A comparison of parameters below the limit of detection in geochemical analyses by substitution methods

Primerjava ocenitev parametrov pod mejo določljivosti pri geokemičnih analizah z metodo nadomeščanja

Timotej Verbovšek[1, *]

[1]University of Ljubljana, Faculty of Natural Science and Engineering, Department for Geology, Aškerčeva 12, SI-1000 Ljubljana, Slovenia

*Corresponding author. E-mail: timotej.verbovsek@ntf.uni-lj.si

Received: November 9, 2011 **Accepted:** November 25, 2011

Abstract: Paper focuses on the analysis of geochemical data with values below the limit of detection (LOD). Such values are treated as text and are difficult to use in further calculations of mean, standard deviation and other statistical parameters. To estimate several methods for substitution of values below the LOD with fractions of LOD (zero, LOD/2, LOD/√2, LOD and *no data* values), a large dataset of generated values with normal and lognormal distributions was tested for different percent of censoring from 1 % to 50 %, plus the censored data of five selected geochemical parameters. Results indicate that the best substitution method is by LOD/√2, as it produces the smallest errors. The greatest errors are found for substitution methods with zero or *no data*. This is valid both for normally and lognormally distributed data. Median is not affected by most methods for censoring level below 50 %. For real geochemical parameters, the interpretation is more complex. For datasets with low amount of censoring (NO_3, O_2), the errors are small. For others (Sr, F, Mn) the errors are larger, as several LODs exist for each parameter and the LOD is sometimes larger than the mean value.

Izvleček: V prispevku je predstavljena in analizirana problematika geokemičnih podatkov pod mejo določljivosti (MD). Ti so obravnavani kot tekst in se s težavo uporabljajo v nadaljnjih statističnih izračunih (povprečje, standardni odklon ipd). Primerjane so bile različne metode nadomeščanja vrednosti pod MD s petimi deleži: nič, MD/2,

Original scientific paper

MD/√2, MD in vrednosti *brez podatkov*. Sprva je bil analiziran velik nabor generiranih podatkov idealne normalne in lognormalne porazdelitve za različne stopnje okrnjenosti podatkov (od 1 % do 50 %), nato pa še pet izbranih geokemičnih parametrov. Rezultati kažejo, da je najboljše uporabiti metodo nadomeščanja z vrednostjo MD/√2, ker daje najmanjše napake, največje napake pa dajeta metodi nadomeščanja z nič ali *brez podatkov*. To velja tako za normalno in lognormalno porazdeljene podatke. Na mediano ne vpliva večina metod, če je okrnjenih manj kot 50 % podatkov. Za izmerjene geokemične parametre je interpretacija bolj zapletena. Za parametre z manjšim deležem okrnjenosti (NO_3, O_2) so napake majhne, za druge (Sr, F, Mn) pa večje zaradi različnih mej določljivosti za vsak parameter in nekaterih vrednosti MD, večjih od povprečja.

Key words: hydrogeochemistry, limit of detection, statistics
Ključne besede: hidrogeokemija, meja detekcije, statistika

Introduction

Parameters, calculated from geochemical analyses, often lie below some limit which occludes true values. Such a limit is called *limit of detection* (LOD) or *method detection limit* (MDL) and is written with a symbol "<", i.e. "<0.005 mg/L". There are many reasons for laboratories to present the values below the limit, the most obvious being the non-ability of instruments to detect the low concentrations of parameters. Signal from the analyzed parameter can be too small for the instruments to discriminate it from the background noise and several other factors can influence the laboratory to report the values below the limit of detection (Lambert et al., 1991). Limit of detection is usually defined as the level at which a measurement has a 95 % probability of being different than zero (Croghan & Egeghy, 2003), but sometimes in the reports no indication at all is given what a detection limit is (Lambert et al., 1991).

The major problem of such low reported values lies in further statistical analysis of data. First, low concentrations are reported as text values ("<0.005 mg/L") and consequently such values are not recognized as numbers in the analyses. Only in specialized databases (like AquaChem software) it is possible to enter and treat the values as the ones below the LOD. Second, the calculation of statistical moments (mean value, standard deviation ...) is problematic, as low values are truncated or *censored*. Such calculations are critical when they are used to predict the water quality and one needs to report whether the concentrations of toxic elements lie below or above some critical level.

