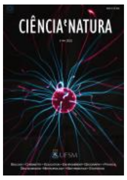
UFSM

CIÊNCIA e NATURA



ISSN 2179-460X
Acesso aberto

Ci. e Nat., Santa Maria, v. 44, e24, 2022• https://doi.org/10.5902/2179460X63368
Submitted: 02/12/2020•Approved: 15/11/2021•Published: 06/06/ 2022

Statistics

OCCURRENCE OF CHEMICAL SUBSTANCES IN WATER SUPPLY SYSTEMS OF BRAZIL: A NONPARAMETRIC APPROACH FOR STATISTICAL ANALYSIS OF SISAGUA DATA

Ocorrência de substâncias químicas em sistemas de abastecimento de água do Brasil: Uma abordagem não-paramétrica para a análise estatística de dados do Sisagua

Fernanda Bento Rosa Gomes[I], **Guilherme Bento Nicolau**[I], **Emanuel Manfred Freire Brandt**[I], **Samuel Rodrigues Castro**[I,II], **Renata de Oliveira Pereira**[I,II], **Taciane de Oliveira Gomes de Assunção**[II] **Pedro Fialho Cordeiro**[III]

[I] Federal University of Juiz de Fora, Graduate Program in Civil Engineering, Juiz de Fora, MG, Brazil
[II] Federal University of Juiz de Fora, Departament of Sanitary and Environmental Engineering, Juiz de Fora, MG, Brazil
[III] SENAI FIEMG Innovation and Technology Center, Belo Horizonte, MG, Brazil

ABSTRACT

The objective of this work was to develop a method for statistical analysis of monitoring data of chemical substancesin drinking water supply systems in Brazil using data from Sisagua (Drinking Water Quality Surveillance Information System). A procedure to check the consistency of the database was proposed and solutions toeach inconsistency were described. Then, descriptive statistics were estimated using the Kaplan-Meier (KM) method, assessingits applicability to different censored data sets. Descriptive parameters estimated by the KM method were compared with those obtained by the substitution method. Substitution method showed susceptibility to biased estimates, especially for datasets withalarge percentage of censored data and high limits of quantification ordetection, estimating higher descriptive parameters than the obtained by the KM method. This work reinforces the need to use appropriate methods for analyzing environmental data and evidences that the analysis of this type of data may be complex. The methods proposed here can help environmental scientists to deal with this issue, providing a systematic procedure to check and solve consistency problems, as well as presenting a nonparametric approach for computing descriptive statistics for environmental monitoring data.

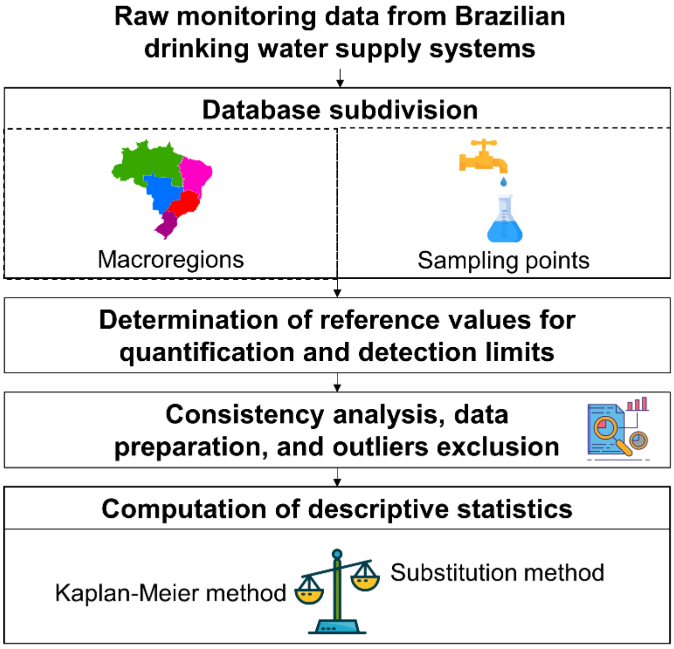
Keywords: Censored data; Drinking water; Environmental data; Kaplan-Meier estimators; Nondetects

RESUMO

Este estudo teve como objetivo o desenvolvimento de uma metodologia para o tratamento estatístico de dados de monitoramento de substâncias químicas em sistemas de abastecimento de água do Brasil, utilizando-se dados do SISAGUA (Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano). Propôs-se uma metodologia para a análise de consistência da base de dados, bem soluções para todas as inconsistências identificadas. Em seguida, estatísticas descritivas foram estimadas pelo método de Kaplan-Meier (KM), avaliando-se a sua aplicabilidade a diferentes proporções de dados censurados. Os parâmetros descritivos obtidos pelo método de KM foram comparados aos obtidos pelo método de substituição. De modo geral, o método de substituição demonstrou maior suscetibilidade a estimativas enviesadas, notadamente com o aumento do percentual de censura e em meio a elevados limites de quantificação e detecção, conduzindo àestimativa de parâmetros descritivos mais altos em relação aos estimados pelo método de KM. O estudo reforça a necessidade do uso de métodos apropriados para a análise de dados ambientais, além de evidenciar que o tratamento desse tipo de dado pode ser uma tarefa complexa. Dessa forma, a metodologia proposta pode ser útil a pesquisadores, uma vez que apresenta um processo sistemático de identificação e correção de inconsistências, bem como uma abordagem não paramétrica para a obtenção de estatísticas descritivas para dados de monitoramento ambiental.

Palavras-chave: Água potável; Dados ambientais; Dados censurados; Estimadores de Kaplan-Meier; Não detectados

GRAPHICAL ABSTRACT


Raw monitoring data from Brazilian drinking water supply systems
Database subdivision
Macroregions
Sampling points
Determination of reference values for quantification and detection limits
Consistency analysis, data preparation, and outliers exclusion
Computation of descriptive statistics
Kaplan-Meier method
Substitution method

1 INTRODUCTION

Water supply is a primary action for public health protection. In this regard, Brazilian drinking water standard, Ordinance GM/MS nº 888/2021 (BRASIL, 2021), establishes physical, chemical, and biological characteristics that must be guaranteed by service providers of water supply aiming at public health protection. Additionally, the National Drinking Water Quality Surveillance Program (VIGIAGUA) uses the Drinking Water Quality Surveillance Information System (Sisagua) to systematize monitoring data for control and surveillance of water potability in Brazil. Sisagua is fed with information on control done by water supply service providers and water quality surveillance exercised by State and Municipal Health Departments. Currently, Sisagua is the main instrument for assessing and monitoring drinking water quality in Brazil. However, due to the presence of inconsistencies and left-censored data (data not quantified which its concentration is less than one or more values/limits determined by analytical methods), the analysis of data from this database maybe a difficult process.

Several approaches can be used to compute descriptive statistics of data sets with left-censored data. According to Christofaro and Leão (2014), these approaches can be subdivided into at least four classes: substitution, parametric, robust, and nonparametric methods. Due to its simplicity, the substitution method is the most commonly used (HELSEL, 2006).

On the other hand, nonparametric methods do not require that data fit a previously known probability density function. These methods can be applied fordata setswith a reduced amount of data and tend to be less affected by outliers (CHRISTOFARO and LEÃO, 2014; ANTWEILER and TAYLOR, 2008). Kaplan-Meier method is one of the most widely usednonparametric methods (ANTWEILER and TAYLOR, 2008). It is based on estimatingthe censored data from the distribution of quantified data (KAPLAN and MEIER, 1958; SINGH *et al*., 2006).

Applying various statistical methods, several authors have been assessed the occurrence of chemical substancesin water bodies (MARIMON *et al*., 2013;

CHRISTOFARO and LEÃO, 2014; SABINO *et al*., 2014; MELO GURGEL *et al*., 2016; CASSANEGO and DROSTE, 2017; DALZOCHIO *et al*., 2017; STAPLES *et al*., 2018). However, in order to assure a robust analysis, particularitiestypically observed in environmental data must be considered. As reported by Helseland Hirsch (2002), environmental datasets frequently do not follow the Gaussian distributionand usually have outliers and left-censored data. Neglecting these aspects canseriously affect the reliability of the estimation of descriptive parameters, hypothesis tests, regression models, and trend analyses (HELSEL, 2006; HELSEL and HIRSCH, 2002; SINGH *et al.,* 2006; CHRISTOFARO and LEÃO, 2014).

Therefore, this work aimed at developing a methodology for statistical analysis of Sisagua data, assessingits applicability in different proportions of censored data and comparing it to another methodological approach (substitution method) which is usually used to obtain descriptive statistics of environmental data.

2 MATERIAL AND METHODS

2.1 Data collection and characterization

Sisaguahas monitoring data of Brazilian water sources (raw water) and drinking water (finished and tap water) quality (BRASIL, 2020). Through Sisagua, 2,569,234 monitoring data from 2014 to 2018 referring to 79 chemical variables were obtained. The Brazilian drinking water standard (former Consolidation Ordinance MS nº 5/2017, Annex XX) was reviewed in 2021 (current Ordinance GM/MS nº 888/2021) (BRASIL, 2021). However, currently Sisagua only has data regarding to substances addressed in the previous drinking water standard (Consolidation Ordinance nº 5/2017, Annex XX) (BRASIL, 2017).

