WARSAW UNIVERSITY OF TECHNOLOGY Index 351733
FACULTY OF CIVIL ENGINEERING
COMMITTEE FOR CIVIL AND WATER ENGINEERING
POLISH ACADEMY OF SCIENCES ISSN 1230-2945

DOI: 10.24425/ace.2020.136471

ARCHIVES OF CIVIL ENGINEERING

Vol. LXVII ISSUE 1 2021

CENSORED RANDOM VARIABLE AS A FORM OF COPING WITH MISSING DATA IN STUDYING THE LEACHABILITY OF HEAVY METALS FROM HARDENING SLURRIES

Ł. SZAREK[1], Z. KLEDYŃSKI[2]

Missing data in test result tables can significantly impact the analysis quality, especially in relation to technical sciences, where the mechanism generating missing data is often non-random, and their presence depends on the non-observed part of studied variables. In such cases, the application of an inappropriate method for dealing with missing data will lead to bias in the estimated distribution parameters.

The article presents a relatively simple method to implement in dealing with missing data generated as a result of the MNAR mechanism, which utilizes the censored random variable. This procedure does not modify the variable distribution form, which is why it ensures objective and efficient estimation of distribution parameters within studies affected by certain restrictions of technical or physical nature (censored distribution), with a relatively low workload. Furthermore, it does not require the application of specialized software. A prerequisite for using this method is the knowledge of the frequency and cause of missing data.

The method for estimating the random variable censored distribution parameters was shown based on the example of studying the leachability of selected heavy metals from a hardening slurry. The analysis results were compared with classical methods for dealing with missing data, such as, ignoring missing data observations (listwise or pairwise deletion), single imputation and stochastic regressive imputation.

Keywords: missing data; censored random variable; censored distribution; heavy metals; leachability; hardening slurries;

1 PhD., Eng.,Warsaw University of Technology, Faculty of Building Services, Hydro and Environmental Engineering, Nowowiejska 20, 00-653 Warsaw, Poland, ORCiD: 0000-0002-5794-9271, e-mail: lukasz.szarek@pw.edu.pl

2 Prof., DSc., PhD., Eng., Warsaw University of Technology, Faculty of Building Services, Hydro and Environmental Engineering, Nowowiejska 20, 00-653 Warsaw, Poland, ORCiD: 0000-0002-8387-3119, e-mail: zbigniew.kledynski@pw.edu.pl

1. INTRODUCTION

Missing data are frequent in various types of studies, which can result from the properties of the subject matter, as well as the nature of the very research. Despite attempts at achieving data completeness, in practice we should expect missing data. Missing data can significantly impact the analysis results, and even lead to erroneous conclusions in extreme cases.

There are three basic mechanisms that are distinguished, which generate missing data [1–4]:

1) MCAR (missing completely at random) – a fully random mechanism. Missing data are defined as completely random, when the probability of missing data neither depends on the observed or missing values of the studied variable nor the unstudied variables. It is a particular case of the MAR mechanism, in which the population parameters can be estimated based on the acquired data set, and its gaps can be ignored. However, the statistical power of testing is lost in this case. An example of it can be a situation, where the results were not saved in random moments due to connection failure.
2) MAR (missing at random) – a random mechanism. Missing data are defined as missing at random, when the probability of their occurrence depends solely on the observed values and external factors. It is assumed that there is no link between the missing data and the values of non-tested variables. In this case, when the cause of missing data is taken into account in the analysis, the population distribution parameters can be estimated in an unbiased way, just like with MCAR. An example is studying two variables, one of which (e.g. kinetic energy of an electron swept from a metal surface) can be observed only above a certain limit value of the second variable (in this case, the limit frequency in the photoelectric phenomenon). Missing data depends on a variable without gaps.
3) MNAR (missing not at random) – a non-random mechanism. Missing data are defined as missing not at random, when the probability of their occurrence depends on the part of non-observed values of studied variable and external factors. In the event of ignoring missing data, this mechanism leads to estimating biased population distribution parameters. A situation, where the values of the studied variable (e.g. mass) can exceed the measurement range of the applied instruments can be used an example here.

MAR and MNAR mechanisms do not only apply to data, but also to their analysis manner [2]. The presented missing data mechanisms are not mutually exclusive, and are practically not found in pure

form. This is why it is most convenient to treat missing data as an outcome of both mechanisms, MAR and MNAR.

Missing data leads to numerous issues in terms of analysis and inference, including distorted distributions of analysed variables and increased estimator bias [5]. Various manners of observing their impact on the test results are applied, depending on the mechanisms generating missing data.

In terms of MCAR, the easiest approach is excluding all observations indicating gaps from the analysis (listwise deletion). However, this leads to the loss of test statistical power, and if the missing data mechanism is different than MCAR, the distribution is truncated, which in turn is associated with encumbering estimated parameters.

Another solution is removing observations in pairs (pairwise deletion); however this hinders the multidimensional analysis because it makes estimating degrees of freedom difficult (each variable pair can have a different number), which in turn impairs the application of statistical tests.

Various imputations, which involve replacing the non-observed variable parts with specific values are also used. They affect the precision estimates, e.g. single imputation leads to encumbering estimated parameters, primarily lowering the variance [5, 6].

The aforementioned classic methods of dealing with missing data, though uncomplicated in terms of implementation, are not universal and their effectiveness depends on the missing data mechanism.