Several methods exist to "replace" the unknown values below the LOD with such values that the committed errors are minimized when performing the statistical analyses (Chastain, 2007, Gilliom & Helsel, 1986, Glass & Gray, 2001, Gochfeld et al., 2005, Helsel, 1990, Helsel & Cohn, 1988, Helsel & Gilliom, 1986, Smith et al., 2006, Succop et al., 2004), and each method has some advantages and disadvantages:

- Values below the limit of detection are replaced with a *constant of zero (0)*. Calculated mean values are in this case lower than the real ones, as we create a set of artificially low numbers. This approach is not recommended.
- Values are replaced with the values of *limit of detection* (LOD). Consequently, the mean values are higher than the real ones. This approach is also not recommended, as both methods represent the extreme possible values of true mean.
- Values are replaced with some *fraction of LOD*. Usually, the replacement is performed with LOD/2 and LOD/√2 (Croghan & Egeghy, 2003). The error is much lower than in previous methods, and this approach is very common due to its simplicity. There is no agreement which substitution value is the correct one, for following reasons. Of a great importance is the distribution of data, as replacement with LOD/2 is by some authors (Hornung & Reed, 1990, Succop et al., 2004) recommended for normally distributed data and LOD/√2 for lognormal distribution. Another suggestion is to use the LOD/2 substitution for datasets with much censored data and LOD/√2 for datasets with relatively few data below the detection limit (Glass & Gray, 2001). Substitution with LOD/2 is used in Slovenia for statistical calculations for water quality reports in Decree on groundwater status (Uradni list RS, 25/2009).
- Values are simply *ignored* and are not included in the analysis. Values below the LOD are replaced with *no data* values. Such an approach is not recommended, as calculated mean values are always higher than the real ones (Gochfeld et al., 2005).
- Unknown values are *estimated or extrapolated* from the distribution curve or calculated from regression. These methods are known to perform best, but the distribution of data should be generally known and there is no agreement on the most suitable method. Several methods exist for the estimation (Succop et

al., 2004), and only some can be used for multiple detection limits (Helsel & Cohn, 1988). Compared to these methods, substitution with any fraction values between zero and LOD is generally not recommended, as real data do not have exactly such values (Helsel & Cohn, 1988). However, the usage of substitution methods is still permissible for data with a few censored values (Croghan & Egeghy, 2003).

- Other methods are seldom used. A possible approach is also *not to use any statistical methods at all*, but just to report the values being lower than LOD (Gochfeld et al., 2005). The last two approaches are not presented here, as they require special statistical methods and software and are therefore not comparable to substitution of LOD fractions, discussed in this paper.

The goal of this study is first to analyze the ideal normal and lognormal distribution of generated data with different amounts of censored data (from 1 % to 50 %) and to compare the errors produced in all methods. Secondly, to use the replacement methods on an actual dataset of five geochemical parameters obtained from groundwater analyses and discuss their deviations. Other authors have used some other rather uncommon distributions (bimodal lognormal, gamma and delta; (Gilliom & Helsel, 1986), but generally lognormal distribution is regarded as more realistic than normal distribution for environmental and geochemical data (Helsel, 2005, Hornung & Reed, 1990).

The novel approach is a systematic comparison of both normal and lognormal distributions (mean and median values) with various substitution methods of replacements of values below the detection limit by five constants: zero, LOD/2, LOD/√2, LOD and *no data* values, all for differently censored data (from 1 % to 50 %), along with a comparison of five geochemical parameters.

Materials and Methods

For the simulation of influence of different detection limits on the statistical calculations, a statistical dataset of normally distributed data was generated first. Data was generated in Microsoft Excel with an internal function NORMINV. Number of data was chosen as N = 65 536, being the maximum possible to calculate further in the program Statistica (Statsoft, Inc.) for the calculations and histograms. Mean value was chosen as 1.00 and standard deviation as 0.25. These data were later transformed to create a lognormal distribution.

Values were later censored at different levels, first by discarding the lowest 1 % data. These were replaced with values of 0, LOD/2, LOD/√2, LOD and with blank values (*no data*). Datasets with censored 5 %, 10 %, 25 % and 50 % were analyzed as well (Figure 1). Some authors have used data with up to 60 % of censored values (Hornung & Reed, 1990), up to 80 % censored values (Gilliom & Helsel, 1986) or even more than 90 % (Glass & Gray, 2001), but any conclusion based on the analysis of such dataset can be considered as highly inaccurate.