The database is grouped in the following categories according to the Brazilian drinking water standard: (i) pesticides, (ii) inorganic substances, (iii) organic compounds, (iv) disinfection by-products, and (v) organoleptic substances. Data collected in this study are from the Sisagua Control module which includes monitoring data from water supply systems (SAA) and alternative collective solutions (SAC) (OLIVEIRA JÚNIOR et al., 2019).

The original spreadsheets from Control module include the following information:

- Name and code of the monitoring point;
- Name and code of the municipality and federative unit;
- Name of the water treatment plant (WTP) and the type of water supply (SAA or SAC);
- Date, year, and semester of sampling;
- Date of the chemical analysis;
- Name and category of the substance;
- The maximum allowed value (MAV) for the compound according to the Brazilian drinking water standard;
- Dates of registration and data upload;
- The values of the analytical limits (LD and LQ); and the analytical result.

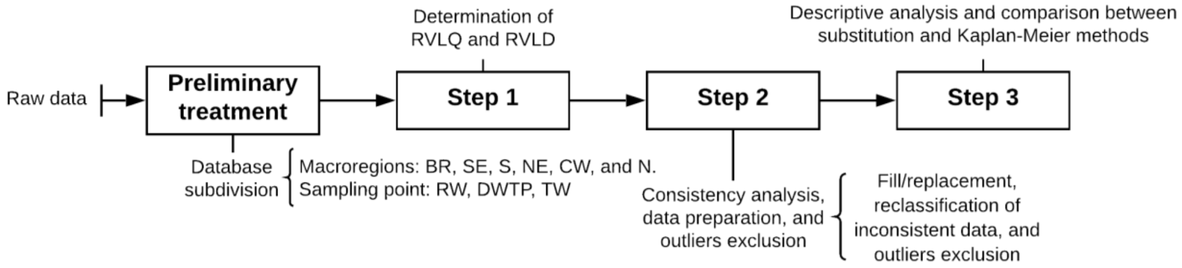
Results and the classification of the data are expressed in one column of the spreadsheets. When the concentration of the sample was quantified by the chemical analysis, the cell shows its numerical value. On the other hand, if the concentration was not quantified or detected by the analytical method, the cell shows the classification of the result (less than LQ or less than LD).

2.2 Statistical analysis

The original spreadsheetswere subdivided in macroregions (Brazil – BR; Southeast – SE; South – S; Northeast – NE; Central-West – CW; and North – N) and according to the type of the sampling point, as follow: (i)raw water source – RW;(ii) drinking water treatment plant– DWTP (finished waters); and (iii)drinking water distribution systems and tap waters – DW.

As the informationareregistered manually in Sisagua, some inconsistencies may occur in this database. In this sense, reference values for analytical limits of detection and quantification (Step 1) had to be established prior to the solution of consistency problems. Subsequently, steps to fill, replace, and reclassify inconsistent data were proposed (Step 2) and, finally, descriptive analyses were carried out (Step 3). A general scheme of the data analysisprocedureis shown in Figure 1.

Figure 1 - General flowchart of the methodological steps proposed for analysis of Sisagua data.


Raw data
Preliminary treatment
Determination of RVLQ and RVLD
Step 1
Step 2
Descriptive analysis and comparison between substitution and Kaplan-Meier methods
Step 3
Database subdivision
Macroregions: BR, SE, S, NE, CW, and N.
Sampling point: RW, DWTP, TW
Consistency analysis, data preparation, and outliers exclusion
Fill/replacement, reclassification of inconsistent data, and outliers exclusion

Legend: RW: raw water source; DWTP: drinking water treatment plants (finished water); DW: drinking water distribution systems and tap waters; RVLD: reference value detection limits; RVLQ: reference value of quantification limits.

Determination of reference values for analytical limits of detection and quantification– Step 1

In this step, reference values (RVs) for unreported LDs and LQs were established. For this purpose, all the analytical limits reported in the spreadsheetfor each chemical variable were grouped into two data sets: one for LD values and other for LQ values.

Two tests for detection of outliers were compared: Grubbs' test and Tukey test. The first one, the Grubbs' test (GRUBBS, 1979), is used to identify the presence of an outlier in an approximately normal distribution (excluding outliers' values). Grubbs' test statistic for a two-tailed test is defined according to Equation 1 (URVOY and AUTRUSSEAU, 2014).

(Equation 1)

$$G = \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{(t_{\alpha/(2N),N-2})}{N-2+(t_{\alpha/(2N),N-2})^2}}$$

Where N is the number of samples; $t_{\alpha}/_{(2N),N-2}$ is the critical value of Student's t distribution with N-2 degrees of freedom and a significance level of α_G/(2N). Grubbs' test can be used iteratively to detect multiple outliers.

Tukey test (1977), also known as the boxplot method, is based on the interval between the 25th percentile (Q1) and 75th percentile (Q3) (Equation 2). In this method, an observation outside the range of inner fences (Equation 3; Equation 4) can be classified as an outlier.

(Equation 2)

$$\text{Interquartile range (IQR)} = (\text{Q3} - \text{Q1})$$

(Equation 3)

$$\text{Upper inner fence} = \text{Q3} + 1.5\ \text{IQR}$$

(Equation 4)

$$\text{Lower inner fence} = \text{Q1} - 1.5\ \text{IQR}$$

Comparing the Grubbs' and Tukey methods, it was observed that the use of the Grubbs' method resulted in a low percentage of purged data. Given this fact, the Grubbs' method was used to detect outliers for LDs and LQs considering also the following aspects:

(i) Removing a minor amount of data implies inhigh maximum values for LD and LQ. It is important to consider that these maximum values will be the upper limits for the estimation of the censored data andwill directly affect the estimation of descriptive statistics. This choice results in high values for the statistical parameters, which is a more conservative decision in the perspective of environmental management;

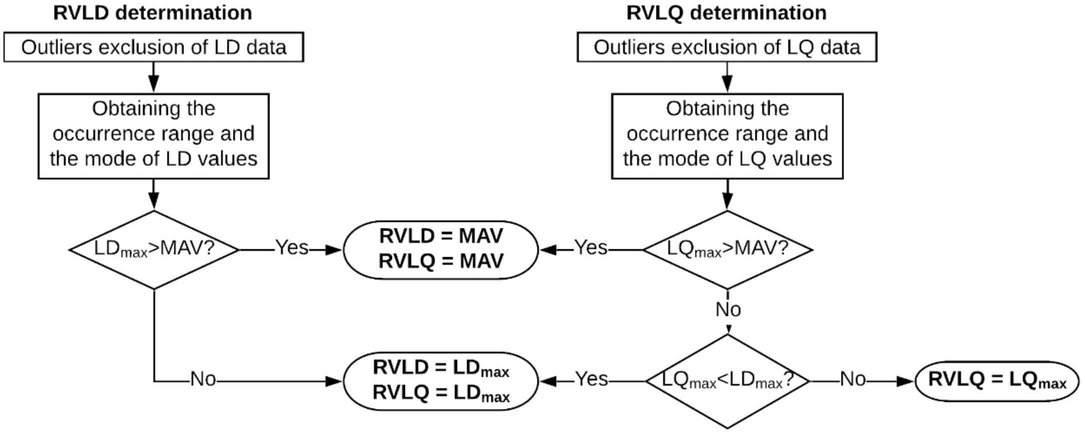
(ii) Grubbs' test is widely used in analytical methods (ANALYTICAL METHODS COMMITTEE, 2015); and

(iii) Analytical methods are standardized procedures. Therefore, it was assumed that the LDs and LQs of the methods practiced in Brazil float around a mean.

Thus, lower and upper outliers were excluded through an iterative process with the Grubbs' test (1979) in the STATISTICA 8 software (STATSOFT, 2007) at 95% confidence level. After removing outliers, ranges of occurrence and the modes of the reported LDs and LQs were obtained. For each chemical variable, two RVs were established, one for LDs (RVLD) and another for LQs (RVLQ). These RVs were defined as the maximum values for each range of reporting limits (CETESB, 2001). When the upper concentration of the ranges of LDs and/or LQs exceeded the maximum allowed value (MAV) for the compound, the MAV was defined as the RV.

For some cases, the RV initially assigned for the LQs was lower than the RVLD. As for definition the LQ must be equal to or higher than the LD, in these cases the RVLQ was set with the same value as the RVLD. Figure 2 shows the flowchart of the Step 1.

Figure 2 - Determination of RVLDs and RVLQsusedin database consistency analysis (Step 1).

