Radioprotection, vol. **46**, n° 6 (2011) S85–S90

DOI: 10.1051/radiopro/20116728s

Limit of detection values in data analysis: Do they matter?

M.D. Wood[1], N.A. Beresford[2] and D. Copplestone[3]

[1] *School of Science & the Environment, Manchester Metropolitan University, Manchester, M1 5GD, UK*
[2] *Centre for Ecology & Hydrology, Bailrigg, Lancaster, LA1 4AP, UK*
[3] *School of Biological & Environmental Sciences, University of Stirling, Stirling, FK9 4LA, UK*

Abstract. Data sets containing values below the limit of detection (LOD) are known as 'censored data sets'. Such data sets are encountered regularly in most fields of environmental contaminant research. The current norm within environmental radioactivity research is to use substitution methods when analysing data sets that include values below the LOD, commonly replacing each LOD value with a value equal to half the LOD (LOD/2). However, this approach has no statistical basis and has implications when summarising or comparing data sets because it can lead to underestimates or overestimates of both the mean and the standard deviation. To remove the need to apply substitution methods, over the last four decades other fields of environmental science have been adopting statistical techniques developed for medical research applications. Despite the long history of applying these techniques in other fields and two recent environmental radioactivity publications that have used survival analysis techniques, there still seems to be reluctance within the environmental radioactivity research community to adopt these 'new' methods. This paper introduces the statistical techniques that can be used in place of LOD substitution, presents some guidance on the applicability of these techniques for different levels of data censoring and provides some examples of the use of these methods in various contexts. It is hoped the present paper will contribute to the evidence-base supporting the use of survival analysis within the field of environmental radioactivity research and go some way to changing the current norm of substitution using LOD/2.

1. INTRODUCTION

The limit of detection (LOD) is an analytical threshold defining the smallest true value that can be distinguished from an analytical blank sample, with a specified probability of error [1, 2]. Data sets containing values below the LOD are known as 'censored data sets' [3]. Such data sets are encountered regularly in most fields of environmental contaminant research.

When analysing data sets containing censored data, it is necessary to decide how values below the LOD should be treated. A range of approaches have been proposed, and adopted to varying degrees, within the peer-reviewed literature. These include: (i) treating the LOD value as the absolute value (i.e. if the LOD value is <0.5 then the value of 0.5 will be used); (ii) treating the LOD values as zero; (iii) excluding the LOD values from the data set; and (iv) substituting the LOD value with a value between zero and the LOD value. Whichever of these methods is employed, the resultant non-censored data set will have certain limitations and deficiencies when it comes to subsequent data analysis and interpretation (Table 1, [2–5]). For example, treating LOD values as absolute values may be viewed as conservative because the mean is overestimated (Table 1). However, the standard deviation is underestimated. If summary statistics are being used as probabilistic analysis input parameters, such as the modelling of radionuclide transfer to wildlife within the ERICA Tool [6], the underestimated standard deviation may result in predictions that are less conservative. At best, substitution approaches serve as a simplistic means for approximating reality but, at worst, they may result in unjustified conclusions being drawn [7].

Table 1. Implications of using substitution methods to estimate summary statistics (adapted from [4]).

Substitution method	Effect on Mean[1]	Effect on Standard Deviation[1]
Treating the LOD value as the absolute value	Overestimated	Underestimated
Treating the LOD values as zero	Underestimated	Overestimated
Excluding the LOD values from the data set	Overestimated	Underestimated[2]
Substituting the LOD value with a value between zero and the LOD value	May be overestimated or underestimated	May be overestimated or underestimated

Notes: [1]The extent of the overestimate or underestimate and the implications of this will be reduced when a low percentage of the data set is censored data (see text in Section 2); [2]If the lowest value in the distribution is an absolute value rather than an LOD value, the standard deviation may not be an underestimate.

