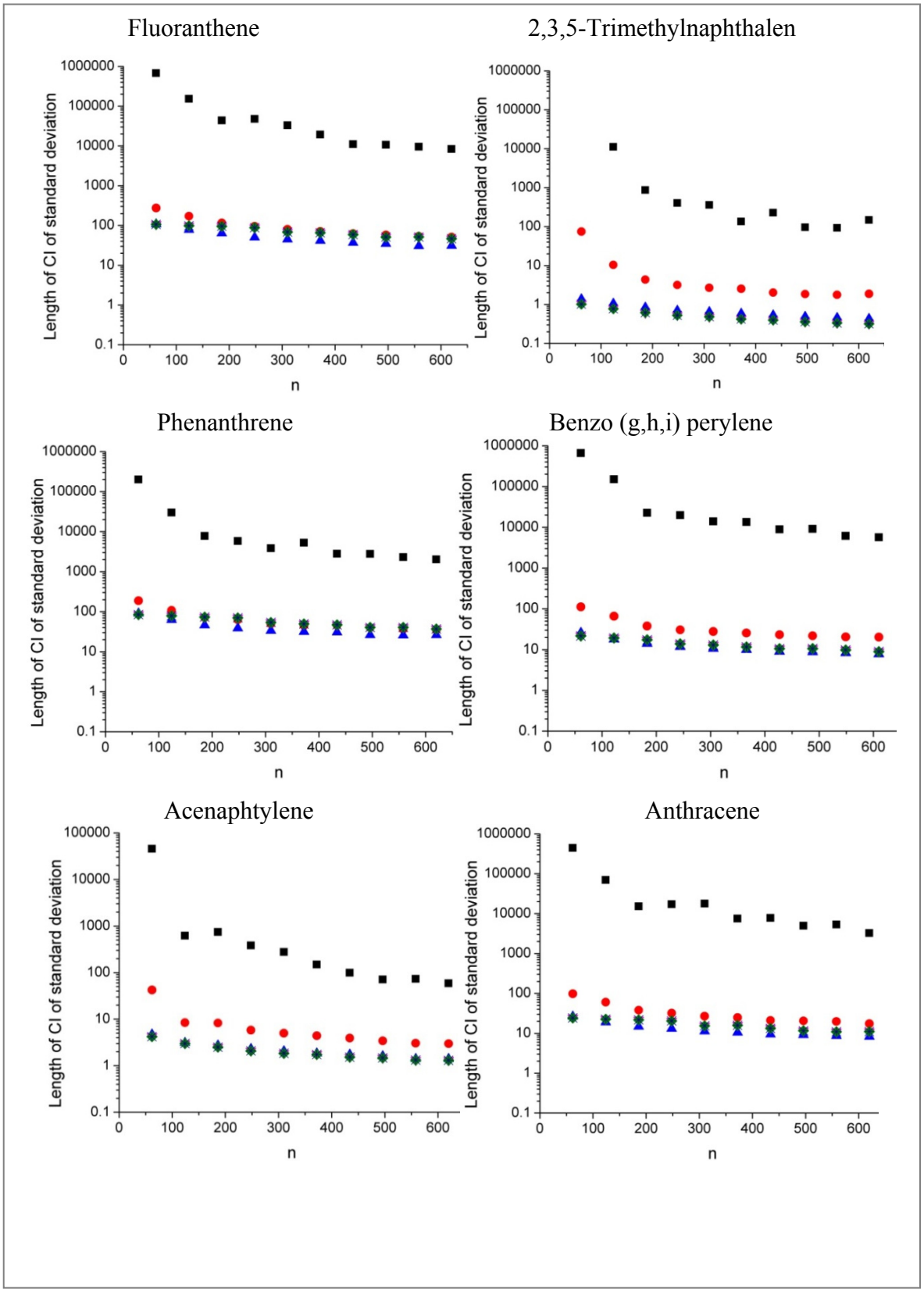
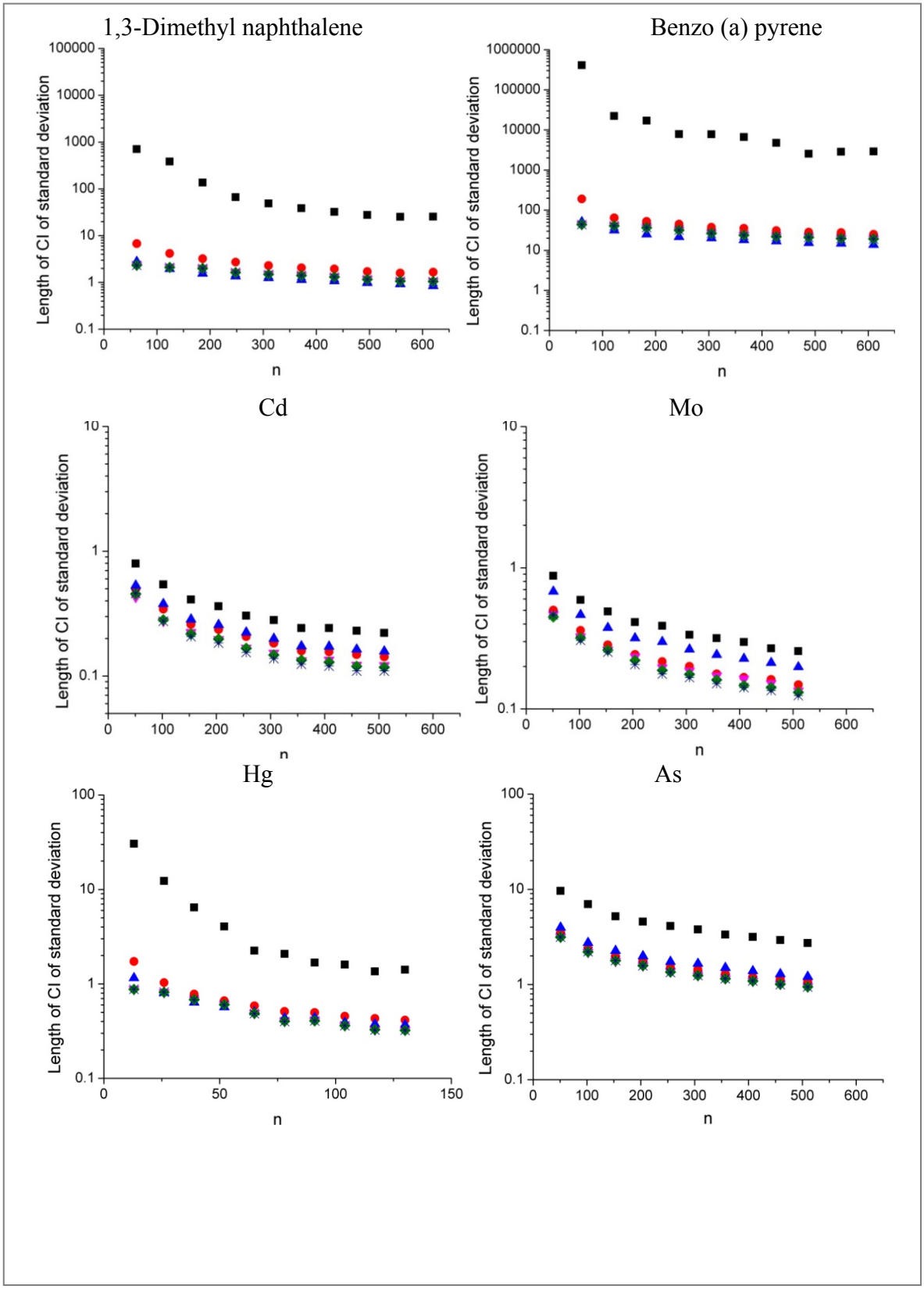


Figure-A V-3 Bootstrap confidence interval lengths around the mean estimate


Pyrene
1-Methylnaphthalene
2-Methylnaphthalene
Benzo (a,h) anthracene
Benzo (b,j,k).fluoranthene
Indeno (1,2,3-cd) pyrene
Length of CI of standard deviation
n


Fluoranthene
2,3,5-Trimethylnaphthalen
Phenanthrene
Benzo (g,h,i) perylene
Acenaphtylene
Anthracene
Length of CI of standard deviation
n


1,3-Dimethyl naphthalene
Length of CI of standard deviation
n
Benzo (a) pyrene
Length of CI of standard deviation
n
Cd
Length of CI of standard deviation
n
Mo
Length of CI of standard deviation
n
Hg
Length of CI of standard deviation
n
As
Length of CI of standard deviation
n