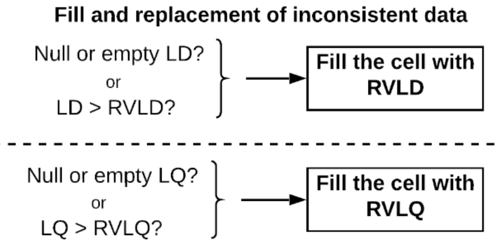

RVLD determination
Outliers exclusion of LD data
Obtaining the occurrence range and the mode of LD values
LDmax>MAV?
Yes
No
RVLD = MAV
RVLQ = MAV
RVLD = LDmax
RVLQ = LDmax
RVLQ determination
Outliers exclusion of LQ data
Obtaining the occurrence range and the mode of LQ values
LQmax>MAV?
Yes
No
LQmax<LDmax?
Yes
No
RVLQ = LQmax

Legend: RVLD: reference value for limits of detection; RVLQ: reference value forlimits of quantification; LD_{max}: maximum limit of detection used in Brazil; LQ_{max}: maximum limit of quantificationused in Brazil; MAV: maximum allowed value according to the Brazilian drinking water standard.

Consistency analysis, data preparation, and outliersexclusion – Step 2

Initially, cells regarding to analytical limits (LD and LQ) that were null or empty were filled with the RVLDs and RVLQs established in Step 1. Furthermore, cells in which the reported LD was higher than the RVLD were replaced by the RVLD. Correspondingly, cells with a LQ higher than the RVLQ were replaced bythe RVLQ (Figure 3).

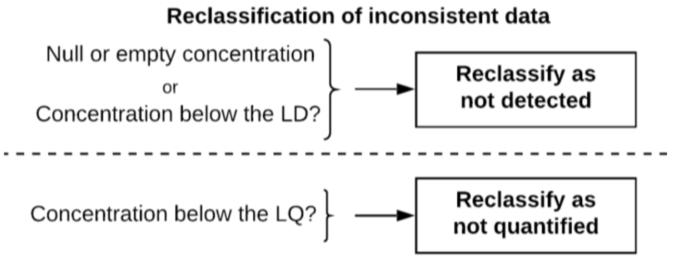
Figure 3 - Flowchart of fill/replacement of inconsistent data



Legend: RVLD: reference value of detection limits; RVLQ: reference value of quantification limit; LD: limit of detection; LQ: limit of quantification.

For the case of quantified data, outlier observations were excluded through the Tukey test (TUKEY, 1977), as reported by Jeong *et al*. (2017), considering the fact that environmental data usually are not normally distributed. Subsequently, inconsistent quantified data were reclassified according to Figure 4.

Figure 4 - Flowchart of reclassification of inconsistent data.


Reclassification of inconsistent data
Null or empty concentration
or
Concentration below the LD?
Reclassify as not detected
Concentration below the LQ?
Reclassify as not quantified

Legend: LD: limit of detection; LQ: limit of quantification

Descriptive analysis – Step 3

Left-censored data (also called nondetects) are low level analytical results which could not be measured accurately by an analytical method. Due to this, their concentrations are reported by laboratories as less than a given analytical threshold (i.e., less than LD or less than LQ) (HELSEL, 2006; SINGH *et al*., 2006).

Chemical analyses in environmental matrices frequently include left-censored data. However, these data are often associated with serious interpretation problems. Frequently, censored data are excluded from the data sets or substituted by an arbitrary value leading to biased (overestimated or underestimated) statistical parameters (GEORGE *et al*., 2021; MCGRORY *et al.*, 2020; SINGH *et al*., 2006; HELSEL and HIRSCH, 2002; ANTWEILER and TAYLOR, 2008).

Several methods for handling censored data have been reported in literature. These techniques include parametric (e. g., maximum likelihood estimation), robust (e. g., robust regression on order statistic), and nonparametric (e.g., Kaplan-Meier) methods, as well as a simple substitution of the reporting limits by a given constant (CHRISTOFARO and LEÃO, 2014; HELSEL, 2005; GEORGE *et al*., 2021; MCGRORY *et al*., 2020; SINGH *et al*., 2006; HELSEL and HIRSCH, 2002; ANTWEILER and TAYLOR, 2008; GILLESPIE *et al*., 2010).

Substitution method consists in attributing a constant value to the censored data. In environmental sciences researches usually use zero, the value of the analytical limit (LD or LQ), or one-half of the analytical limit as reference values for left-censored data (HELSEL, 2004; HELSEL, 2006; ANTWEILER and TAYLOR, 2008). Substitution method is the simplestand still the most commonly used (CHRISTOFARO and LEÃO, 2014; SINGH *et al*., 2006). However, in the last years, this method has been discouraged by the technical community (USEPA, 2010; USEPA, 2016) and other approaches such as the Kaplan-Meier method has been gained attention (HELSEL, 2005).

Kaplan–Meieris a nonparametric method which was initially developed for computing descriptivestatistics of right-censored survival data. In environmental sciences, the Kaplan-Meier estimators are used with a reverse scale for analyzing left-censored data (FLIKKEMA, 2016; GILLESPIE *et al*., 2010; HUYNH *et al*., 2014).

The original Kaplan-Meier method (for right-censored data) computes a left-continuous cumulative distribution function (survival function) to estimate the probability of an individual surviving beyond a given time x. For environmental data, the Kaplan-Meier method estimates a right-continuous cumulative distribution function according to Equation 5. The Kaplan-Meier estimator is given by Equation 6, where the quantified (uncensored) are denoted by x_j, n_j is the number of values (accounting censored and quantified data) less than or equal tox_j, and d_j is the number of quantified values equal to x_j.

(Equation 5)

$$F(x) = P(X \leq x)$$

$$\hat{F}(x) = \begin{cases} 1 & x_j \leq x \\ \prod_{j;x_j>x} \frac{(n_j - d_j)}{n_j} & x < x_j \end{cases}$$

(Equation 6)

It is worthwhile to mention that Kaplan-Meier methodcan handle multiple distinct reporting limits and, as a nonparametric method, does not require an assumed distribution and is less affected by outliers (FLIKKEMA, 2016; HELSEL, 2005; CHRISTOFARO and LEÃO, 2014).

Kaplan-Meier method is available in United States Environmental Protection Agency's (USEPA) ProUCL software for estimating the mean, its upper confidence limit (95% UCL), and the standard deviation for data sets with left-censored data (USEPA, 2010; USEPA, 2016). Furthermore, the Kaplan-Meier method for environmental data (including for computing percentiles, confidence intervals and performing hypothesis tests) is also available in some statistical packages in R environment (DELIGNETTE-MULLER *et al*., 2018; LEE, 2017). Additionally, users can also adapt the algorithms developed for right-censored data as reported by Huynh *et al*. (2014).

In this study, the Kaplan-Meier estimators (KAPLAN and MEIER, 1958), as described above, were used as a nonparametric approach to compute descriptive statistics of water quality monitoring data from Sisagua. This decision was based on the percentage of censored data in the data sets, the presence of multiple distinct LD and LQ, and on the fact that mostly of data sets were not approximately normally distributed (BOLKS *et al*., 2014; HELSEL, 2012; LEE and HELSEL, 2007). Additionally, descriptive statistics were also obtained in this work through the substitution method. Results from these two methods were compared through the Mann-Whitney test (MANN and WHITNEY, 1947) at a 95% confidence level.

Considering the data sets for 79 chemical substances obtained in Sisagua database, the following variables were selected for discussion in this work: (i)Pesticides: pendimethalin and terbufos; (ii) Inorganic substances: uranium and nitrate; (iii)Organic compounds: 1,2-dichloroethene and dichloromethane; (iv)

Organoleptic compounds: 1,2-dichlorobenzene and total hardness; and (v)Disinfection by-products: chlorite and trihalomethanes.

These parameters were selected because they presented the minimum and maximum percentages of censored data in raw water sources for each chemical category (Supplementary Material). Results regarding to these substances were presented in more detail in order to discuss the applicability of the developed method for data sets with different censoring levels.

Fitdistrplus statistical package (DELIGNETTE-MULLER *et al*., 2018) was used to assess the fit to the normal distribution. This package has specific functions to deal with censored data. In this work, the distribution parameters were estimated by maximum likelihood estimators by the fitdistrplus package. The estimation of descriptive statistics, including means (and their 95% upper and lower confidence limits), standard deviations, percentiles (and their 95% upper and lower confidence limits) with the Kaplan-Meier (KM) approach was made using NADA statistical package (Nondetects and data analysis for environmental data) (LEE, 2017). Both fitdistrplus and NADA packages are available in R software.

Descriptive statistics were also performed with the substitution method (LIM/2) applying one-half of LD and LQ concentrations since this technique is frequently used in environmental studies (HELSEL, 2006; SABINO *et al*. 2014). These analyses were made using MS Excel (Microsoft Office).

3 RESULTS AND DISCUSSION

Figure 5 depicts the spatial distribution of data declared by Brazilian service providers of water supply to Sisagua between 2014 and 2018. Since information fromindividual water supply solutions are not available in Sisagua, there is a low data coverage in North and Northeast regions (24% and 28% of the municipalities in 2018, respectively), where such alternatives are widely used (INSTITUTO TRATA BRASIL, 2018).