Beside the mean value, the comparison of medians is also presented, as this statistical value is often used in non-parametric statistics for non-normally distributed data.

For the analysis of real geochemical data, a subset of geochemical monitoring data (kindly provided by Slovenian Environmental Agency) for groundwaters in karstic and fractured aquifers was used. Analyses have been performed in years 1990–2009, according to national monitoring program and Decree on groundwater status (Uradni list RS, 25/2009). Number of analyses was $N = 942$, and from available dataset, the following five were chosen for the analysis: NO_3, O_2, Sr, F and Mn. Selection was based on two criteria; first that the percentage of censored data was approximately the same as the vales of generated data (from 1 % to 50 %), with values of NO_3: 0,5 %, O_2: 5 %, Sr: 17 %, F: 25 % and Mn: 45 %, with intention to compare the real data

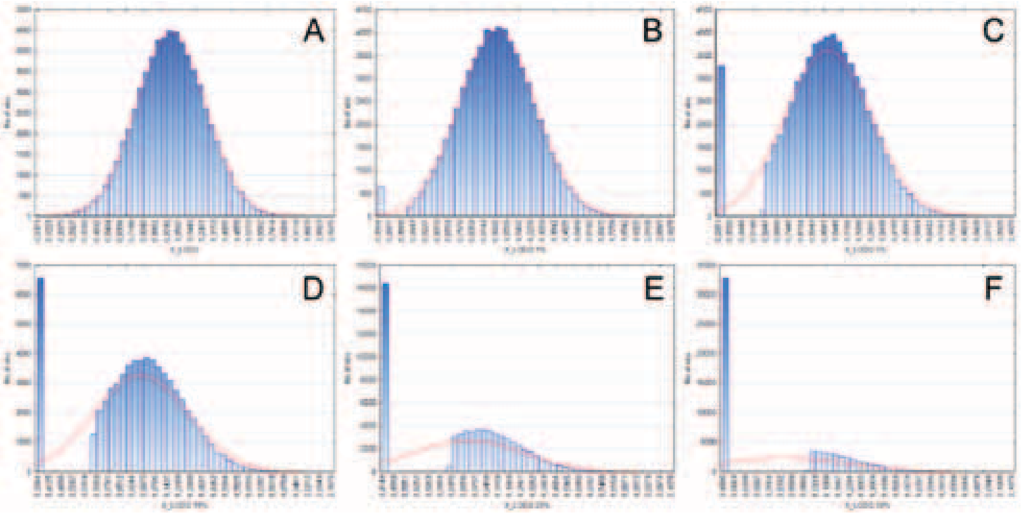


Figure 1. Histograms for generated ideal normal data ($N = 65\ 536$) with data below the detection limit substituted by LOD/2 and censored for: A: no censoring, B: 1 %, C: 5 %, D: 10 %, E: 25 %, F: 50 % of all data.

with generated ones. Secondly, the actual number of data must have been high, at least 70 % of all analyses (NO_3: N = 942, O_2: N = 942, Sr: N = 678, F: N = 858, Mn: N = 942).