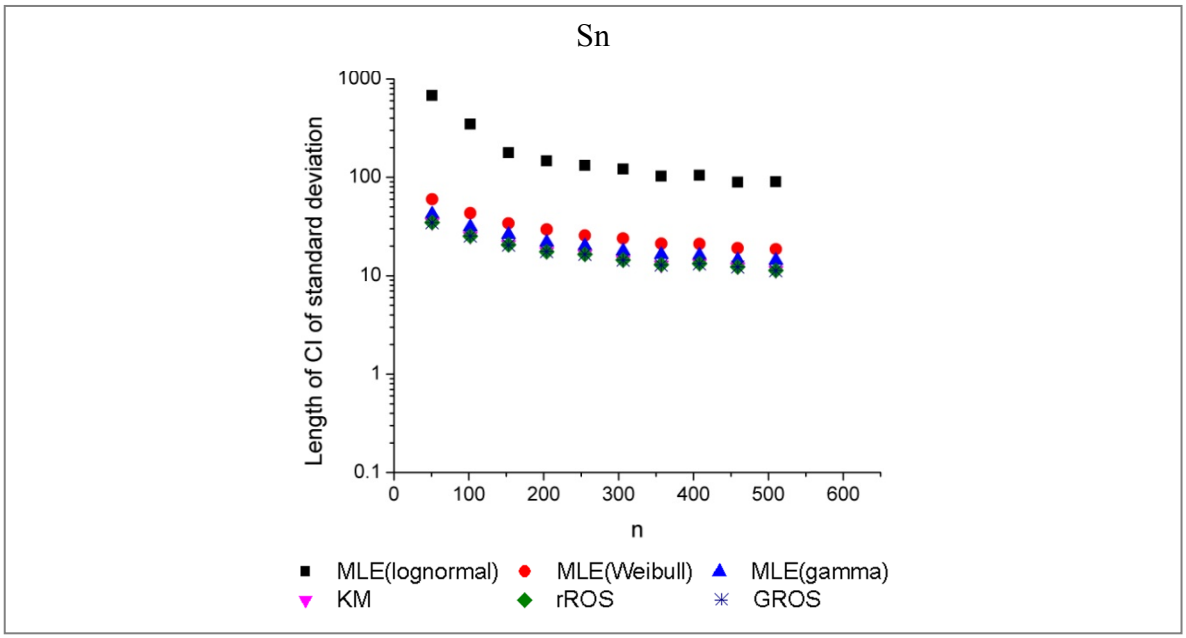
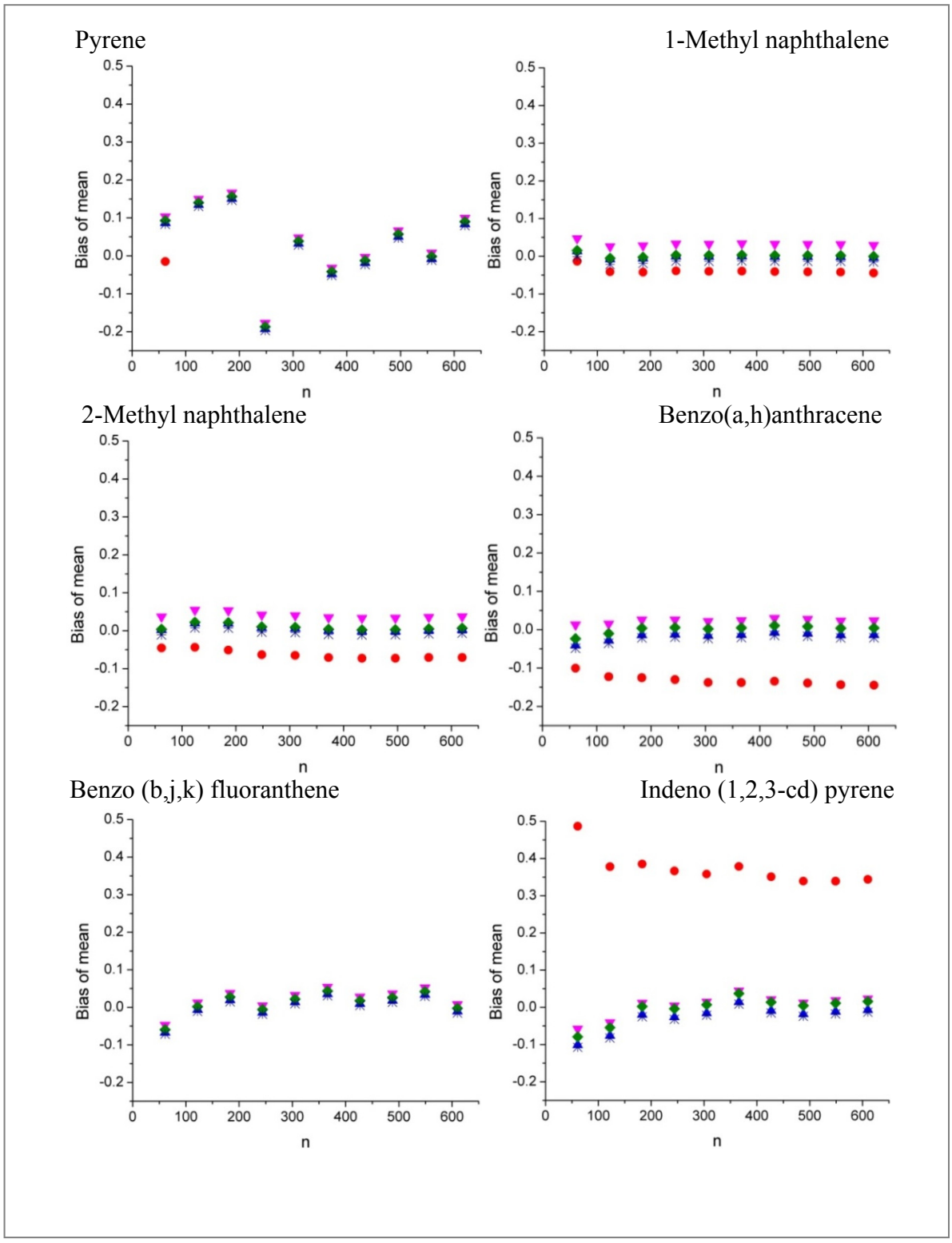
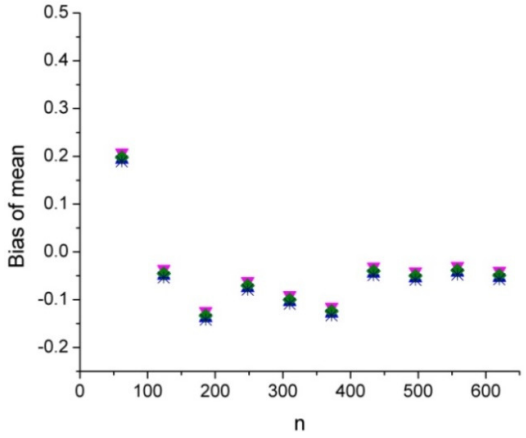


Figure-A V-4 Bootstrap confidence interval lengths around the standard deviation estimate

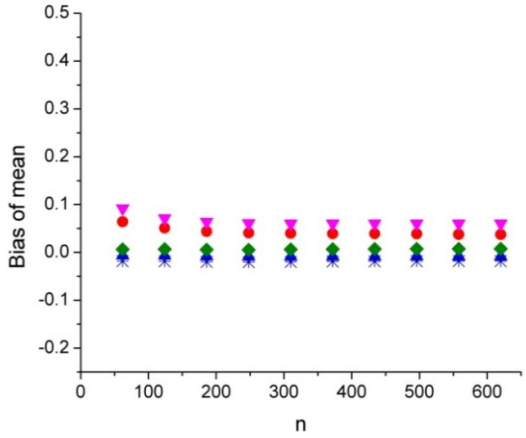
Figures related to the approximated bias of the mean and standard deviation estimates are illustrated below.