Figure 5 - Brazilian municipalities with monitoring data of chemical variables recorded in Sisagua between 2014 and 2018 and percentage of Brazilian municipalities with data available in Sisagua in the Brazilian macroregions

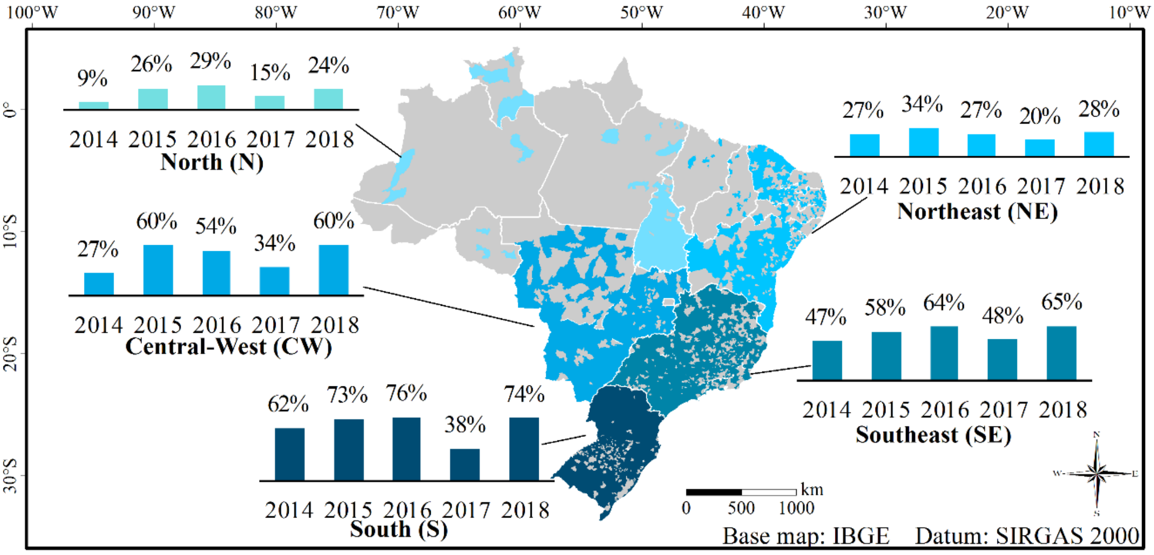


Figure 5 also suggests the non-diligence of Brazilian service providers of water supplyin recording information generated by water quality control in the Sisagua database. According to Figure 5, the percentage of municipalities declaring information in Sisagua increased between 2014 and 2015 in Northeast, and Central-West regions and between 2014 and 2016 in North, South, and Southeast regions. In the other years, these percentages werepractically constant, a fact that was also observed by Barbosa *et al*. (2015) for the period from 2007 to 2010. Despite this, Oliveira Júnior *et al*. (2019) reported an increasing trend in the aggregated information on drinking water supply (type of water supply and water quality monitoring) in Sisagua between 2014 and 2017 in terms ofthe percentage of served population. This fact signals a gradual expansion in the registration of basic information about the Brazilian drinking water supply systems in Sisagua, although increases in the insertion of monitoring data over the years are not so expressive.

3.1 LD and LQ of analytical methods used in Brazil

Table 1 shows the ranges of LDs and LQs of the analytical methods used in Brazil, as well as the percentages ofoutliers of LDs and LQs that were excluded from the datasets, and the RVLD and RVLQ establishedaccording to the methodological step 2.

Table 1 - Analytical limits of detection (LDs) and quantification (LQs) used in Brazil and determination of reference values of limits of detection (RVLDs) and quantification (RVLQs) usedinthe consistency analysis of Sisagua database.

	LD				LQ			
Chemical parameter	**N**	**% outliers**	**Range**	**RVLD**	**N**	**% outliers**	**Range**	**RVLQ**
Pendimethalin (μg.L^{-1})	5541	2.65	0.000001-10.0	10.0	27,351	1.91	0.00001-10.0	10.0
Terbufos (μg.L^{-1})	2921	3.93	0.000001-1.20	1.20	24,121	0.19	0.00002-2.00	1.20*
Uranium (mg.L^{-1})	5256	0.70	0.00001-0.04	0.03*	33,802	0.82	0.000007-0.04	0.03*
Nitrate (mg.L^{-1})	9251	8.83	0.0001-0.74	0.74	33,669	1.79	0.00005-5.35	5.35
1,2-Dichloroethene (μg.L^{-1})	3942	4.49	0.00008-13.3	13.3	29,122	0.04	0.000004-50.0	50.0
Dichloromethane (μg.L^{-1})	9419	1.33	0.00002-11.0	11.0	37,584	1.34	0.000003-13.1	13.1
1,2-Dichlorobenzene (mg.L^{-1})	3652	4.76	0.000001-0.01	0.01	23,661	3.06	0.000001-2.00	0.01*
Total hardness (mg.L^{-1})	4953	6.50	0.00016-52.0	52.0	20,289	1.59	0.0001-527.2	500.0*
Chlorite (mg.L^{-1})	65	0.06	0.0002-0.04	0.04	296	0.34	0.003-0.20	0.20
Trihalomethanes (mg.L^{-1})	7461	4.01	0.000012-0.43	0.10*	29,100	0.05	0.000001-1.02	0.10*

Legend: N: amount of data *The maximum value of the range of occurrence was higher than the maximum allowed value (MAV) of the Brazilian drinking water standard. Due to this, the MAV was adopted as the reference value. The MAV considered in this work were those established in the Consolidation Ordinance MS nº 5/2017.

One of the consistency problems identified in establishing RVLD and/or RVLQ was maximum values of LDs and/or LQs for a given chemical variable higher than the MAV established for this variable according to the Brazilian drinking water standard even after outliers exclusion. In these cases, the MAV had to be used as RVLD and/or RVLQ, considering the fact that the monitoring is carried out to check the compliance with the MAV. For this reason, RVLDs and RVLQs established for uranium and trihalomethanes were equal to their MAVs. Similarly, the MAVs for terbufos, uranium, 1,2-dichlorobenzene, total hardness, and trihalomethanes were used in establishing their RVLQs. For the other parameters presented in Table 1, the RVLDs and RVLQs were determined based on the maximum values of the reported LDs and LQs, respectively. It is also worthwhile to mention the low percentage of outliers of LDs and LQs identified in Sisagua, especially for the data sets regarding to LDs (1.1% for LQ and 3.7% for LD on average).

Additionally, Table 2 shows the MAVs considered in this work (according to the Consolidation Ordinance MS nº 5/2017), the modes obtained for LDs and LQs, as well as LDs and LQs reported in literature for each parameter. No reference values were found in the literature for LDs and LQs of analytical methods for dichloromethane and total hardness. The same occurred for LQs of uranium, 1,2-dichloroethene, 1,2-dichlorobenzene, chlorite, and trihalomethanes.

As shown in Table 1 and Table 2, pendimethalin, terbufos, uranium, nitrate, 1,2-dichloroethene, and chlorite had RVLDs higher than the LD reported in literature. However, chlorite had an equal value for the mode of LD reported in Brazil and for the LD described by USEPA (1993) (Table 2). Furthermore, the modes of LD of nitrate from Sisagua data remained at the same order of magnitude of values from literature (Table 2).

The compound 1,2-dichlorobenzene presented a RVLD (Table 1) lower than the LD found in literature (Table 2). However, the LD described by USEPA (1984) is higher than the MAV established in Consolidation Ordinance MS nº 5/2017. The RVLD established for trihalomethanes (Table 1) remained in the range reported by the World Health Organization (WHO 2004) (Table 2). Additionally, the RVLQ for pendimethalin, terbufos, and nitrate were lower than the LQ reported in literature (Table 1, Table 2). Nevertheless, the modes obtained for LQ of the analytical methods for these variables were comparable to the values reported in literature (Table 2). The compatibility between the RVLD/RVLQ or the mode of LD/LQ with values described in international literature demonstrates the consistency of the data regarding LD and LQ declared in the Sisagua and the proposed methodology.

Table 2 - Maximum allowed values (MAV) of the Consolidation Ordinance MS nº 5/2017 (Brazilian drinking water standard until 2021), the modes of the analytical limits of detection (LD) and quantification (LQ) reported in Sisagua, and LD and LQ reported in international literature

		LD		**LQ**	
Chemical parameter	**MAV** [(a)]	**Mode**	**Literature**	**Mode**	**Literature**
Pendimethalin (μg.L^{-1})	20.0	0.88	0.004 [(b)]	0.01	0.02 [(b)]
Terbufos (μg.L^{-1})	1.20	0.10	0.02 [(b)]	0.10	0.5 [(b)]
Uranium (mg.L^{-1})	0.03	0.01	0.001 [(c)]	0.01	-
Nitrate (mg.L^{-1})	10.0	0.01	0.05 [(d)]	0.01	0.15 [(d)]
1,2-Dichloroethene (μg.L^{-1})	50.0	1.00	0.03 [(e)]	1.00	-
Dichloromethane (μg.L^{-1})	20.0	0.49	-	1.00	-
1,2-Dichlorobenzene (mg.L^{-1})	0.01	0.001	1.14 [(f)]	0.00004	-
Total hardness (mg.L^{-1})	500,0	0.50	-	5.0	-
Chlorite (mg.L^{-1})	1.00	0.01	0.01 [(g)]	0.04	-
Trihalomethanes (mg.L^{-1})	0.10	0.003	0.02 – 1.0 [(h)]	0.001	-

(a) Brasil (2017); (b) USEPA (2020); (c) MAE (2020); (d) UMCES (2020); (e) ATSDR (1999) apud WHO (2003); (f) USEPA (1984); (g) USEPA (1993); (h) WHO (2004).