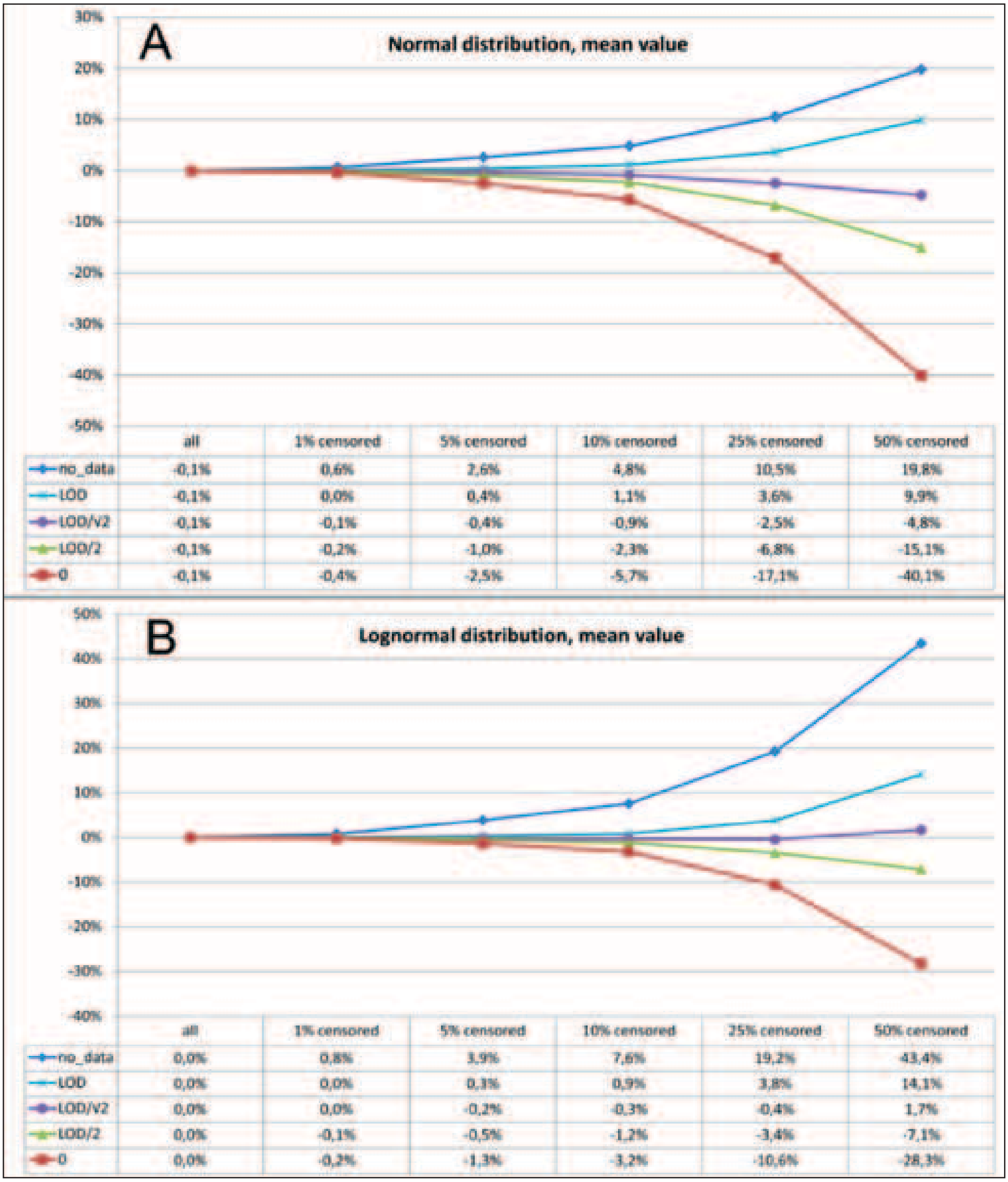
Regarding the analyzed parameters, it must be mentioned that from year 2003 on some major ions (Ca^{2+}, Mg^{2+}, HCO_3^-, ...) are missing from the analyses, as they are not anymore required to analyze according to Rules on drinking water (Uradni list RS, 19/2004). This is a major information loss, as geochemical modeling cannot be performed with such missing data, and the cost of including these parameters into a complete analysis is relatively small.

Results and discussion

Analysis of normally distributed generated data (Figure 1A, Table 1) clearly shows that the replacement of censored data with values of LOD/√2 is the best among all substitution methods, as the error is lowest for this approach for all censoring levels (1 %, 5 %, 10 %, 25 % in 50 %). Error is here represented as a ratio between the calculated mean value and true mean value (1.00), in percent. Average error is –1.7 % for normal distribution (Table 1). Replacement with zero produces the greatest errors and should be avoided. Other methods lie in between and also should not be used.

Replacement of values below the limit of detection with LOD/√2 underestimates the true mean value only slightly, substitution by LOD/2 produces greater underestimation and substitution by zero the greatest underestimation, which can be seen by the greatest deflection of the curve in Figure 2A. Contrarily, the replacement with LOD or *no data* values overestimated the true mean values.