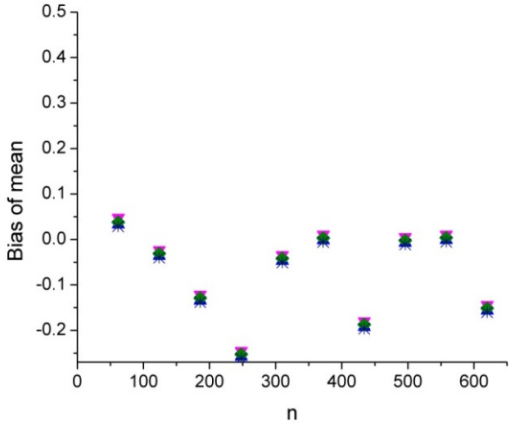
Fluoranthene



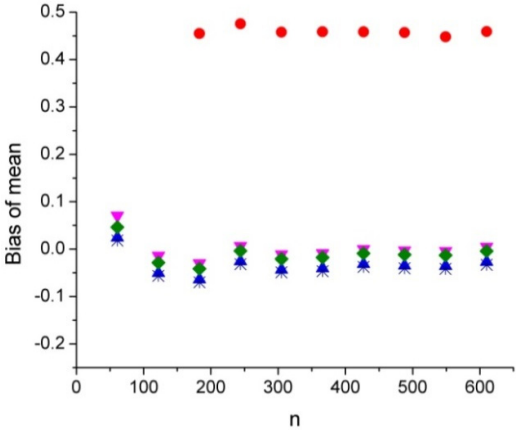
2,3,5-Trimethyl naphthalene



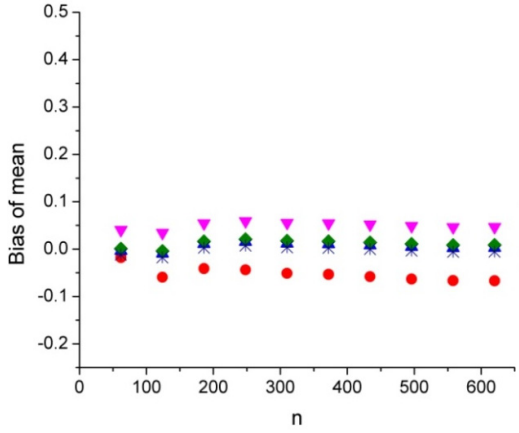
Phenanthrene



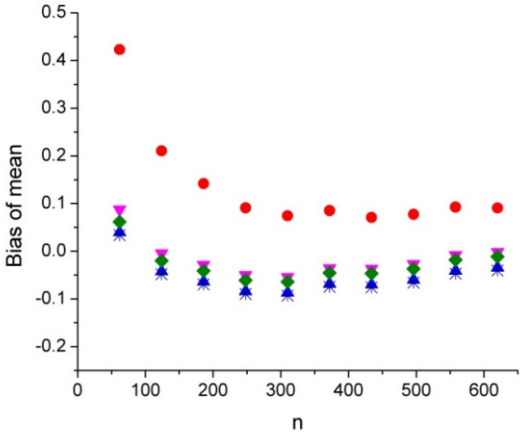
Benzo (g,h,i) perylene

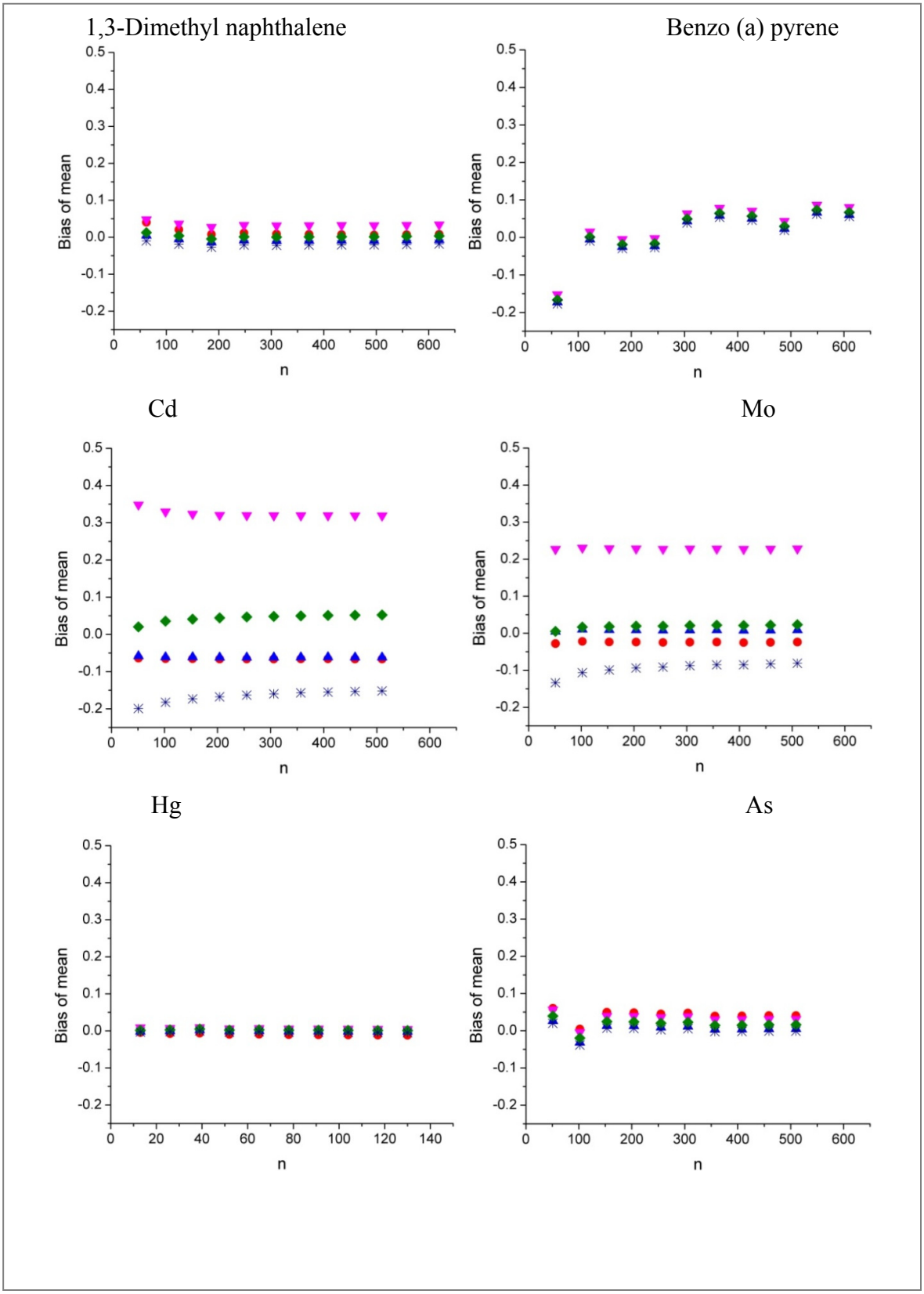


Acenaphtylene



Anthracene




1,3-Dimethyl naphthalene
Benzo (a) pyrene
Cd
Mo
Hg
As
Bias of mean
n

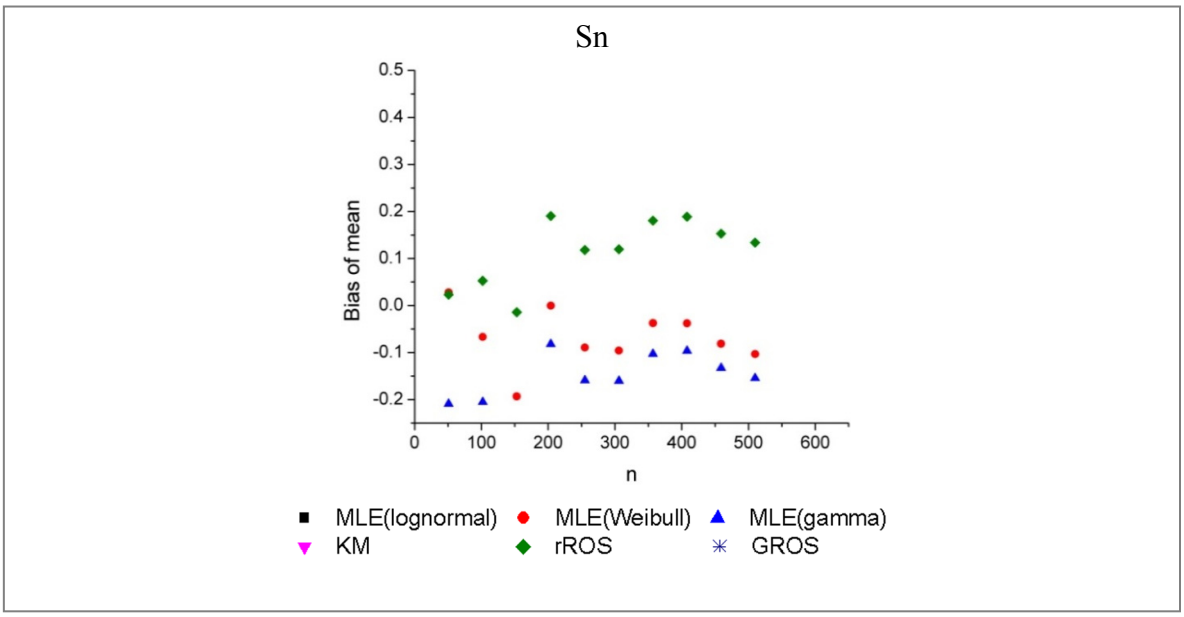
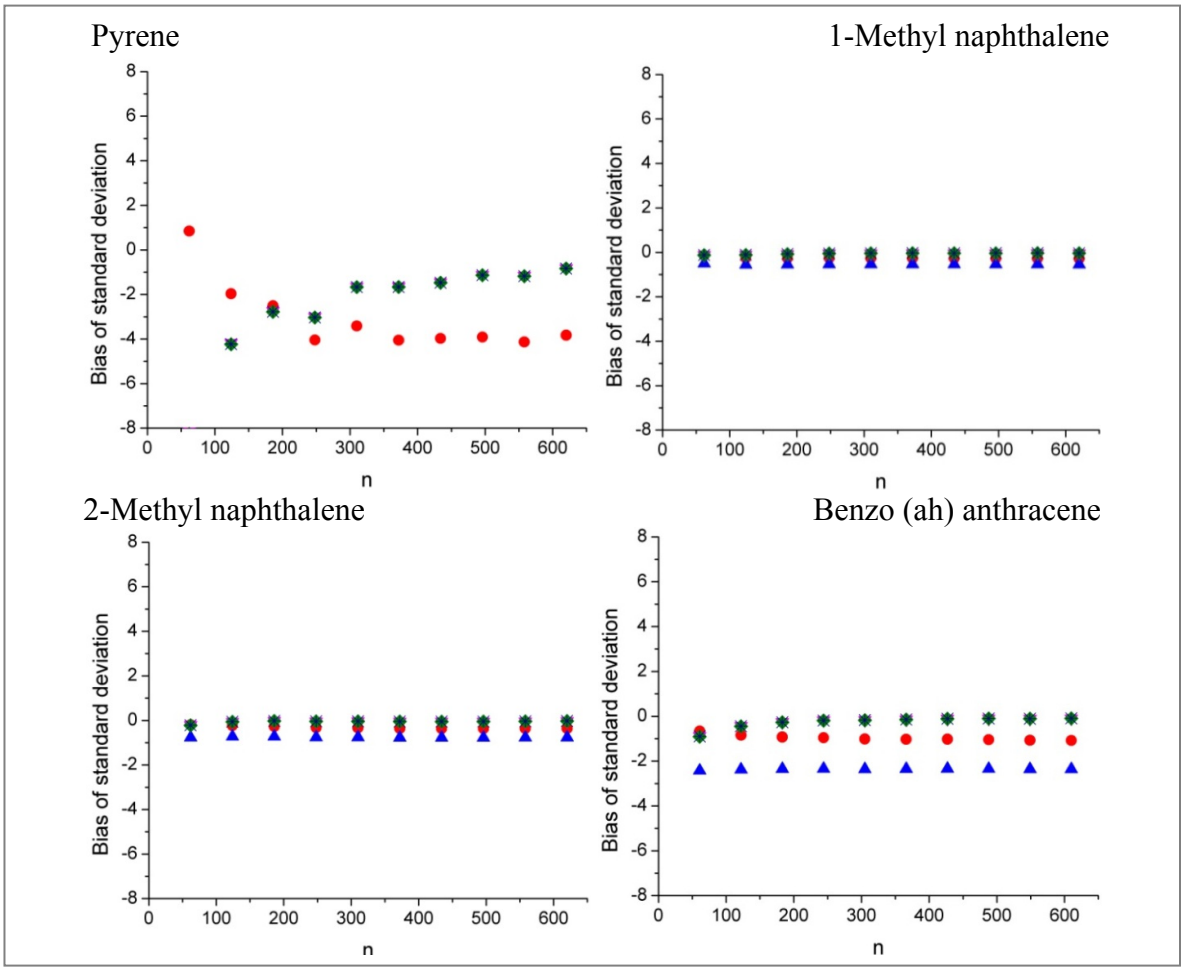
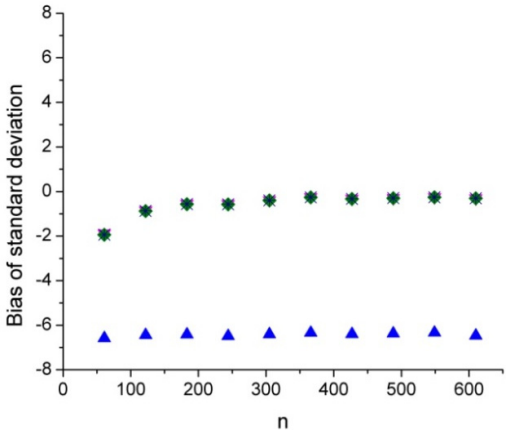