3.2 Consistency analysis, data preparation, and outliers exclusion

Table 3 shows the percentages of filled/replaced and reclassified data during the solution of inconsistencies in the data sets for each chemical variable. It is worthwhile to mention the percentages of null or empty LDand/or LQ especially for parameters with high quantification percentages (Supplementary Material), such as nitrate, total hardness, and trihalomethanes.

Table 3 - Percentage of modified data during the Step 2 of the methodology developed in this work (consistency analysis, data preparation, and outliers exclusion), according to the identified inconsistency types

	Fill/replaced (%)					Reclassified (%)	
Chemical parameter	**Null or empty LD[a]**	**LD >RVLD[a]**	**Null or empty LQ[b]**	**LQ >RVLQ[b]**	**Null or empty conc.[c]**	**Conc. below the LD[c]**	**Conc. below the LQ[c]**
Pendimethalin	23.35	0.42	12.66	1.50	5.80	0.12	0.49
Terbufos	24.06	0.02	14.66	0.30	2.74	0.05	0.46
Uranium	28.08	0.10	15.94	0.79	2.74	0.15	0.93
Nitrate	59.27	1.49	34.46	1.10	2.60	0.37	1.03
1,2-Dichloroethene	26.24	0.47	17.31	0.03	1.73	0.12	0.94
Dichloromethane	25.49	0.26	17.37	2.03	2.54	0.19	0.63
1,2-Dichlorobenzene	38.65	0.46	23.52	9.21	9.97	0.02	0.32
Total hardness	81.53	0.64	57.86	0.02	0.76	0.06	0.43
Chlorite	34.89	0.00	14.08	0.00	0.00	4.69	5.67
Trihalomethanes	46.83	1.96	26.58	5.94	5.61	0.37	0.99

a: below the detection limit; b: below the quantification limit; c: quantified data in original dataset. Legend: conc.: concentration; LD: limit of detection; LQ: limit of quantification. RVLD: reference value of detection limits; RVLQ: reference value of quantification limits.

Reported LDs higher than the RVLDs occurred in the data sets of all chemical variable except for chlorite (Table 3; Table 1) and, due to this, the values of the cells with these LDs were replaced with the RLVD. In the case of uranium and trihalomethanes, replacements with the RVLD were also carried out for cells in which the reported LD exceeded the MAV (Table 1; Table 3).

For the case of LQ, filling and/or replacement of data occurred both for outliers and reporting LQs higher than MAV in data sets regarding to terbufos, uranium, 1,2-dichlorobenzene, total hardness, and trihalomethanes (Table 3). For the other variables, there were no reported LQ higher than the MAV (except for chlorite which did not show such inconsistency).

Table 3 also shows that the most frequent reclassification occurred for the quantified data with null or empty concentration, which were reclassified as not detected. The highest percentages of this inconsistency occurred for 1,2-dichlorobenzene (9.97%), pendimethalin (5.80%), and trihalomethanes (5.61%), whose cells were filled with the RVLD: 0.01 mg.L^{-1}, 10.0μg.L^{-1}, and 0.10 mg.L^{-1}, respectively (Table 1).

Reclassification of quantified data whosethe concentration reported was lower than their reported LD or LQ was more expressive (4.69% and 5.67%, respectively) for the case of chlorite (Table 3). This aspect resulted in a slightly increase in the already high percentages of censorship of data sets regarding to chlorite, which presented a mean of 94.5% of samples below the analytical limits (Supplementary Material).

In spite of the considerable number of inconsistencies identified in Sisagua database, according to Oliveira Júnior *et al*. (2019), the current version of Sisagua can receive data automatically from laboratories or information systems of service providers, the current version of Sisaguacan receive data automatically from laboratories or information systems of service providers. This function may substantially reduce the occurrence of inconsistencies resulting from typing errors or missing data. Thus, its use must be encouraged.

The percentages of outliers identified and excluded from the data sets of each chemical variable were shown in Table 4. As expected, the percentages of outliers varied spatially according to the region, sampling point and chemical variable. These outliers may be a result of several aspects such as errors in sampling, analysis and/or typing, but they also can be real concentrations from atypical events (VON SPERLING *et al*., 2020).The largest percentages of outliers were identified for terbufos in water sources of Brazilian South region, nitrate at DWTPs from Northeast, and dichloromethane in distribution systems and tap waters of Northeast region (Table 4).

High outlier levels were identified for data sets regarding to finished waters from DWTPs and samples from drinking water distribution systems and tap waters. However, according to the Brazilian drinking water standard (BRASIL, 2021), for the case of compounds that cannot be introduced water during/after treatment, monitoring of finished and tap waters can be conditioned to the occurrence of the substance in water sources and finished waters. This fact may result in data sets that represent mainly observations with high concentrations.

Table 4 - Percentages of outliers of quantified data from different macroregions and sampling points of the Brazilian drinking water supply systems

Sampling point	**Chemical parameter**	**Macroregion**					
		BR	**SE**	**S**	**NE**	**CW**	**N**
Raw water catchments (RW) (%)	**Pendimethalin**	4.46	4.63	0.00	0.00	1.26	0.00
	Terbufos	0.01	0.01	58.85	6.25	0.00	0.00
	Uranium	8.02	8.31	4.21	0.00	1.98	0.00
	Nitrate	12.63	12.62	14.75	14.82	27.22	0.00
	1,2-Dichloroethene	2.84	2.73	3.25	0.00	1.56	0.00
	Dichloromethane	4.85	9.98	0.96	0.00	2.59	0.00
	1,2-Dichlorobenzene	6.93	6.69	6.87	10.00	6.31	0.00
	Total hardness	3.45	3.39	2.37	4.57	0.56	0.00
	Chlorite	0.91	0.00	0.00	0.00	0.00	0.00
	Trihalomethanes	5.58	5.53	5.66	22.22	3.24	0.00

Table 4 - Continue

Table 4 - Conclusion

Drinking water treatment plants (DWTP) (%)	**Pendimethalin**	8.24	8.33	0.00	0.00	1.89	0.00
	Terbufos	13.08	0.03	0.03	3.07	1.47	0.00
	Uranium	9.35	14.23	3.98	3.06	3.42	0.31
	Nitrate	15.50	14.16	10.79	44.96	9.29	23.52
	1,2-Dichloroethene	1.63	18.06	4.94	3.15	0.00	0.00
	Dichloromethane	0.88	4.37	0.99	2.61	2.18	0.00
	1,2-Dichlorobenzene	5.42	10.32	10.45	0.06	1.95	0.52
	Total hardness	3.03	3.05	2.31	12.22	2.94	1.56
	Chlorite	0.00	0.00	0.00	0.00	0.00	0.00
	Trihalomethanes	16.16	13.55	14.31	4.74	17.23	5.68
Drinking water distribution systems and tap waters (DW) (%)	**Pendimethalin**	5.68	4.97	0.18	8.75	1.71	0.00
	Terbufos	0.10	0.10	0.00	20.00	1.24	0.00
	Uranium	6.92	18.21	16.22	17.89	0.74	0.00
	Nitrate	18.21	7.37	8.96	23.42	2.97	22.22
	1,2-Dichloroethene	7.58	7.87	8.50	1.98	0.55	0.00
	Dichloromethane	7.23	6.09	3.74	33.00	5.07	0.00
	1,2-Dichlorobenzene	16.44	19.39	8.00	4.08	2.3	0.00
	Total hardness	3.44	1.34	2.07	7.26	2.26	0.00
	Chlorite	0.00	0.00	0.00	0.00	0.00	0.00
	Trihalomethanes	9.95	5.96	13.38	8.25	11.12	6.39

3.3 Descriptive analysis of Sisagua data

The percentage of censored data varied between 0% (as occurred in data sets regarding to 1,2-dichlorobenzene and total hardness) and 100% (as in data sets of pendimethalin, terbufos, 1,2-dichoroethene, and chlorite) (Supplementary Material).