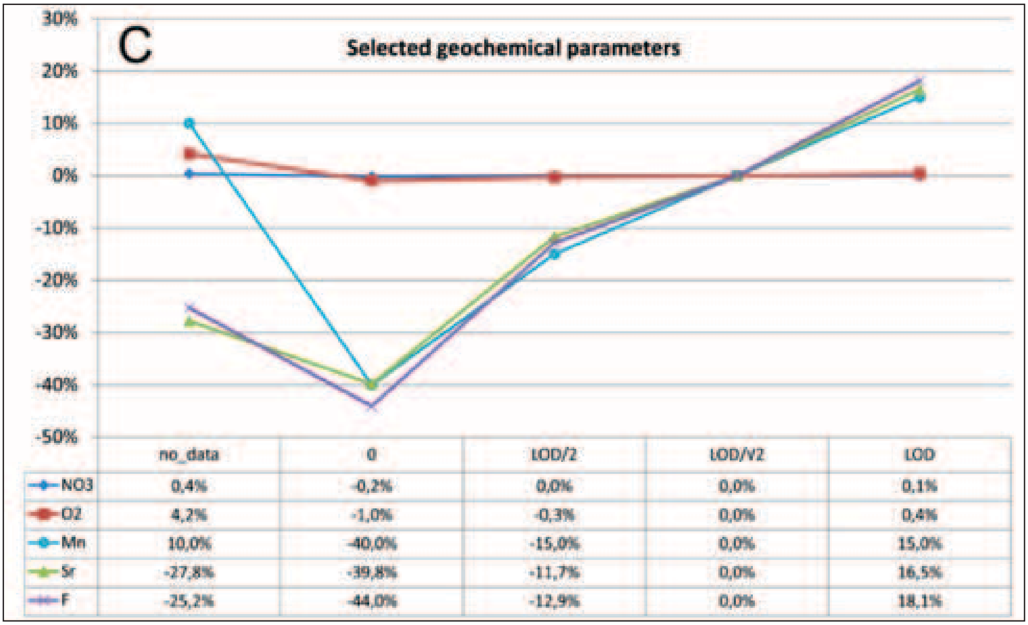
Similar conclusions are found for the lognormally distributed data (Figure 2B), where the distribution with LOD/√2 is again considered the best (average error is only 0.2 % for this method) and therefore recommended method for the substitution, and replacement with *no data* values behaves as the worst method. From the comparison of curves for both distributions, the major difference lies in the fact which method performed worst. For normal distribution, this is substitution by zero and for lognormal distribution, the substitution by *no data* values. The reason lies in the skewness of data, as much more data lies on the left side of the histogram for the lognormal data. As seen from both figures, error is steeply increasing with the percentage of censored data, up to 40 % when a half of data are censored. The increase is not linear and this holds for all methods.



	all	1% censored	5% censored	10% censored	25% censored	50% censored
no_data	-0,1%	0,6%	2,6%	4,8%	10,5%	19,8%
LOD	-0,1%	0,0%	0,4%	1,1%	3,6%	9,9%
LOD/√2	-0,1%	-0,1%	-0,4%	-0,9%	-2,5%	-4,8%
LOD/2	-0,1%	-0,2%	-1,0%	-2,3%	-6,8%	-15,1%
0	-0,1%	-0,4%	-2,5%	-5,7%	-17,1%	-40,1%

	all	1% censored	5% censored	10% censored	25% censored	50% censored
no_data	0,0%	0,8%	3,9%	7,6%	19,2%	43,4%
LOD	0,0%	0,0%	0,3%	0,9%	3,8%	14,1%
LOD/√2	0,0%	0,0%	-0,2%	-0,3%	-0,4%	1,7%
LOD/2	0,0%	-0,1%	-0,5%	-1,2%	-3,4%	-7,1%
0	0,0%	-0,2%	-1,3%	-3,2%	-10,6%	-28,3%

Figure 2A, 2B. Errors (ratios between the calculated mean value and true value in %) for replacement of values with no data, zero, LOD/2, LOD/√2 and LOD. A. Normal data, mean values. B. Lognormal data, mean values.



	no_data	0	LOD/2	LOD/√2	LOD
NO3	0,4%	-0,2%	0,0%	0,0%	0,1%
O2	4,2%	-1,0%	-0,3%	0,0%	0,4%
Mn	10,0%	-40,0%	-15,0%	0,0%	15,0%
Sr	-27,8%	-39,8%	-11,7%	0,0%	16,5%
F	-25,2%	-44,0%	-12,9%	0,0%	18,1%

Figure 2C. Errors (ratios between the calculated mean value and true value in %) for replacement of values with no data, zero, LOD/2, LOD/√2 and LOD. C. Errors for selected geochemical parameters: NO_3, O_2, Sr, F and Mn.