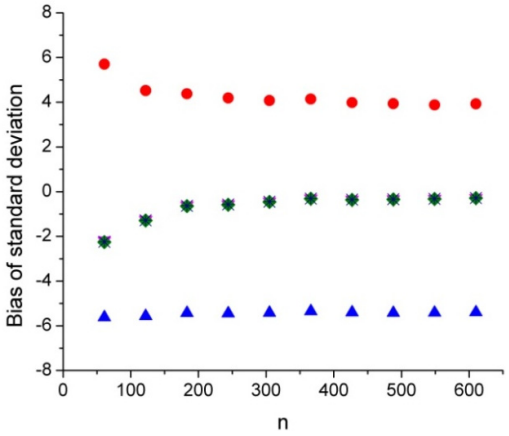
Figure-A V-5 Bootstrap approximated bias of the mean estimate



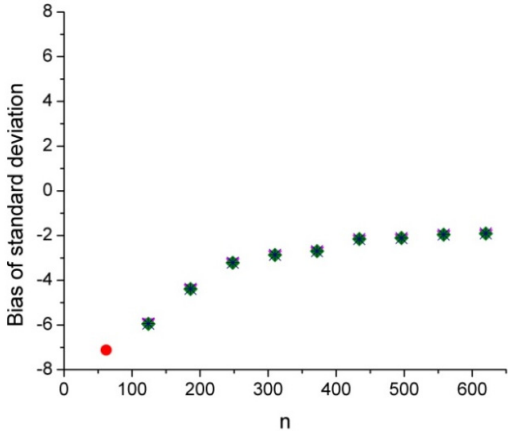
Benzo (b,j,k).fluoranthene



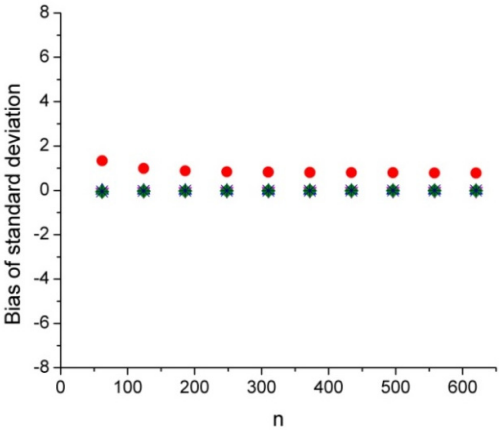
Indeno (1,2,3-cd) pyrene



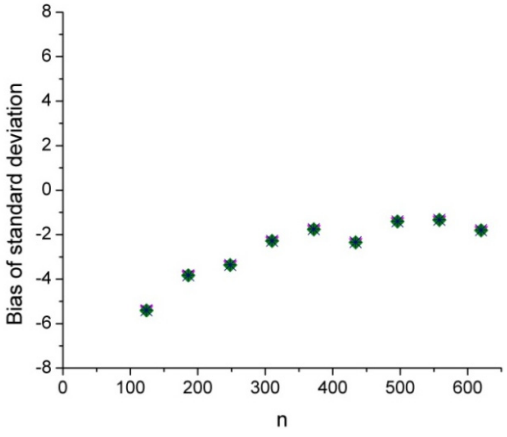
Fluoranthene



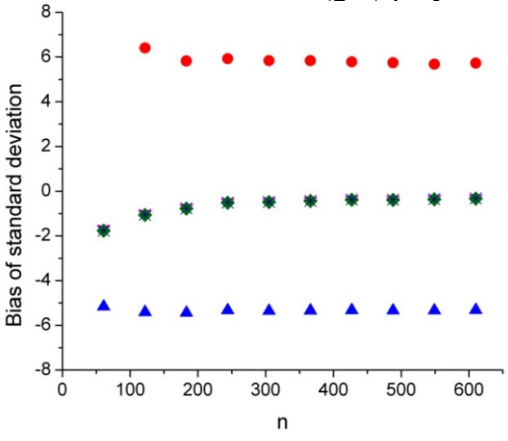
2,3,5-Trimethyl naphthalene

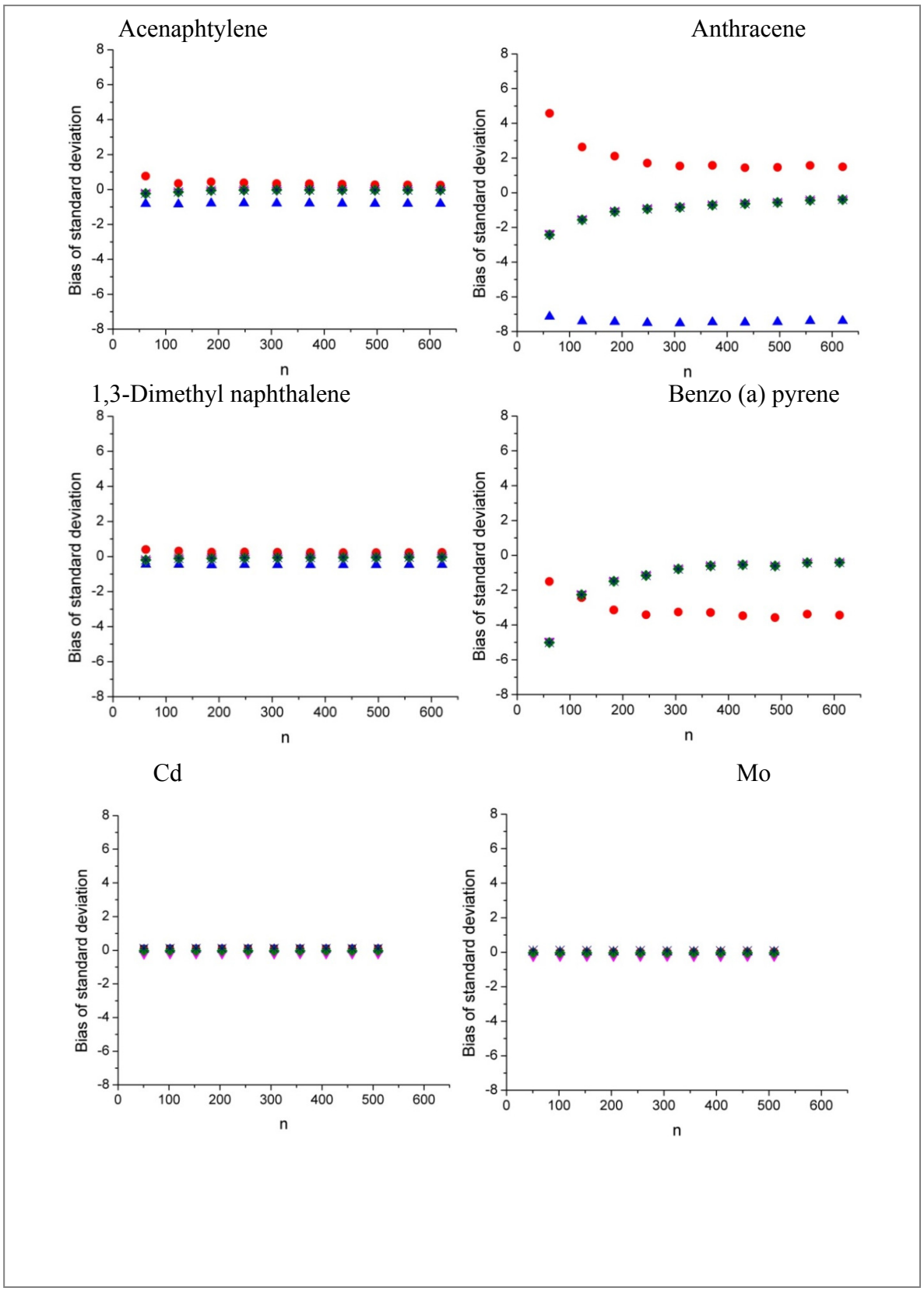


Phenanthrene



Benzo (ghi) perylene




Acenaphtylene
Anthracene
1,3-Dimethyl naphthalene
Benzo (a) pyrene
Cd
Mo
Bias of standard deviation
n

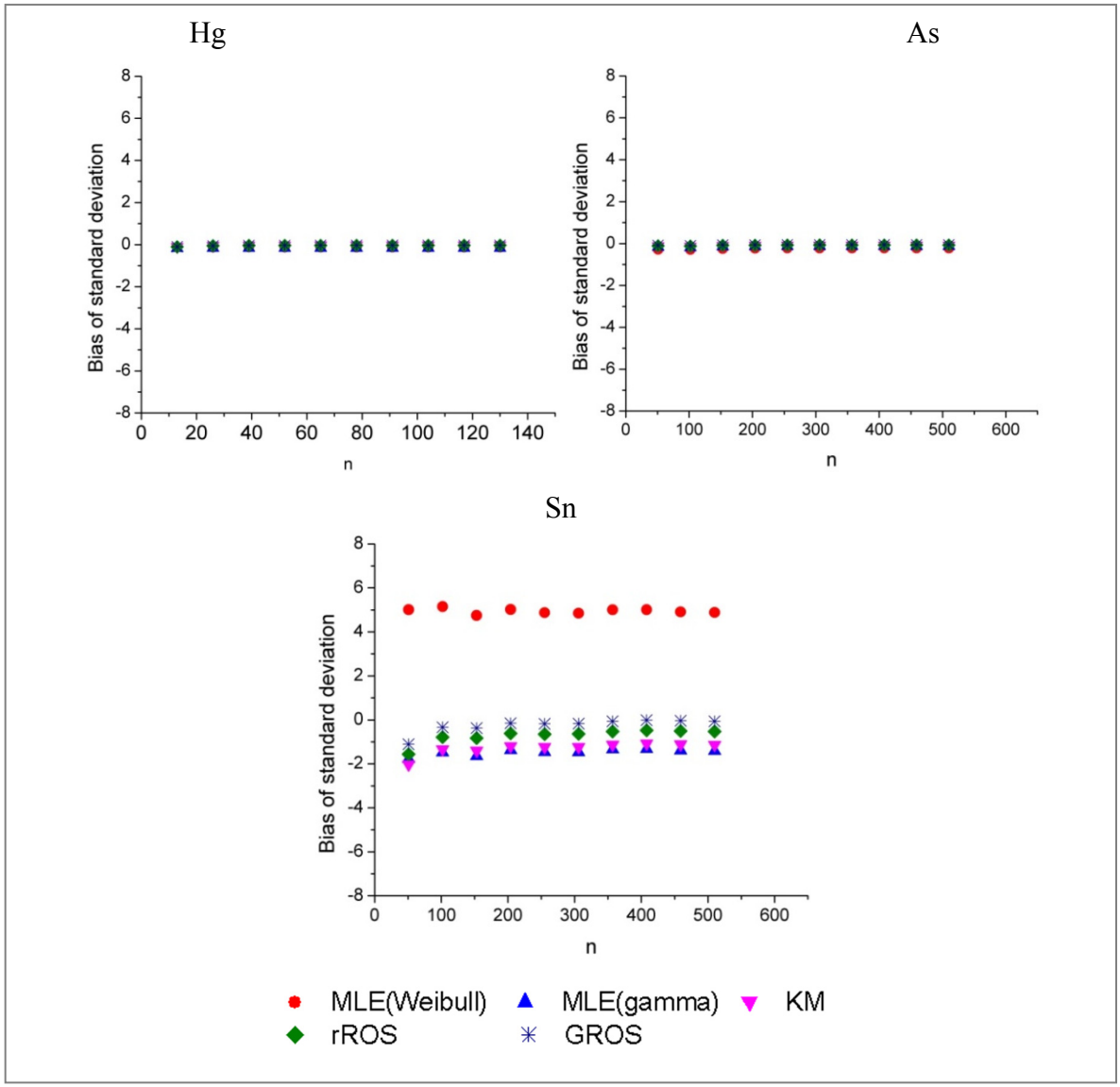


Figure-A V-6 Bootstrap approximated bias of the standard deviation estimate

APPENDIX VI

SUPPLEMENTARY MATERIAL OF ARTICLE 4

Table-A VI-1 The estimates of GM and 95% confidence intervals after fitting mixed-effects models to data after censored values are substituted[a]

Inorganic contaminants			
	Copper censoring=18%	Lead censoring=30%	Cadmium censoring=67%
Within borehole variance σ^2	0.93	1.71	0.30
Between borehole variance σ_b^2	0.19	0.72	0.62
IBC	0.17	0.29	0.68
GM (waste)	5.03 [4.73,5.33]	5.24 [4.80,5.69]	0.59 [0.31,0.88]
GM (crushed stones)	2.99 [2.64,3.35]	2.53 [2.02,3.05]	-0.57 [-0.83,-0.31]
GM (backfill)	3.95 [3.82,4.09]	3.48 [3.28,3.68]	-0.36 [-0.49,-0.23]
GM (natural soil)	3.32 [3.08,3.56]	2.35 [1.99,2.70]	-0.62 [-0.81,-0.43]
Organic contaminants			
	Benzo(a)pyrene censoring=51%	Naphthalene censoring=57%	
Within borehole variance σ^2	1.55	1.45	
Between borehole variance σ_b^2	0.64	0.50	
IBC	0.29	0.26	
GM (waste)	-1.29 [-1.65,-0.92]	-1.47 [-1.82,-1.12]	
GM (crushed stones)	-2.77 [-3.32,-2.23]	-2.54 [-3.06,-2.02]	
GM (backfill)	-1.50 [-1.69,-1.31]	-1.88 [-2.06,-1.70]	
GM (natural soil)	-2.35 [-2.60,-2.09]	-2.44 [-2.68,-2.20]	

[a] The material is considered as fixed effects in the model

LIST OF BIBLIOGRAPHICAL REFERENCES