Alternative methods that could be used to overcome this problem have been proposed (e.g. [3]). Many of these methods belong to a family of statistical techniques known as survival analysis methods, which were originally developed for analysing medical data. They offer a means of analysing censored data sets without the need for modifying the LOD values. Despite two recent papers recommending the use of survival analysis methods in the field of environmental radioactivity research [8, 9], substitution methods are still being used extensively. There is a need to build a case for the adoption of survival analysis methods within environmental radioactivity research and to provide guidance on their use.

In this paper we consider the pros and cons of current norms for handling censored data and highlight the influence of LOD treatment choice on the results obtained. We then introduce the range of survival analysis and allied methods and question whether these methods should be adopted more widely within our research community instead of the more conventional data fabrication approaches that are commonly used. We discuss the evidence base for the adoption of survival analysis techniques in other areas of environmental science. We also provide guidance to facilitate environmental radioactivity researchers in selecting justifiable techniques for analysis of censored data sets in the future.

2. CURRENT NORMS IN ENVIRONMENTAL RADIOACTIVITY RESEARCH

The most commonly used approach for handling LOD values in environmental datasets is to substitute the LOD value with a value equivalent to half of the LOD (LOD/2) (e.g. [10–12]). The substitution approach has been used in the development of recent databases of radionuclide transfer parameters for wildlife where $<20\%$ of the data were LOD values, compiled under European Commission and International Atomic Energy Agency programmes [13–15]. The use of the substitution approach in the development of these transfer databases is in line with the guidance provided by the United States Environmental Protection Agency (US EPA), which suggests that substitution may be an adequate approach when up to 15% of the data are censored [16]. However, it has been commented that the 15% value is based on judgement rather than being scientifically derived [7].

Substitution methods are simple to apply [17] and, for situations where LOD values comprise a minor proportion of the total dataset, substitution methods may also be viewed as a pragmatic approach. However, there is little, if any, statistical rationale behind the substitution method[1] and the approach performs worst in situations where there are multiple detection limits [2], which is a situation often encountered in radioecological data sets and especially those obtained from gamma spectrometry [9]. There are alternative methods that can be used in these circumstances and are statistically justifiable.

[1] It has been commented that data substitution methods could be referred to more honestly as data fabrication approaches [2]. However, the same could be said of approaches that assume a distribution and use this to effectively assign values to LOD values.

Table 2. Approaches for estimating summary statistics for censored data, where n is the number of observations in the data set (adapted from Helsel [2]).

Degree of censoring	$n = < 50$	$n = > 50$
< 50% censored[1]	• Kaplan-Meier	• Kaplan-Meier
50 – 80% censored[2]	• Robust MLE or ROS	• MLE
> 80% censored[2]	• Only report the % of data above a threshold [3]	• May also report high sample percentiles (e.g. 90th and 95th)

Notes: [1]Kaplan-Meier only reports the median when less than 50% of the data set are censored so, beyond 50% censoring, a parametric approach is recommended; [2]80% cut-off is based on findings from previous studies which suggest that estimates of summary statistics when more than 80% of data are censored are effectively guesses; [3]The maximum LOD may be a suitable threshold in this case.

3. ALTERNATIVE METHODS

Although some authors (e.g. [18, 19]) continue to support the use of substitution as an 'adequate' method for handling LOD values, there are many theoretical (e.g. [2, 3, 7, 20]) and empirical (e.g. [4, 21]) studies that demonstrate substitution to be inferior to survival analysis methods and to non-parametric survival analysis methods in particular (e.g. [22]). For example, there are several methods that can be used for calculating summary statistics, including Maximum Likelihood Estimation (MLE), Regression on Order Statistics (ROS) [2] and the non-parametric Kaplan-Meier method [23]. Helsel [2] proposed a set of rules for selecting a suitable method for estimating summary statistics based on the number of observations in the data set and the proportion of the total data set that is censored (Table 2). These rules provide a useful guide but should be seen as broad indicators given that other authors have suggested different censoring cut-off criteria. For example, Antweiler and Taylor [18] concluded that Kaplan-Meier gave the most accurate estimate of summary statistics for censored data sets containing up to 70% censored data. As a non-parametric technique, the Kaplan-Meier method is the most widely applicable of these survival analysis methods (e.g. [18, 24]).