As expected, the Kaplan-Meier method was able to estimate means for all data sets that had uncensored observations (GILLESPIE *et al*., 2010). In this work, this corresponded to data sets with censoring levels of up to 99.8%, such as the case of terbufos in finished waters samples from DWTPs from Brazilian North region (Table S.2). This data set consists in two quantified observations (0.05 µg.L^{-}

^{1} and 1.0 μg.L^{-1}) in an array of 938 samples with left-censored data ranging from 0.001 μg.L^{-1} to 1.2 μg.L^{-1} and the mean estimated by the Kaplan-Meier method was equal to 0.05 μg.L^{-1}). It is worthwhile to mention that if a quantified data is equal to a censored observation, the value estimated by the Kaplan-Meier method for the censored data will be lower than the quantified concentration (GILLESPIE *et al*., 2010).

On the other hand, the capacity to estimate percentiles depends on some factors. For estimating probabilities using the Kaplan-Meier method, quantified and censored data are ordered together from the smallest to the largest, but the method does not estimate percentiles for censored data (FLIKKEMA, 2016; GILLESPIE *et al*., 2010; HELSEL, 2005). In this sense, a value for a given percentile only will be estimated if there is an uncensored observation near to the position of the percentile of interest. Furthermore, it is also important to consider that even if there are censored observations larger than all the quantified, the Kaplan-Meier method will distribute the probability between the values smaller than maximum quantified concentration (GILLESPIE *et al*., 2010).

The influence of these aspects can be observed for all chemical variables (Supplementary Material). Kaplan-Meier method was able (or not) to estimate values for different percentiles depending on the arrangement of the censored and uncensored data in the probability function as well as the density of censored data near to the percentiles.

Examples were the data sets for pendimethalin in drinking water distribution systems and tap waters from the South region of Brazil and for terbufos in finished waters from DWTPs of the Northeast region. The first data had the maximum quantified value near to the 75% percentile, and, due to this, percentiles higher than the 75% could not be estimated. On the other hand, the 50% percentile could not be estimated for the second data set due to the fact that minimum quantified observation was near to the 75% percentile of the empirical cumulative distribution function obtained by the Kaplan-Meier method (Table S.2).

Percentiles were estimated for data sets with up to 99% of censored observations, which was the case of dichloromethane in water distribution systems and tap waters of the North region of Brazil (Table S.6). However, in this case, the algorithm was not able to calculate a confidence interval for such estimates.

Furthermore, the intervals between the 95% upper and lower confidence limits (95% LCL and 95% UCL) obtained for the percentiles (which are estimated based on the position of the quantified observations) indicate that the Kaplan-Meier method produces better estimates for data sets with small censoring levels. This was the case of the data sets regarding total hardness (Table S.8), which had censoring levels varying between 0 and 22.5% and presented the smallest intervals between the 95% LCLs and 95% UCLs.

Descriptive statistics obtained by the Kaplan-Meier (KM) method, proposed in this study, were compared to the substitution method using one half of the analytical limits (LIM/2)(Figure 6). These results are regarding to total hardness, trihalomethanes, uranium, terbufos, and 1,2-dichloroethene in water sources (RW), finished waters sampled at DWTPs, or samples from drinking water distribution systems and tap waters(DW), whose percentages of censorship were 11.06%, 47.20%, 64.45%, 70.99% and 88.24%, respectively (Supplementary Material).

As can be seen in Figure 6, the discrepancy between descriptive parameters obtained by each method increased with rising thecensoring level. In all cases, the estimation of descriptive parameters by the Kaplan-Meier method led to lower concentrations compared to the substitution method, as reported by Gillespie *et al*. (2010).

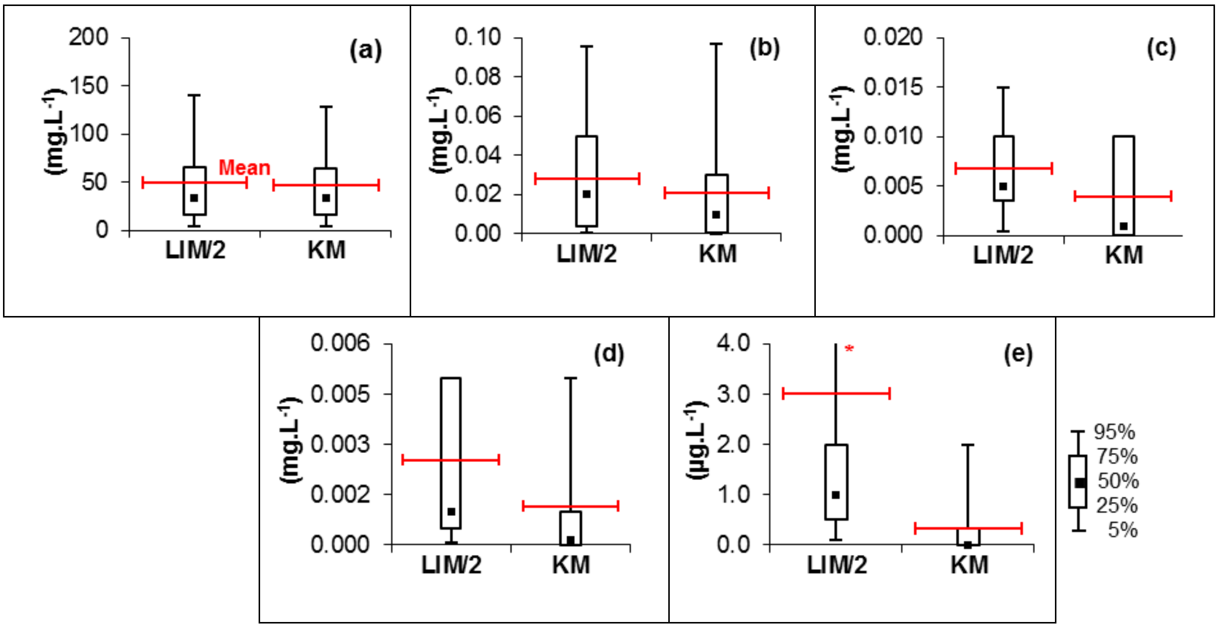
As can be seen in Figure 6, the discrepancy between descriptive parameters obtained by each method increased with rising thecensoring level. In all cases,

the estimation of descriptive parameters by the Kaplan-Meier method led to lower concentrations compared to the substitution method, as reported by Gillespie *et al*. (2010).

No statistically significant difference at a 95% confidence level (p-value = 0.374) was evidenced comparing results regarding the analyses of total hardness in samples from DWTPs of Brazil (percentage of censorship of 11.06%) obtained by the two methods (Figure 6a). Percentiles estimated by the substitution method were equal (5% percentile) or slightly (1.1 times) higher than those estimated using the Kaplan-Meier method (95% percentile).In this case, the mean obtained by the substitution method also exceeded by 5.5% themean obtained by the Kaplan-Meier method.

On the other hand, for 1,2-dichloroethene in Brazilian water sources (88.24% of censoring level), the substitution method estimated descriptive parameters 6.2 (75% percentile) to 500 (median) times higher than the Kaplan-Meier method. The mean estimated through the Kaplan-Meier method was 9.0 times lower than the obtained by the substitution method. Furthermore, the Wilcoxon-Mann-Whitney test evidenced a statistically significant difference at a 95% confidence level between the descriptive parameters estimated through the two methods (p-value = 0.0269).

Figure 6 - Comparative analysis between the descriptive statistics obtained by substituting censored data by one-half of their limits (LIM/2) and the Kaplan-Meier method (KM);



6a: Total hardness in finished waters from Brazilian drinking water treatment plants (DWTP); 6b: Trihalomethanes in Brazilian drinking water distribution systems and tap waters (DW); 6c: Uranium in Brazilian drinking water distribution systems and tap waters (DW); 6d: 1,2-Dichlorobenzene inBrazilian drinking water treatment plants (DWTP); 6e: 1,2-Dichloroethene in Brazilian raw water sources (RW). * 95% percentile = 25 mg.L^{-1}

With the advancement of softwares and analytical tools, the USEPA has been encouraging the use of more sophisticated statistical methods, no longer recommending substitution methods in its most recent technical guides (USEPA, 2010; USEPA, 2016).

Previous researches reported that substitution methods may perform worse even at small censoring levels (5~11%) (SINGH *et al*., 2006; LEITH *et al*., 2010). Several studies also pointed out that the Kaplan-Meier method can perform better than the substitution methods (ANTWEILER and TAYLOR, 2008; SHE, 1997; HEWETT and GANSER, 2007; FLIKKEMA, 2016; LEITH *et al*., 2010; GILLESPIE *et al*., 2010), being less susceptible to biased estimates. Findings of this work also reinforce this aspect. The massive presence of censored data, often with values higher than the quantified data, raised the values estimated for the statistical parameters, especially for the case of 1,2-dichloroethene.

Literature suggest that the Kaplan-Meier method performs better than the other methods especially for data sets with censoring levels ≤50%~70% (HELSEL, 2005; ANTWEILER and TAYLOR, 2008; SINGH *et al*., 2006). In spite of this, findings of Gillespie *et al*. (2010) show that the Kaplan-Meier method may produce informative estimates even for high censoring levels (up to 97%), working better than the substitution method.