Aboueissa, A. A & Stoline M.R. (2004). Estimation of the mean and standard deviation from normally distributed singly-censored samples. *Environmetrics, 15* (7), 659-673.

Annan, S. Y., Liu, P. & Zhang Y. (2009). Comparison of the Kaplan-Meier, Maximum Likelihood, and ROS estimators for left-censored data using simulation studies.

Antweiler, R. C. (2015). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets. II. group comparisons. *Environmental science & technology, 49* (22), 13439-13446.

Antweiler, R.C. & Taylor, H.E. (2008). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. *Environmental science & technology, 42* (10), 3732-3738.

Austin, P. C., Escobar, M. & Kopec J.A. (2000). The use of the Tobit model for analyzing measures of health status. *Qual Life Res Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Res, 9* (8), 901-910.

Babamoradi, H., van den Berg F. & Rinnan, Å. (2013). Bootstrap based confidence limits in principal component analysis—A case study. *Chemometrics and Intelligent Laboratory Systems, 120*, 97-105.

Baccarelli, A., Pfeiffer, R., Consonni, D., Pesatori, A.C., Bonzini, M., Patterson,D.G., … Landi M.T. (2005). Handling of dioxin measurement data in the presence of non-detectable values: overview of available methods and their application in the Seveso chloracne study. *Chemosphere, 60* (7), 898-906.

Bakke, B., Ulvestad, B.,Thomassen, Y., Woldbæk, T. & Ellingsen, D.G. (2014). Characterization of occupational exposure to air contaminants in modern tunnelling operations. *Annals of Occupational Hygiene, 58* (7), 818-829.

Barghi, M., Choi, S. D., Kwon, H. O., Lee, Y. S. & Chang, Y. S. (2016). Influence of non-detect data-handling on toxic equivalency quantities of PCDD/Fs and dioxin-like PCBs: A case study of major fish species purchased in Korea. *Environ. Pollut. Environmental Pollution, 214*, 532-538.