Helsel [2] also identified equivalent statistical methods for uncensored and censored data sets to facilitate selection of the appropriate statistical analysis technique when a dataset includes values below the LOD (Table 3). These techniques are available in most of the standard statistical analysis software packages, such as SPSS and Minitab. There is also a freely-available package that has been developed for running survival analysis with the free statistics software package R.

Survival analysis methods were developed for statistical analysis in medical research. The key difference between medical and environmental research data is the type of censoring; medical statistics tend to be right censored (Fig. 1a, censoring is in the form of 'greater than' values), whereas environmental datasets are generally left censored (Fig. 1b, censoring is in the form of 'less than' values) [2]. Therefore, to utilise medical statistical analysis techniques for environmental data, a left-censored environmental data distribution must be transformed by subtracting all values in the dataset from a constant that is larger than the maximum value in the dataset. This process is known as 'flipping'. Depending on the software and the analytical technique used, it may be necessary to reverse flip the result upon completion of the analysis.

4. APPLICATIONS IN ENVIRONMENTAL SCIENCE

The application of statistical analysis techniques to handling censored environmental datasets dates back to the early 1970s (e.g. [25]). Over the last four decades, survival analysis methods have been used to summarise and analyse censored data sets in many fields of environmental science. The extent of interest

Table 3. Comparable methods for uncensored and censored data sets (adapted from Helsel [2]).

Purpose of analysis	Uncensored data sets	Censored data sets
Calculation of summary statistics	• Standard descriptive statistics	• Kaplan-Meier, MLE or ROS
Comparison of two groups	• t-test	• Censored regression (with 0/1 group indicator[1])
	• Wilcoxon rank sum	• Generalised Wilcoxon test
	• Paired t-test	• Censored confidence interval (CI) on differences
	• Paired sign or signed-rank test	• Paired Prentice–Wilcoxon (PPW) test
		• Akritas test
Comparison of three or more groups	• Analysis of Variance (ANOVA)	• Censored regression (with 0/1 group indicator1)
	• Kruskal-Wallis test	• Generalised Wilcoxon test
Correlation	• Pearson's r	• Likelihood r
	• Kendall's tau	• Kendall's tau b
Linear regression	• Regression	• Censored regression
		• Logistic regression
	• Robust regression	• Proportional hazards (Cox) regression
	• Theil-Sen median line	• Akritas-Theil-Sen median line

Notes: [1]It is necessary to indicate in the datasheets whether values are above or below the LOD. This can be done by adding a dummy variable to the datasheet and using 0 to categorise LOD values and 1 to categorise absolute values.

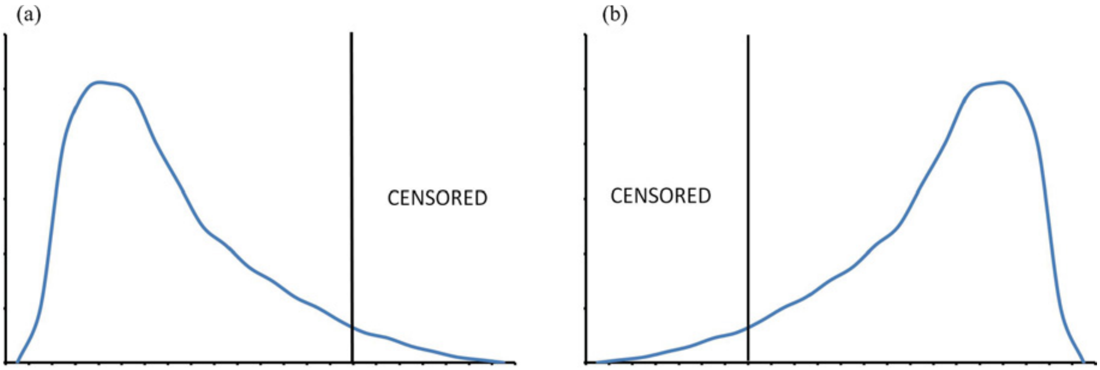

Figure 1. Example data distributions showing (a) right censored data and (b) left censored data. The process of 'flipping' converts left censored data to right censored data, which can then be analysed using survival analysis techniques.