However, uncertainties associated with the use of statistical methods in censoring levels greater than 80%, which are common in environmental data, are also highlighted inliterature (HELSEL, 2012; BOLKS *et al*., 2014; CHRISTOFARO and LEÃO, 2014). In these cases, Helsel (2012) recommended to report only the highest percentiles or even the maximum quantified concentration.

Other authors conclude that the substitution method can be a good choice for lognormal distributions with large sample sizes or to estimate means (HUYNH *et al*., 2014; HEWETT and GANSER, 2007; MIKKONEN *et al*., 2018). Nonetheless, it is also important to consider that when using a LIM/2 substitution approach it is assumed that the censored data follow a uniform distribution (GILLESPIE *et al*., 2010; CHRISTOFARO and LEÃO, 2014). In this sense, given the fact that environmental data often do not follow a known probability density function (e. g. uniform, normal, lognormal), a nonparametric approach for estimating descriptive statistics can be a safe decision (GILLESPIE *et al*., 2010), especially thinking in a consistent systematic procedure for environmental data analysis.

4 CONCLUSIONS AND RECOMMENDATIONS

This work proposed a method for statistical analysis of environmental data which was successfully applied for water quality monitoring data from Brazilian drinking water supply systems obtained through the Brazilian Drinking Water

Quality Surveillance Information System (Sisagua). A list of possible consistency problems present in Sisagua as well as strategies for their solution were provided. Data reported without their limit of detection (LD) and/or limit of quantification (LQ) were the main inconsistency observed in Sisagua. The outlier levels for censored and uncensored data varied for regions, sampling points and chemical variables. For the case of quantified data, drinking water samples had outlier levels higher than those from water sources. However, concentrations from drinking water samples may be biased high, since they can be monitored in a high frequency only if high concentrations are observed in water sources. LDs and LQs obtained from Sisagua were comparable to values reported in literature, evidencing the robustness of the proposed methodology. Descriptive parameters obtained by the substitution method were larger than the estimated by Kaplan-Meier method and demonstrated a high statistical bias. Results reaffirm the applicability of the Kaplan-Meier method when compared with the substitution method and reinforce the need to use appropriate method for data analysis, especially for the case of environmental data which are often not normally distributed and may present multiple distinct censored observations. Further studies can evaluate the performance of the parametric and robust methods for the descriptive analysis of the Sisagua database, as some authors mentioned the good performance of such methods in high censoring levels.

ACKNOWLEDGEMENTS

The authors thank the Pro-Rectory of Postgraduate and Research (PROPP) from Federal University of Juiz de Fora (UFJF) and the Postgraduate Programme in Civil Engineering (PEC – UFJF).

REFERENCES

ANALYTICAL METHODS COMMITTEE. Using the Grubbs and Cochran tests to identify outliers. **Analytical Methods**, v. 7, n. 19, p. 7948-7950, 2015.

ANTWEILER, R. C.; TAYLOR, H. E.Evaluation of Statistical Treatments of Left-Censored Environmental Data using Coincident Uncensored Data Sets: I. **Summary Statistics. Environmental Science and Technology**, v. 42, p. 3732-3738, 2008.

ATSDR - Agency for Toxic Substances & Disease Registry. Toxicological profile for 1,2-dichloroethane. **Atlanta: US Department of Health and Human Services**, 1999.

BARBOSA, A. M. C.; SOLANO, M. L. M.; UMBUZEIRO, G. A. Pesticides in drinking water – the Brazilian monitoring program. **Frontiers in Public Health**, v. 3, p. 246, 2015.

BRASIL. **Ministério da Saúde**. Manual de procedimentos de entrada de dados do sistema de informação de vigilância da qualidade da água para consumo humano (Sisagua). 2016. Available from: https://www.saude.go.gov.br/images/imagens_migradas/upload/arquivos/2016-03/manual-de-procedimentos-de-entrada-de-dados-do-sisagua-08-01-2016-1.pdf Accessed May 2022.

BRASIL. **Ministério da Saúde**. Portaria de Consolidação nº 5, de 28 de setembro de 2017 – ANEXO XX. Diário Oficial [da] República Federativa do Brasil, Poder Executivo, Brasília, DF, 03 out. 2017. Seção 1, p. 360.

BRASIL. **Ministério da Saúde**. Portaria GM/MS nº 888, de 4 de maio de 2021. Diário Oficial [da] República Federativa do Brasil, Poder Executivo, Brasília, DF, 7 mai. 2021. Seção 1, p. 127.

BRASIL. **Ministério da Saúde**. Sisagua. Available from: http://sisagua.saude.gov.br/sisagua/paginaExterna.jsf. Accessed October 2020.

BOLKS, A.; DeWIRE, A.; HARCUM, J. B. Baseline assessment of left-censored environmental data using R. **USEPA**. Tech Notes 10. 2014. Available from: https://www.epa.gov/sites/production/files/2016-05/documents/tech_notes_10_jun2014_r.pdf. Accessed October 2020.

CASSANEGO, M. B. B.; DROSTE, A. Avaliação do padrão espacial da qualidade da água de um rio no Sul do Brasil por meio da análise multivariada de indicadores biológico e químicos, **Brazilian Journal of Biology**, v. 77, n. 1, p. 118-126, 2017.

CETESB - Companhia de Tecnologia de Saneamento Ambiental**.** Relatório de estabelecimento de valores orientadores para solos e águas subterrâneas no estado de São Paulo. São Paulo**: CETESB**, 2001.

CHRISTOFARO, C.; LEÃO, M. D. Tratamento de dados censurados em estudos ambientais. **Química Nova**, v. 37, n. 1, p. 104-110, 2014.

DALZOCHIO, T. *et al*. Water quality parameters, biomarkers and metal bioaccumulation in native fish captured in the Ilha River, Southern Brazil. **Chemosphere**, 189, p. 609-618, 2017.

DELIGNETTE-MULLER, M. L.; DUTANG, C. fitdistrplus: An R Package for Fitting Distributions. 2018. Available from: https://cran.r-project.org/web/packages/fitdistrplus/vignettes/paper2JSS.pdf. Accessed October 2020.

FLIKKEMA, R. M. Statistical methodology for data with multiple limits of detection. **Tese (Doutorado).** Western Michigan University, Michigan, 2016.

GEORGE, B.J.; GAINS-GERMAIN, L.; BROMS, K.; BLACK, K.; FURMAN, M.; HAYS, M.D.; THOMAS, K.W.; SIMMONS, J.E. Censoring Trace-Level Environmental Data: Statistical Analysis Considerations to Limit Bias. **Environmental Science and Technology**, v. 55, p. 3786-3795, 2021.

GILLESPIE, B. W.; CHEN, Q.; REICHERT, H.; FRANZBLAU, A.; HEDGEMAN, E.; LEPKOWSKI, J.; ADRIAENS, P.; DEMOND, A.; LUKSEMBURG, W.; GARABRANT, D. H. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. **Epidemiology**, v. 21, n. 4, S64–S70, 2010.

GRUBBS, F. Procedures for detecting outlying observations in samples. **Technometrics**, p. 11, n. 1, p. 1-21, 1979.

HELSEL, D. R. Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. **Chemosphere**, v. 65, p. 2434-2439, 2006.

HELSEL, D. R.; HIRSCH, R. M.; Statistical Methods in Water Resources. **Washington: U.S. Geological Survey**. 2002.

HELSEL, D. R. More than obvious: Better methods for interpreting nondetect data. **Environmental Science and Technology**, v. 39, n. 20, p. 419-423, 2005.

HELSEL, D. R. Nondetects and data analysis statistics for censored environmental data. **New York: John Wiley and Sons**. 2004.

HELSEL, D. R. Statistics for Censored Environmental Data Using Minitab and R. 2 ed. **New York: John Wiley and Sons**. 2012.

HEWETT, P.; GANSER, G. H. A comparison of several methods for analyzing censored data. **Annals of Occupational Hygiene**, v. 51, n. 7, p. 611-632, 2007.

HUYNH, T.; RAMACHANDRAN, G.; BANERJEE, S.; MONTEIRO, J.; STENZEL, M.; SANDLER, D. P.; ENGEL, L. S.; KWOK, R. K.; BLAIR, A.; STEWART, P. A. Comparison of Methods for Analyzing Left-Censored Occupational Exposure Data.**The Annals of Occupational Hygiene**, v. 58, n. 9, p. 1126–1142, 2014.

INSTITUTO TRATA BRASIL. Acesso à água nas regiões norte e nordeste do Brasil: desafios e perspectivas. 2018. Available From: https://tratabrasil.org.br/images/estudos/acesso-agua/tratabrasil_relatorio_v3_A.pdf. Accessed May 2022.

JEONG, J.; PARK, E.; HAN, W. S.; KIM, K.; CHOUNG, S.; CHUNG, I. M. Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends. **Journal of Hydrology**, v. 548, p. 135-144, 2017.

KAPLAN, E. L.; MEIER, O. Nonparametric Estimation from Incomplete Observations. **Journal of the American Statistical Association**, v. 53, p. 457-481, 1958.