Barr, D.J., Levy, R., Scheepers, C. & Tily H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68* (3), 255-278.

Bogner, C., Gaul, D., Kolb, A., Schmiedinger, I. & Huwe, B. (2010). Investigating flow mechanisms in a forest soil by mixed-effects modelling. *European journal of soil science, 61* (6), 1079-1090.

Boudreault, J.-P., Dubé, J.-S, Sona, M. & Hardy, É. (2012). Analysis of procedures for sampling contaminated soil using Gy's Sampling Theory and Practice. *STOTEN Science of the Total Environment,* 425, 199-207.

Buckley, J. & James, I. (1979). Linear Regression with Censored Data. *Biometrika, 66* (3), 429-436.

Burnham, K. P., & Anderson D.R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Canadian Council of Ministers of the Environment. (2016). *Guidance manual for environmental site characterization in support of environmental and human health risk assessment*. CCME. Retrieved from http://www.ccme.ca/en/resources/contaminated_site_management/assessment.html. on June 08, 2016.

Caudill, S.P., Wong, L.-Y., Turner, W. E., Lee, R., Henderson, A. & Patterson, D.G. (2007). Percentile estimation using variable censored data. *Chemosphere, 68* (1), 169-180.

Chen, M., Qin, X., Zeng, G. & Li, J. (2016). Impacts of human activity modes and climate on heavy metal "spread" in groundwater are biased. *Chemosphere, 152*, 439-445.

Clarke, J. U. (1998). Evaluation of censored data methods to allow statistical comparisons among very small samples with below detection limit observations. *Environmental science & technology, 32* (1), 177-183.

Clinical and Laboratory Standards Institute. (2004). *Protocols for determination of limits of detection and limits of quantitation; approved guideline*. Wayne, PA USA.

Cohen, A. C. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics, 1* (3), 217-237.

Cohen, A. C. (1961). Tables for maximum likelihood estimates: singly truncated and singly censored samples. *Technometrics, 3* (4), 535-541.

Coronel-Brizio, H.F. & Hernandez-Montoya, A.R. (2010). The Anderson–Darling test of fit for the power-law distribution from left-censored samples. *Physica A: Statistical Mechanics and its Applications, 389* (17), 3508-3515.

D'Agostino, R.B. & Stephens, M.A. (1986). *Goodness-of-fit techniques*. New York: M. Dekker.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge; New York, NY, USA: Cambridge University Press.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Dessau. (2009). *Évaluation de la représentativité et de la précision statistique des données de caractérisation des sols. Site Turcot*. Technical report.

Dien, N.T., Hirai, Y., Miyazaki, T. & Sakai, S.-I. (2016). Factors influencing atmospheric concentrations of polybrominated diphenyl ethers in Japan. *Chemosphere,* 144, 2073-2080.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review/Revue Internationale de Statistique*, 67-82.

Dubé, J.-S., Boudreault, J.-P., Bost, R., Sona, M., Duhaime F. & Éthier, Y. (2015). Representativeness of laboratory sampling procedures for the analysis of trace metals in soil. *Environ Sci Pollut Res Environmental Science and Pollution Research, 22 (*15), 11862-11876.

Eastoe, E.F., Halsall, C.J., Heffernan, J. E. & Hung, H. (2006). A statistical comparison of survival and replacement analyses for the use of censored data in a contaminant air database: A case study from the Canadian Arctic. *Atmospheric Environment, 40* (34), 6528-6540.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *statscie Statistical Science, 1* (1), 54-75.

Efron, B. (1981). Censored Data and the Bootstrap. *jamerstatasso Journal of the American Statistical Association, 76* (374), 312-319.

Efron, B. & Tibshirani, R. (1994). *An introduction to the bootstrap*. New York: Chapman & Hall.

El-Shaarawi, A. H. & Piegorsch, W.W. (2002). *Encyclopedia of environmetrics*. Chichester; New York: Wiley.

El-Shaarawi, A.H. & Esterby, S.R. (1992). Replacement of censored observations by a constant: an evaluation. *Water Research, 26* (6), 835-844.

El-Shaarawi, A.H. & Naderi, A. (1991). Statistical inference from multiply censored environmental data. In *Statistical Methods for the Environmental Sciences*. 261-269. Springer.

El-Shaarawi, A.H. (1989). Inferences about the mean from censored water quality data. *Water Resources Research, 25* (4), 685-690.

European Food Safety Authority. (2010). *Management of left-censored data in dietary exposure assessment of chemical substances*. 8.

Farnham, I.M., Singh, A.K., Stetzenbach, K.J. & Johannesson, K.H. (2002). Treatment of nondetects in multivariate analysis of groundwater geochemistry data. *Chemometrics and Intelligent Laboratory Systems, 60* (1), 265-281.

Feigelson, E.D. & Babu, G. J. (2012). *Modern statistical methods for astronomy : with R applications*. Cambridge; New York: Cambridge University Press.

Finkelstein, M.M. (2008). Asbestos fibre concentrations in the lungs of brake workers: another look. *Annals of Occupational Hygiene, 52* (6), 455-461.

Finkelstein, M.M. & Verma, D.K. (2001). Exposure estimation in the presence of nondetectable values: Another look. *AIHAJ - American Industrial Hygiene Association, 62* (2), 195-198.

Flynn, M. R. (2010). Analysis of censored exposure data by constrained maximization of the Shapiro–Wilk W statistic. *Annals of Occupational Hygiene, 54* (3), 263-271.

Frey, H. C. & Zhao, Y. (2004). Quantification of variability and uncertainty for air toxic emission inventories with censored emission factor data. *Environmental science & technology, 38* (22), 6094-100.

Frome, E. L. & Wambach, P.F. (2005). Statistical methods and software for the analysis of occupational exposure data with non-detectable values. United States Department of Energy.

Ganser, G.H. & Hewett, P. (2010). An accurate substitution method for analyzing censored data. *Journal of occupational and environmental hygiene, 7* (4), 233-244.

Gardner, K.K. & Vogel, R.M. (2005). Predicting ground water nitrate concentration from land use. *Ground water, 43* (3), 343-352.

Gardner, M. (2012). Improving the interpretation of 'less than'values in environmental monitoring. *Water and Environment Journal, 26* (2), 285-290.

Gbaguidi-Haore, H., Roussel, S., Reboux, G., Dalphinand, J.-C. & Piarroux, R. (2009). Multilevel analysis of the impact of environmental factors and agricultural practices on the concentration in hay of microorganisms responsible for farmer's lung disease . *Annals of Agricultural and Environmental Medicine, 16* (2), 219-225.

Gerlach, R.W., Dobb, D. E., Raab, G. A. & Nocerino, J. M. (2002). Gy sampling theory in environmental studies. 1. Assessing soil splitting protocols. *CEM Journal of Chemometrics, 16* (7), 321-328.

Gilbert, R. O. (1987). *Statistical methods for environmental pollution monitoring*. New York: Van Nostrand Reinhold Co.

Gillespie, B. W., Chen, Q., Reichert, H., Franzblau, A., Hedgeman, E., Lepkowski, J., … Garabrant, D. H. (2010). Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. *Epidemiology (Cambridge, Mass.), 21*, 64-70.

Gilliom, R.J. & Helsel, D. (1986). Estimation of distributional parameters for censored trace level water quality data: 1. Estimation techniques. *Water Resources Research, 22* (2), 135-146.

Giri, S., Nejadhashemi, A.P., Zhang, Z. & Woznicki, S.A. (2015). Integrating statistical and hydrological models to identify implementation sites for agricultural conservation practices. *Environmental Modelling & Software, 72*, 327-340.

Groupe Qualitas inc. (2010). *Étude géotechnique et caractérisation environnementale complémentaire. Résidences phase IV. Rues De La Montagne et William, Montréal (Québec)*. 16039-GE2. Longueuil, QC: Groupe Qualitas inc.

Guo, S. (2005). Analyzing grouped data with hierarchical linear modeling. *Children and Youth Services Review, 27* (6), 637-652.

Gurka, M.J., Edwards, L.J., Muller, K.E. & Kupper, L.L. (2006). Extending the Box–Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169* (2), 273-288.

Haas, C. N. & Scheff, P.A. (1990). Estimation of averages in truncated samples. *Environ. Sci. Technol. Environmental Science & Technology, 24* (6), 912-919.

He, J. (2013). Mixture model based multivariate statistical analysis of multiply censored environmental data. *Advances in Water Resources Advances in Water Resources,* 59 (2), 15-24.

Health Canada. (2010). *Federal contaminated site risk assessment in Canada, Part V: guidance on complex human health detailed quantitative risk assessment for*

chemicals (DQRACHEM). Ottawa. Retrieved from http://www.hc-sc.gc.ca/ewh-semt/pubs/contamsite/chem-chim/index-eng.php on August 22, 2016.

Hébert, J. & Bernard, J. (2013). *Bilan sur la gestion des terrains contaminés au 31 Decembre 2010. Gouvernement du Québec*. Retrieved from http://www.mddep.gouv.qc.ca/sol/terrains/bilan/bilan2010.pdf on June 30, 2016.

Heck, R.H. & Thomas, S.L. (2015). *An Introduction to Multilevel Modeling Techniques: MLM and SEM Approaches Using Mplus*. Routledge.