in these methods is indicated by the 186 citations[2] that Helsel's 1990 paper [3] on methods for handling LOD values in environmental science has received[3]. Examples of the use of statistical techniques when analysing censored data include studies on pesticides, metals, polycyclic aromatic hydrocarbons (PAHs),

[2] Results of an ISI Web of Knowledge search performed on 17th June 2011.
[3] A few of the 186 references use Helsel's 1990 paper [3] as a reference to support their use of substitution, which is strange given that he argues strongly *against* the use of substitution and states that using substitution methods "leads to undesirable and unnecessary errors".

dioxins and, more recently, radionuclides. The following examples illustrate some of the applications of these techniques.

Millard & Deverel [26] compared censored data sets of trace element concentrations in groundwater from two areas of different geology in the San Joaquin Valley of California. Using survival analysis techniques they were able to identify significant differences between the Zn concentrations in the two areas.

In analysing pyrene concentrations in water samples collected from Puget Sound in Washington, She [21] compared the performance of four techniques for handling censored data. These techniques were substitution, MLE, probability plotting and Kaplan-Meier. The study demonstrated that Kaplan-Meier provided the most reliable estimates of summary statistics and, because it does not rely on a defined distribution shape, was more robust than methods such as substitution.

Wood *et al*. [9] developed a database of transfer parameters for reptiles. These data can be used to predict whole-body activity concentrations in reptiles from environmental media (soil or water). There were limited data available on radionuclide transfer to reptiles and the transfer parameter database contained some values below the LOD. Using the guidance outlined in Table 2, the Kaplan-Meier method was used to estimate the mean and standard deviation of reptile transfer datasets that included LOD values [9].

Fievet and Della Vedova [8] analysed time-series ^{137}Cs activity concentration data from seaweed samples collected at La Hague. They used both the Kaplan-Meier method and the ROS method, selecting the appropriate method for the degree of censoring and number of observations within each time period based on the guidance presented in Table 2.

Leith *et al*. [27] analysed data on contaminant concentrations in bald eagle eggs. The data were for polychlorinated biphenyl (PCB) and p,p'-dichlorodiphenyldichloro-ethylene (p,p'DDE) concentrations and included LOD values. They tested substitution, imputation and Kaplan-Meier methods for analysing the data and found Kaplan-Meier to perform best overall.

These are just a few examples of the use of alternative approaches to the analysis of censored environmental data, rather than relying on substitution methods with their inherent deficiencies.

5. CONCLUSION

Given the regularity with which LOD values are encountered in environmental radioactivity datasets, the ease with which survival analysis methods can be applied and the extent to which these methods have been applied in other fields of environmental research, it is surprising that survival analysis methods do not appear to have been widely adopted within environmental radioactivity research. There seems to be reluctance amongst researchers to re-think the established norms. This results in the ongoing re-enforcement of substitution methods in the peer-reviewed literature, even though the failings of these methods have been recognised for the last few decades. Changing the mindset of researchers in the field of environmental radioactivity is a challenging task. Fievet and Della Vedova [8] and Wood *et al*. [9] have published environmental radioactivity research recent papers that utilise survival analysis techniques. It is hoped that the present paper, in addition to these two previous papers [8, 9], will contribute to the evidence-base supporting the use of survival analysis within the field of environmental radioactivity research and go some way to changing the current norm of substitution using LOD/2.

References

[1] EC, Commission Recommendation of 18 December 2003 on standardised information on radioactive airborne and liquid discharges into the environment from nuclear power reactors and reprocessing plants in normal operation. *Official Journal of the European Union* (2004).