From the results of geochemical data (Figure 2C) it is obvious that there are vast differences both among the methods and among the parameters. As the true mean values for all populations are not known (they are influenced by censored values), all comparisons are normalized to the results of method of LOD/√2, as it turned out to be the most accurate. Major differences are visible between the group of NO_3 and O_2 and the group of Sr, F and Mn. Differences within the first group are very small (mostly below 1 %), and very big in the second (up to 40 %). There are several reasons for such behavior compared to generated ideal datasets:

- The detection limits in presented real dataset are often bigger than the values themselves (taking into account only uncensored data). For example, limit of detection for strontium is equal to 500 μg/L, but the mean value of uncensored data is equal to 101 μg/L. Any substitution of LOD/2, LOD/√2 or LOD thus gives much bigger values than the average, so the estimated means are much higher and obviously unacceptable for any further calculations. Censored values should be obviously smaller than the mean. Such an example of a large deviation for strontium is evidently presented in the histogram in Figure 3. The left histogram (Figure 3A) is based on all data with no censored values (*no data*), and the right one (Figure 3B) for the same data with inclusion of values substituted by LOD/√2. A clear peak is visible for the latter data, with values much above the mean values. Such substitution cannot be used as it overestimates the mean by a great value.

- A large fraction of parameters Sr, F and Mn are highly censored, so the estimation of mean values is more problematic than the estimation of less censored NO_3 and O_2.
- Data distribution can also cause some deviations from the ideal datasets. Distributions were not tested for normality by special methods, but were visually estimated from histograms and are mostly lognormal.
- For some parameters, there exist several limits of detection in different time periods. Consequently, different substitutions can influence the statistical calculations. A possible approach to overcome such problem (Helsel & Cohn, 1988) is to use the largest limit of detection for all censored data, but the information about the lower limits is lost in such cases.

Again, the greatest deviations are found for replacements with zero or *no data* values, and this is much more pronounced for parameters Sr, F and Mn. For NO_3 and O_2, the difference between the methods is very small, which can be attributed to very small censoring levels (0.5 % and 5 %).

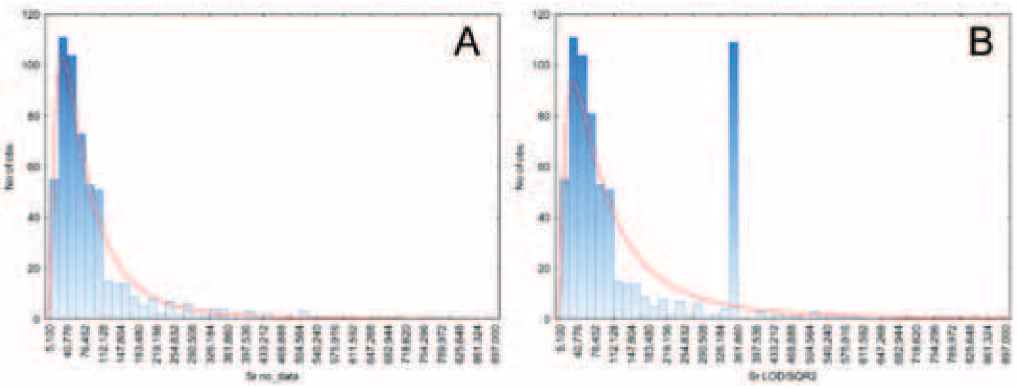


Figure 3. A. Histogram of strontium distribution with substitution of no data values. B. Histogram of strontium distribution with visible outliers, created by substitution of LOD/√2.

Table 1. Average method errors for all censored data - mean and median values for normal and lognormal distributions.

Method and distribution	no data	0	LOD/2	LOD /√2	LOD
mean values - normal distribution	7.7 %	–13.2 %	–5.1 %	–1.7 %	3.0 %
mean values - lognormal distribution	15.0 %	–8.7 %	–2.4 %	0.2 %	3.8 %
median - normal distribution	5.9 %	0 %	0 %	0 %	0 %
median - lognormal distribution	157.1 %	0 %	0 %	0 %	0 %

Table 1 summarizes the results for produced errors committed by several substitution methods for mean values and medians, for both normal and lognormal distributions. Replacement with LOD/√2 gives the best results, the replacement with LOD/2 or LOD causes larger deviations and replacements with zero or *no data* the highest errors. Interesting fact is that the median value is not sensible to any method of substitution except for replacement by blank values (ignoring the censored data). In all other cases, there is no deviation from the true data. The reason for such behavior is relatively obvious, as median is by definition the value separating the higher half of the sample from the lower half. Consequently, for data censored up to 50 %, there is no difference between the medians for different methods. For higher censored values, the median is nevertheless increasingly influenced by higher censoring level.