LEE, L.; HELSEL, D. Statistical analysis of water-quality data containing multiple detection limits II: S-language software for nonparametric distribution modeling and hypothesis testing. **Computers and Geosciences**, v. 33, n. 5, p. 696-704, 2007.

LEE, L. Package 'NADA'. 2017. Available from: https://cran.r-project.org/web/packages/NADA/NADA.pdf. Accessed October 2020.

LEITH, K. F.; BOWERMAN, W. W.; WIERDA, M. R.; BEST, D. A.; GRUBB, T. G.; SIKARSKE, J. G. 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**, v. 80, n. 1, p. 7-12, 2010.

MAE. Department of Municipal Affairs and Environment. Government of Newfoundland and Labrador. Drinking Water Quality Database - Detection Limits. Available from: https://www.gov.nl.ca/ecc/files/waterres-quality-drinkingwater-pdf-detect-limits.pdf. Accessed May 2022.

MCGRORY E.; HOLIAN E.; MORRISON L. Assessment of groundwater processes using censored data analysis incorporating non-detect chemical, physical, and biological data. **Journal of Contaminant Hydrology**, v. 235, p. 103706, 2020.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **Annals of Mathematical Statistics**, v. 18, n. 1, p. 50-60, 1947.

MARIMON, M. P. C.; ROISENBERG, A.; SUHOGUSOFF, A. V.; VIERO, A. P. Hydrogeochemistry and statistical analysis applied to understand fluoride provenance in the Guarani Aquifer System, Southern Brazil. **Environmental Geochemistry and Health**, v. 35, n. 3, p. 391–403, 2013.

MELO GURGEL, P. et al. Ecotoxicological water assessment of an estuarine river from the Brazilian Northeast, potentially affected by industrial wastewater discharge. **Science of the Total Environment**, v. 572, p. 324-332, 2016.

MIKKONEN, H. G. et al. Evaluation of methods for managing censored results when calculating the geometric mean. **Chemosphere**, v. 191, p. 412-416, 2018.

OLIVEIRA JÚNIOR, A. et al. Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano (Sisagua): características, evolução e aplicabilidade**. Epidemiologia e Serviços de Saúde**, v. 28, n. 1, n. p., 2019.

SABINO, C. V. S.; LAGE, L. V.; ALMEIDA, K. C. B. Uso de métodos estatísticos robustos na análise ambiental, **Engenharia Sanitária e Ambiental**, v. 19 (spe), p. 87-94, 2014.

SHE, N. Analyzing censored water quality data using a non-parametric approach. **Journal of the American Water Resources Association**, v. 33, n. 3, p. 615-624, 1997.

SINGH, A.; MAICHLE, R.; LEE, S. E. On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations. **Washington, DC: USEPA**, 2006 (EPA/600/R-06/022).

STAPLES, C. *et al*. Distributions of concentrations of bisphenol A in North American and European surface waters and sediments determined from 19 years of monitoring data. **Chemosphere**, v. 201, p. 448-458, 2018.

STATSOFT. **Statistica (data analysis software system**), version 8.0. 2007.

TUKEY, J. W. Exploratory Data Analysis**. Massachusetts: Addison-Wesley**, 1977.

UMCES – University of Maryland Center for Environmental Science. Standard Operating Procedure for Determination of Total Dissolved Nitrogen (TDN) and Total Nitrogen (TN) in Fresh/Estuarine/Coastal Waters Using Alkaline Persulfate Digestion of Nitrogen to Nitrate and Measured Using Cadmium Reduction (References EPA 353.2, Standard Methods #4500-N C, 4500-NO3 F). Available from: https://www.umces.edu/sites/default/files/TDN%20Nitrate%20Method%202018-1_1.pdf. Accessed October 2020.

URVOY, M.; AUTRUSSEAU, F. Application of Grubbs' test for outliers do the detection of watermarks. *In:* ACM Workshop on Information Hiding and Multimedia Security, 2., 2014, Salzburg. **Proceedings... Salzburg: ACM**, 2014. p. 49-60.

USEPA – United States Environmental Protection Agency. 2018 Edition of the Drinking Water Standards and Health Advisories Tables. **Washington, DC: Office of Research and Development**, 2018.

USEPA – United States Environmental Protection Agency. EPA Method 612: Chlorinated Hydrocarbons.Cincinnati, **Ohio: Office of Research and Development**, 1984.

USEPA – United States Environmental Protection Agency. EPA Method 300.0: Determination of Inorganic Anions by Ion Chromatography. Cincinnati, **Ohio: Office of Research and Development**, 1993.

USEPA - United States Environmental Protection Agency. Pesticide Analytical Methods. Environmental Chemistry Methods (ECM). Available from: https://www.epa.gov/pesticide-analytical-methods/environmental-chemistry-methods-ecm-index-d. Accessed October 2020.

USEPA - United States Environmental Protection Agency. ProUCL Version 4.00.05 Technical Guide (Draft). **Washington, DC: Office of Research and Development**, 2010.

USEPA - United States Environmental Protection Agency. ProUCL Version 5.1 Technical Guide. Washington, **DC: Office of Research and Development**, 2016.

VON SPERLING, M.; VERBYLA, M. E.; OLIVEIRA, S. M. A. C. Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners. London: **IWA Publishing**, 2020.

WHO. **World Health Organization**. 1,2-Dichloroethane in Drinking-water. Background document for development of WHO Guidelines for Drinking-water Quality. 2003. Available from: https://cdn.who.int/media/docs/default-source/wash-documents/wash-chemicals/1-2-dichloroethane.pdf?sfvrsn=d7e6c0e0_4. Accessed May 2022.

WHO. **World Health Organization**. Trihalomethanes in Drinking-water. Background document for development of WHO Guidelines for Drinking-water Quality. 2004. Available from: https://cdn.who.int/media/docs/default-source/wash-documents/wash-chemicals/trihalomethanes.pdf?sfvrsn=3d3a90e3_4. Accessed May 2022.

Authorship contributions

1 – **Fernanda Bento Rosa Gomes** (Corresponding Author)
Bachelor of Environmental and Sanitary Engineering, Master's student at the Graduate Program in Civil Engineering of the Federal University of Juiz de Fora
https://orcid.org/0000-0003-1262-5238 • fernanda.bento@engenharia.ufjf.br
Contribution: Conceptualization, Methodology, Formal Analysis, Validation, Writing - Original Draft, Visualization

2 – **Guilherme Bento Nicolau**
Bachelor of Environmental and Sanitary Engineering, Master's student at the Graduate Program in Civil Engineering of the Federal University of Juiz de Fora
https://orcid.org/0000-0002-9320-0196 • guilherme.nicolau@engenharia.ufjf.br
Contribution: Conceptualization, Methodology, Formal Analysis

3 – **Emanuel Manfred Freire Bran**
Doctor of Sanitation, Environment, and Water Resources, Professor at the Graduate Program in Civil Engineering of the Federal University of Juiz de Fora
http://orcid.org/0000-0002-9009-1940 • emanuel.brandt@ufjf.edu.br
Contribution: Conceptualization, Methodology, Supervision, Writing - Review & Editing

4 – **Samuel Rodrigues Castro**
Doctor of Sanitation, Environment, and Water Resources, Professor at theDepartment of Environmental and Sanitary Engineering and at the Graduate Program in Civil Engineering of the Federal University of Juiz de Fora
https://orcid.org/0000-0003-4053-7040 • samuel.castro@ufjf.edu.br
Contribution: Conceptualization, Methodology, Supervision, Writing - Review & Editing

5 – **Renata de Oliveira Pereira**
Doctor of Hydraulics and Sanitation, Professor at the Department of Environmental and Sanitary Engineering and at the Graduate Program in Civil Engineering of the Federal University of Juiz de Fora
https://orcid.org/0000-0002-3414-7292 • renata.pereira@ufjf.edu.br
Contribution: Conceptualization, Methodology, Supervision, Writing - Review & Editing

6 – **Taciane de Oliveira Gomes de Assunção**
Undergraduate student at the Department of Environmental and Sanitary Engineering of the Federal University of Juiz de Fora
https://orcid.org/0000-0002-8342-4727 • taciane.assuncao@engenharia.ufjf.br
Contribution: Conceptualization, Methodology, Formal Analysis, Writing - Original Draft:

7 – **Pedro Fialho Cordeiro**
Researcher at the SENAI FIEMGInnovation and Technology Center, Doctoral student at the Graduate Program in Environmental Systems Analysis and Modeling of the Federal University of Minas Gerais
https://orcid.org/0000-0001-8197-7988 • pedrofialhoc@gmail.com
Contribution: Conceptualization, Methodology

How to quote this article

GOMES, F. B.; *et al*. Occurrence of chemical substances in water supply systems of Brazil: a nonparametric approach for statistical analysis of sisagua data**. Ciência e Natura**, Santa Maria, v. 44, e24, 2022. Available in: https://doi.org/10.5902/2179460X63368.