[2] Helsel, D.R., Nondetects and data analysis: statistics for censored environmental data. (John Wiley & Sons Inc., New Jersey, 2005)

[3] Helsel, D.R., *Environ Sci Technol* 24 (1990) 1766-1774.

[4] Baccarelli, A., Pfeiffer, R., Consonni, D., Pesatori, A.C., Bonzini, M., Patterson, D.G., Bertazzi, P.A., Landi, M.T. *Chemosphere* 60 (2005) 898-906.

[5] Gleit, A., *Environ Sci Technol* 19 (1985) 1201-1206.

[6] Brown, J.E., Alfonso, B., Avila, R., Beresford, N.A., Copplestone, D., Prohl, G., Ulanovsky, A., *J Environ Radioact* 99 (2008) 1371-1383.

[7] Helsel, D.R., *Chemosphere* 65 (2006) 2434-2439.

[8] Fievet, B., Della Vedova, C., *J Environ Radioact* 101 (2010) 1–7.

[9] Wood, M.D., Beresford, N.A., Semenov, D.V., Yankovich, T.L., Copplestone, D., *Radiat Environ Biophys* 49 (2010) 509-530.

[10] Garcia-Fernandez, A.J., Gomez-Ramirez, P., Martinez-Lopez, E., Hernandez-Garcia, A., Maria-Mojica, P., Romero, D., Jimenez, P., Castillo, J.J., Bellido, J.J., *Ecotoxicol Environ Saf* 72 (2009) 557-563.

[11] Tajimi, M., Uehara, R., Watanabe, M., Oki, I., Ojima, T., Nakamura, Y., *Chemosphere* 61 (2005) 1256-1262.

[12] Vitaliano, J.J., Zdanowicz, V.S., *Marine Pollut Bull* 24 (1992) 364-367.

[13] Beresford, N.A., Barnett, C.L., Howard, B.J., Scott, W.A., Brown, J.E., Copplestone, D., *J Environ Radioact* 99 (2008) 1393-1407.

[14] Hosseini, A., Thorring, H., Brown, J.E., Saxen, R., Ilus, E., *J Environ Radioact* 99 (2008) 1408-1429.

[15] IAEA (International Atomic Energy Agency), Handbook of parameter values for the prediction of radionuclide transfer in Wildlife (IAEA, Vienna, 2011) (*Draft in press*)

[16] US EPA, Guidance for data quality assessment. Practical methods for data analysis. (Office of Environmental Information, Washington DC, 2000).

[17] Geras'kin, S., Oudalova, A., Michalik, B., Dikareva, N., Dikarev, V., *Chemosphere* 83 (2011) 1133-1146.

[18] Antweiler, R.C., Taylor, H.E., *Environ Sci Technol* 42 (2008) 3732-3738

[19] Glass, D.C., Gray, C.N., *Ann Occup Hyg* 45 (2001) 275–282.

[20] Hornung, R.W., Reed, L.D. *Appl Occup Environ Hyg* 5 (1990) 46-51.

[21] She, N., *J Am Water Resour Assoc* 33 (1997) 615-624.

[22] Zhang, D.H., Fan, C.P., Zhang, J., Zhang, C.H., *Stat Med* 28 (2009) 700-715.

[23] Kaplan, E.L., Meier, P., *J Am Stat Assoc* 53 (1958) 457-481

[24] Brady, D., Pratt, G.C., *J Air Waste Manag Assoc* 57 (2007) 1091-1102.

[25] Leese, M. N., *Water Resour Res* 9 (1973) 1534-1542.

[26] Millard, S.P., Deverel, S.J., *Water Resour Res* 24 (1988) 2087-2098.

[27] Leith, K.F., Bowen, W.W., Wierda, M.R., Best, D.A., Grubb, T.G., Sikarske, J.G., Chemosphere 80 (2010) 7–12.