Based on presented results, it is therefore recommended that no more than about 20 % of data should be censored, if one should keep the error relatively small - below few percent for both lognormal and normal methods for substitution with LOD/√2 for the ideal data and somewhat larger value (but still permissible) for natural data.

Conclusions

From the presented results it is clear that the substitution of data below the limit of detection is complex and still not adequately used. For substitution of different fractions of LOD, based on results of this study, the method of LOD/√2 is recommended, as the committed errors are the lowest. Other methods perform worse, with replacement with LOD/2 or LOD causing larger deviations and replacements with zero or *no data* the highest deviations. Real data present a challenge, due to problems with multiple limits of detection, unknown distributions and other factors, but still the substitution with LOD/√2 is recommended if neither of such factors is known. Median is not sensible to replacement method except for the replacement with *no data* values, up to 50 % of the censored data and can be used as a reported value, like the mean.

Substitution with estimation of missing data below the detection limit can be obtained by more complex statistical methods, like the maximum likehood estimation or various regressions (not used in this paper), but for such, the distribution of data should generally be known. A better method is to analyze the samples again (perhaps in another laboratory) or use another method, to avoid the censored values as much as possible.

Acknowledgments

Work was supported by Slovenian Research Agency (Javna agencija za raziskovalno dejavnost Republike Slovenije - ARRS), number of research project Z1-3670. Author would also like to thank the Slovenian Environmental Agency for the provided geochemical data.

References

Chastain, J. R. (2007): Censored Data: What's The Average Of Unknown Values. Chastain-Skillman, Inc.

Croghan, C. W. & Egeghy, P. P. (2003) Methods of Dealing with Values Below the Limit of Detection using SAS, paper presented at the *Southeastern SAS User Group*, City, 22–24 September, 2003.

Gilliom, R. J. & Helsel, D. R. (1986): Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 1. Estimation Techniques – *Water Resources Research,* Vol. 22, No. 2, pp. 135.

Glass, D. C. & Gray, C. N. (2001): Estimating mean exposures from censored data: exposure to benzene in the Australian petroleum industry – *The Annals of occupational hygiene,* Vol. 45, No. 4, pp. 275–82.

Gochfeld, M., Burger, J. & Vyas, V. (2005) *Statistical Analysis of Data Sets with Values Below Detection Limits.* Report for (Piscataway, New Jersey).

Helsel, D. R. (1990): Less than obvious - statistical treatment of data below the detection limit – *Environmental Science & Technology,* Vol. 24, No. 12, pp. 1766–1774.

Helsel, D. R. (2005): More than obvious: Better methods for interpreting nondetect data – *Environmental Science & Technology,* Vol. 39, No. 20, pp. 419a–423a.

Helsel, D. R. & Cohn, T. A. (1988): Estimation of descriptive statistics for multiply censored water quality data – *Water Resources Research,* Vol. 24, No. 12, pp. 1997.

Helsel, D. R. & Gilliom, R. J. (1986): Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 2. Verification and Applications – *Water Resources Research,* Vol. 22, No. 2, pp. 147.

Hornung, R. W. & Reed, L. D. (1990): Estimation of Average Concentration in the Presence of Nondetectable Values – *Applied Occupational and Environmental Hygiene,* Vol. 5, No. 1, pp. 46–51.

Lambert, D., Peterson, B. & Terpenning, I. (1991): Nondetects, Detection Limits, and the Probability of Detection – *Journal of the American Statistical Association,* Vol. 86, No. 414, pp. 266–277.

Smith, D., Silver, E. & Harnly, M. (2006): Environmental samples below the limits of detection – comparing regression methods to predict environmental concentrations. http://www.lexjansen.com/wuss/2006/Analytics/ANL-Smith.pdf

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,* Vol. 1, No. 7, pp. 436–441.

Uradni list RS (19/2004): Pravilnik o pitni vodi = Rules on drinking water, Ljubljana.

Uradni list RS (25/2009): Uredba o stanju podzemnih voda = Decree on groundwater status, Ljubljana.