Helsel, D. (2006). Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere, 65* (11), 2434-2439.

Helsel, D. (2010a). Much ado about next to nothing: incorporating nondetects in science. *Annals of Occupational Hygiene, 54* (3), 257-262.

Helsel, D. (2010b). Summing nondetects: Incorporating low-level contaminants in risk assessment. *Integrated environmental assessment and management, 6* (3), 361-366.

Helsel, D. & Cohn T.A. (1988). Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research, 24* (12), 1997-2004.

Helsel, D. (1990). Less than obvious-statistical treatment of data below the detection limit. *Environmental science & technology, 24* (12), 1766-1774.

Helsel, D. (2005).More than obvious: better methods for interpreting nondetect data. *Environmental science & technology, 39* (20), 419A-423A.

Helsel, D. (2012). *Statistics for censored environmental data using Minitab and R*. Hoboken, N.J.: Wiley.

Hewett, P. & Ganser, G.H. (2007). A comparison of several methods for analyzing censored data. *Annals of Occupational Hygiene, 51* (7),. 611-632.

Heydari, S., Miranda-Moreno, L.F. & Fu, L. (2014). Speed limit reduction in urban areas: A before–after study using Bayesian generalized mixed linear models. *Accident Analysis & Prevention, 73*, 252-261.

Higley, K. A. (2010). Estimating transfer parameters in the absence of data. *Radiation and environmental biophysics, 49* (4), 645-656.

Hornung, R. W. & Reed, L.D. (1990). Estimation of average concentration in the presence of nondetectable values. *Applied occupational and environmental hygiene, 5* (1), 46-51.

Hox, J.J. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Hsu, J.-F., Guo, H.-R., Wang, H.W., Liaoand, C.-K. & Liao, P.-C. (2011). An occupational exposure assessment of polychlorinated dibenzo-p-dioxin and dibenzofurans in firefighters. *Chemosphere, 83* (10), 1353-1359.

Huybrechts, T., Thas, O., Dewulfand, J. & Van Langenhove, H. (2002). How to estimate moments and quantiles of environmental data sets with non-detected observations? A case study on volatile organic compounds in marine water samples. *Journal of Chromatography A, 975* (1), 123-133.

Huynh, T., Ramachandran, G., Banerjee, S., Monteiro, J., Stenzel, M., Sandler, D.P., … Stewart, P.A. (2014). Comparison of methods for analyzing left-censored occupational exposure data. *Annals of Occupational Hygiene,* 58 (9), 1126-1142.

Jain, R.B., Caudill, S.P., Wang, R.Y. & Monsell, E. (2008). Evaluation of maximum likelihood procedures to estimate left censored observations. *Analytical chemistry, 80* (4), 1124-1132.

Jain, R.B. & Wang, R.Y. (2008). Limitations of maximum likelihood estimation procedures when a majority of the observations are below the limit of detection. *Analytical chemistry, 80* (12), 4767-4772.

Janssen, D.P. (2012). Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants. *Behavior Research Methods, 44* (1), 232-247.

Jin, Y., Hein, M.J., Deddens, J.A. & Hines, C.J. (2011). Analysis of lognormally distributed exposure data with repeated measures and values below the limit of detection using SAS. *Annals of Occupational Hygiene, 55* (1), 97-112.

Jones, R.J. (2011). Spatial patterns of chemical contamination (metals, PAHs, PCBs, PCDDs/PCDFS) in sediments of a non-industrialized but densely populated coral atoll/small island state (Bermuda). *Marine pollution bulletin, 62* (6), 1362-1376.

Jordan, L., Schimleck, L.R., Clark, A., Hall, D.B. & Daniels, R.F. (2007). Estimating optimum sampling size to determine weighted core specific gravity of planted loblolly pine. *Canadian journal of forest research, 37* (11), 2242-2249.

Knight, K. (2000). *Mathematical statistics*. Boca Raton: Chapman & Hall/CRC Press.

Krapac, I., Dey, W.S., Roy, W.R., Smyth, C.A., Storment, E., Sargent, S.L. & Steele, J.D. (2002). Impacts of swine manure pits on groundwater quality. *Environmental Pollution, 120* (2), 475-492.

Kreft, I. & DeLeeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Krishnamoorthy, K., Mallick, A. & Mathew, T. (2009). Model-based imputation approach for data analysis in the presence of non-detects. *Annals of Occupational Hygiene*, *53* (3), 249-263.

Kroll, C.N. & Stedinger, J.R. (1996). Estimation of moments and quantiles using censored data. *Water Resources Research, 32* (4), 1005-1012.

Kuttatharmmakul, S., Smeyers-Verbeke, J., Massart, D.L., Coomans, D. & Noack, S. (2000). The mean and standard deviation of data, some of which are below the detection limit: an introduction to maximum likelihood estimation. *TrAC Trends in Analytical Chemistry, 19* (4), 215-222.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. Hoboken, N.J: John Wiley & Sons.

Lee, E.T. & Wang, J.W., (2003). *Statistical methods for survival data analysis*. Hoboken, N.J: John Wiley & Sons.

Lee, H. J. & Koutrakis, P. (2014). Daily ambient NO2 concentration predictions using satellite ozone monitoring instrument NO2 data and land use regression. *Environmental science & technology, 48* (4), 2305-2311.

Lee, L. & Helsel, D. (2005). Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics. *Computers & Geosciences, 31* (10), 1241-1248.

Lee, L. & Helsel, D. (2007). Statistical analysis of water-quality data containing multiple detection limits II: S-language software for nonparametric distribution modeling and hypothesis testing. *Computers & Geosciences, 33* (5), 696-704.

Leith, K.F., Bowerman, W.W., Wierda, M.R., Best, D.A., Grubband, T.G. & Sikarske, J.G. (2010). A comparison of techniques for assessing central tendency in left-censored data using PCB and p, p′ DDE contaminant concentrations from Michigan's Bald Eagle Biosentinel Program. *Chemosphere, 80* (1), 7-12.

Liero, H. & Zwanzig, S. (2011). *Introduction to the theory of statistical inference*. Boca Raton, Florida, USA: Chapman and Hall/CRC.

Liu, S., Lu, J.-C., Kolpin, D.W. & Meeker, W.Q. (1997). Analysis of environmental data with censored observations. *Environmental science & technology,* 31 (12), 3358-3362.

Lubin, J.H., Colt, J.S., Camann, D., Davis, S., Cerhan, J.R., Severson, R.K., … Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental health perspectives*, 1691-1696.

Lynn, H.S. (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statistics in medicine, 20* (1), 33-45.

McCarthy, M.C., O'Brien, T.E., Charrier J.G. & Hafner, H.R. (2009). Characterization of the chronic risk and hazard of hazardous air pollutants in the United States using ambient monitoring data. *Environmental health perspectives, 117* (5), 790-6.

Ministère du Développement durable, de l'Environnement et des Parcs du Québec. (2003a). *Guide de caractérisation des terrains*. Les Publications du Québec, 111 p.

Ministère du Développement durable, de l'Environnement et des Parcs du Québec, (2003b). *Soil Protection and Contaminated Sites Rehabilitation Policy. Les publications du Quebec*. Retrieved from http://legisquebec.gouv.qc.ca/en/pdf/cr/Q-2,%20R.%2037.pdf on September 12, 2016.

Newman, M.C., Dixon, P.M., Looney,B.B. & Pinder, J.E. (1989). Estimating mean and variance for environmental samples with below detection limit observations. *Water Resources Bulletin*.

Nocerino, J., Schumacher, B. & Dary,C. (2005). Role of Laboratory Sampling Devices and Laboratory Subsampling Methods in Representative Sampling Strategies. *Environmental Forensics, 6* (1), 35-44.

Ofungwu, J. (2014). *Statistical applications for environmental analysis and risk assessment*. Hoboken, NJ (USA): Wiley.

Ott, W. R. (1990). A physical explanation of the lognormality of pollutant concentrations. *Journal of the Air & Waste Management Association*, *40* (10), 1378-1383.

Pajek, M., Kubala-Kukuś, A., Banaś, D., Braziewicz, J. & Majewska, U. (2004). Random left-censoring: a statistical approach accounting for detection limits in x-ray fluorescence analysis. *X-Ray Spectrometry, 33* (4), 306-311.

Peretz, C., Goren, A., Smid, T. & Kromhout, H. (2002). Application of mixed-effects models for exposure assessment. *The Annals of occupational hygiene, 46* (1), 69-77.

Perez, A. & Lefante, J.J. (1997). Sample size determination and the effect of censoring when estimating the arithmetic mean of a lognormal distribution. *Communications in Statistics-Theory and Methods, 26* (11), 2779-2801.

Persson, T. & Rootzen, H. (1977). Simple and highly efficient estimators for a type I censored normal sample. *biometrika Biometrika, 64* (1), 123-128.

Pinheiro, J. & Bates, D.(2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Powell, J.L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics, 25* (3), 303-325.

Quéformat Ltée. (2004). *Caractérisation préliminaire phase I. Propriétés situées dans le quadrilatère délimité par les rues De la Montagne, Notre-Dame, Eleanor et William. Montréal, Québec*. S-11603-A. Longueuil, QC: Quéformat Ltée.

Röösli, M., Frei, P., Mohler, E., Fahrländer, C.B., Bürgi, A., Fröhlich, J., … Egger, M. (2008). Statistical analysis of personal radiofrequency electromagnetic field measurements with nondetects. *Bioelectromagnetics, 29* (6), 471-478.

Sapkota, A., Heidler, J. & Halden, R.U. (2007). Detection of triclocarban and two co-contaminating chlorocarbanilides in US aquatic environments using isotope dilution liquid chromatography tandem mass spectrometry. *Environmental research, 103* (1), 21-29.

Schäfer, R.B., Paschke, A., Vrana, B., Mueller, R. & Liess, M. (2008). Performance of the Chemcatcher® passive sampler when used to monitor 10 polar and semi-polar pesticides in 16 Central European streams, and comparison with two other sampling methods. *Water Research, 42* (10), 2707-2717.

Scherbaum, C.A. & Ferreter, J.M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12* (2), 347-367.

Schisterman, E.F., Vexler, A., Whitcomb B.W. & Liu, A. (2006). The limitations due to exposure detection limits for regression models. *American journal of epidemiology, 163* (4), 374-383.

Schmitt, J.H.M.M. (1985). Statistical analysis of astronomical data containing upper bounds - General methods and examples drawn from X-ray astronomy. *The Astrophysical Journal, 293*.

Schmoyeri, R. L., Beauchamp, J.J., Brandt, C.C. & Hoffman, F.O. (1996). Difficulties with the lognormal model in mean estimation and testing. *Environ Ecol Stat Environmental and Ecological Statistics, 3* (1), 81-97.

Schneider, H. & Weissfeld, L. (1986). Estimation in Linear Models with Censored Data. *biometrika Biometrika, 73* (3), 741-745.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association Journal of the American Statistical Association, 63* (324), 1379-1389.

She, N. (1997). Analyzing censored water quality data using a non-parametric approach. *JAWRA Journal of the American Water Resources Association, 33* (3), 615-624.

Shoari, N., Dubé, J.-S. & Chenouri, S. (2015). Estimating the mean and standard deviation of environmental data with below detection limit observations: Considering highly skewed data and model misspecification. *Chemosphere, 138*, 599-608.

Shoari, N., Dubé, J.-S. & Chenouri, S. (2016). On the use of the substitution method in left-censored environmental data. *Human & ecological risk assessment, 22* (2), 435-446.

Shumway, R.H., Azari, R.S. & Kayhanian, M. (2002). Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental science & technology, 36* (15), 3345-3353.

Singh, A., Maichle, R. & Lee, S.E. (2006). On the computation of a 95% upper confidence limit of the unknown population mean based upon data sets with below detection limit observations.

Singh, A. & Nocerino, J. (2002). Robust estimation of mean and variance using environmental data sets with below detection limit observations. *Chemometrics and Intelligent Laboratory Systems, 60* (1), 69-86.

Singh, A. & Singh, A.K. (2013). *ProUCL Version 5.0.00 Technical Guide - Statistical software for environmental applications for datasets with and without nondetect observations*. Washington D.C: United States Environmental Protection Agency.

Singh, A., Singh, A.K. & Iaci, R.J. (2002).*Estimation of the exposure point concentration term using a gamma distribution*. Las Vegas, NV: Technology Support Center for Monitoring and Site Characterization.

Singh, K.P., Bartolucci,A.A. & Bae,S. (2001).Mathematical modeling of environmental data. *Mathematical and computer modelling, 33* (6), 793-800.

Sinha, P., Lambert, M.B. & Trumbull, V.L. (2006). Evaluation of statistical methods for left-censored environmental data with nonuniform detection limits. *Environmental toxicology and chemistry, 25* (9), 2533-2540.

Slymen, D.J., De Peyster, A. & Donohoe, R.R. (1994). Hypothesis testing with values below detection limit in environmental studies. *Environmental science & technology, 28* (5), 898-902.

Strucinski, P., Morzycka, B., Góralczyk, K., Hernik, A., Czaja, K., Korcz, W., … Ludwicki, J.K. (2015). Consumer Risk Assessment Associated with Intake of Pesticide Residues in Food of Plant Origin from the Retail Market in Poland. *Human & ecological risk assessment, 21* (8), 2036-2061.

Succop, P. A., Clark, S., Chen, M. & Galke, W. (2004). Imputation of data values that are less than a detection limit. *Journal of occupational and environmental hygiene, 1* (7), 436-441.

Tate, E.L. & Freeman, S.N. (2000). Three modelling approaches for seasonal streamflow droughts in southern Africa: the use of censored data. *Hydrological sciences journal, 45* (1), 27-42.

Theil, H. 1992. A rank-invariant method of linear and polynomial regression analysis. In *Henri Theil's Contributions to Economics and Econometrics*. 345-381. Springer.

Thiébaut, R. & Jacqmin-Gadda, H. (2004). Mixed models for longitudinal left-censored repeated measures. *Computer methods and programs in biomedicine, 74* (3), 255-260.

Thompson, M.L. & Nelson, K.P. (2003). Linear regression with Type I interval-and left-censored response data. *Environmental and Ecological Statistics, 10* (2), 221-230.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24-36.

Tong, L.-I., Chang, C.-W., Jin, S.-E. & Saminathan, R. (2012). Quantifying uncertainty of emission estimates in National Greenhouse Gas Inventories using bootstrap confidence intervals. *AEA Atmospheric Environment, 56*, 80-87.

Tong, L.-I., Saminathan, R. & Chang, C.W. (2016). Uncertainty assessment of non-normal emission estimates using non-parametric bootstrap confidence intervals. *Journal of Environmental Informatics*. *28* (1), 61-70.

Travis, C.C. & Land, M.L. (1990). Estimating the mean of data sets with nondetectable values. *Environmental Science & Technology, 24* (7), 961-962.

Twisk, J. & Rijmen, F. (2009). Longitudinal tobit regression: a new approach to analyze outcome variables with floor or ceiling effects. *Journal of clinical epidemiology, 62* (9), 953-958.

Uh, H.-W., Hartgers, F.C., Yazdanbakhsh, M. & Houwing-Duistermaat, J.J. (2008). Evaluation of regression methods when immunological measurements are constrained by detection limits. *BMC immunology, 9* (1).

USEPA. (2000). *Practical methods for data analysis, Guidance for data quality assessment*. Washington D.C.: United States Environmental Protection Agency.

USEPA. (2004). *Local limits development guidance appendices*. Washington D.C: United States Environmental Protection Agency.

USEPA. (2006). *Data quality assessment:Statistical methods for practitioners*. Washington D.C: United States Environmental Protection Agency. Retrieved from http://www.epa.gov/QUALITY/dqa.html on August 23, 2016.

USEPA. (2009). *Statistical analysis of groundwater monitoring data at RCRA facilities, Unified guidance*. Unite States Environmental Protection Agency.

Vaida, F. & Liu, L. (2009). *"lmec": Linear Mixed-Effects Models with Censored Responses* (Version Responses R Package Version 1. 2012).

Vaida, F. & Liu, L. (2012). Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics*.

Van Breukelen, G.J. & Candel, M.J. (2012). Calculating sample sizes for cluster randomized trials: We can keep it simple and efficient!. *Journal of clinical epidemiology, 65* (11), 1212-1218.

Vassura, I., Passarini, F., Ferroni, L., Bernardi, E. & Morselli, L. (2011). PCDD/Fs atmospheric deposition fluxes and soil contamination close to a municipal solid waste incinerator. *Chemosphere.83* (10), 1366-1373.

Warne, R.T., Li, Y., McKyer, E.L.J., Condie, R., Diep, C.S. & Murano, P.S. (2012). Managing clustered data using hierarchical linear modeling. *Journal of nutrition education and behavior, 44* (3), 271-277.

Watkins, D.J., Fortenberry, G.Z., Sánchez, B.N., Barr, D.B., Panuwet, P., Schnaas, L.,… Meeker, J.D. (2016). Urinary 3-phenoxybenzoic acid (3-PBA) levels among pregnant women in Mexico City: Distribution and relationships with child neurodevelopment. *Environmental research, 147*, 307-313.

West, B.T., Welch, K.B. & Galecki, A.T. (2014). *Linear mixed models: a practical guide using statistical software*. CRC Press.

Wu, L. (2009). *Mixed effects models for complex data*. CRC Press.

Wu, R., Qian, S.S., Hao, F., Cheng, H., Zhu, D. & Zhang, J. (2011). Modeling contaminant concentration distributions in China's centralized source waters. *Environmental science & technology, 45* (14), 6041-6048.

Wu, S., Yang, D., Wei, H., Wang, B., Huang, J., Li, H., … Guo, X. (2015). Association of chemical constituents and pollution sources of ambient fine particulate air pollution and biomarkers of oxidative stress associated with atherosclerosis: A panel study among young adults in Beijing, China. *Chemosphere, 135*, 347-353.

Zhao, Y. & Frey, H.C. (2003). Quantification of uncertainty and variability for air toxic emission factor data sets containing non-detects. In *Proceedings, Annual Meeting of the Air & Waste Management Association*. *53*, 1436-47.

Zhao, Y. & Frey, H.C. (2006). Uncertainty for data with non-detects: Air toxic emissions from combustion. *Human and Ecological Risk Assessment: An International Journal, 12* (6), 1171-1191.