In natural and technical sciences, the MNAR mechanism is often the backbone of missing data (e.g. limitations resulting from the properties of the test equipment, tested sample, etc.). The article presents a relatively simple method for dealing with missing data resulting from the MNAR mechanism, which utilizes the random variable of mixed character [7] in the analysis, often applied in, e.g., medical [8] and economic [9] sciences.

2. CENSORED RANDOM VARIABLE

A censored random variable is the name for the quantitative variable x, the studied characteristics of which are known up to (or beyond) a certain value x_0. Examples of such variables in civil engineering and environmental engineering are all variables defined in studies affected by certain limitations of technical or physical nature, within the test apparatus or experiment sample, e.g., substance concentration (limitation resulting from the determination limit, which the applied test method is characterized by), mechanical strength (limitation resulting from the measuring range of the durometer), fatigue strength (interrupted testing of samples not destroyed after a sufficiently high

number of loading cycles), sample mass (limitation resulting from the measuring range of the scale), maximum depth of water penetrating the sample during material tightness tests [10] (limitation resulting from the dimensions of the sample), etc.

In such cases, the phenomenon of missing data formation can be considered as an MNAR mechanism because the observed variable section is present only within its certain value range, whereas missing data depend on its non-observed part and external factors. This impacts the manner of dealing with missing data [1].

The censored random variable character impacts its empirical distribution form. Observed values of the variable x enable describing the native population only within the range of values lower (or higher) than x_0. This is why, the theoretical random variable (non-censored) population distribution shall be subjected to appropriate modification, called distribution censoring [7, 11].

A censored distribution differs from the distribution of the total population only in that the detailed frequencies are known merely to a certain value x_0 of the random variable (or beyond this value), and the population specific (theoretical) distribution form is maintained, i.e., it is not subject to modification specific to, e.g., the distribution truncating procedure [12]. Assuming that the probability of occurrence of values lower (or higher) than x_0 is concentrated at point x_0 and if the non-censored distribution is continuous (e.g. normal), the censored distribution will be a mixed distribution with its spinode at x_0 [13]. This means that there is a different from zero probability of a censored random variable adopting the value x_0, but apart from that, for $x > x_0$ (or $x < x_0$) it behaves like a continuous variable (Figure 1).

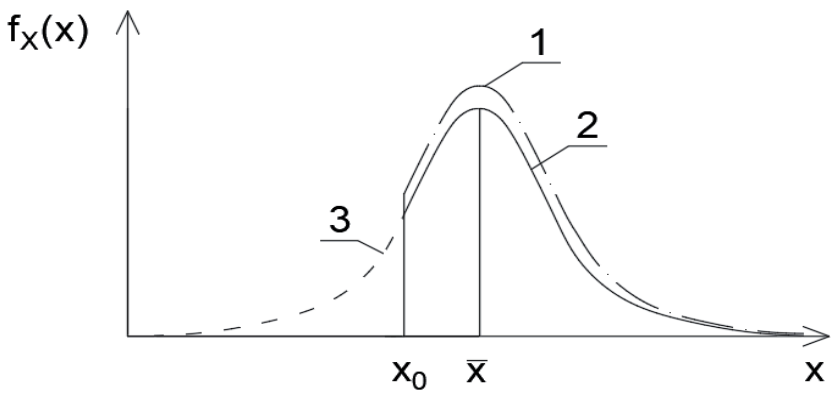


Fig. 1. Compared approaches towards estimating the theoretical distribution of a censored variable. Truncated distribution (1); censored distribution (left-side censoring) observed section (2), non-observed section (3); theoretical distribution of the total population (2+3)

3. ESTIMATING CENSORED DISTRIBUTION PARAMETERS

The method for estimating the random variable censored distribution parameters is presented using an example of the leachability study of selected heavy metals (lead, zinc, copper, chromium, cadmium) in eluates sampled from a hardening slurry composed with an admixture of fly ash derived from the thermal treatment of municipal sewage sludge [14].

The concentration of selected heavy metals in sampled eluates was determined using the flame atomic absorption spectroscopy (FAAS) method. Table 1 shows the determination limits of metals, whose leachability was studied.

Table 1. Method's determination limits and values adopted for imputation

Heavy metal	Determination limit [mg/dm^3]	Concentration value used for imputation [mg/dm^3]
Zinc (Zn)	0.01	0.005
Copper (Cu)	0.02	–
Lead (Pb)	0.03	–
Cadmium (Cd)	0.01	0.005
Chromium (Cr)	0.03	0.015

The concentration of heavy metals in the eluate was treated as the random variable.

The test involving the distribution form of concentration variable c_i was conducted for lead, owing to the concentration of this element in the eluate, which exceeded the determination value in all considered cases (Table Z1 – 1 Attachment 1). Eluates (n = 95) from hardening slurry samples were analysed. The Chi-square goodness-of-fit test with a significance level of $\alpha = 0.05$ was applied. The H_0 null hypothesis was that the random variable c_i had a Lognormal distribution. The form of the null hypothesis results directly from the physical nature of the concentration, which adopts solely positive values. Figure 2 contains a histogram with the test probability p and test statistics value χ^2, which indicate the lack of grounds to reject the null hypothesis at the established level of significance. Table 2 shows the distribution parameters with selected statistics.

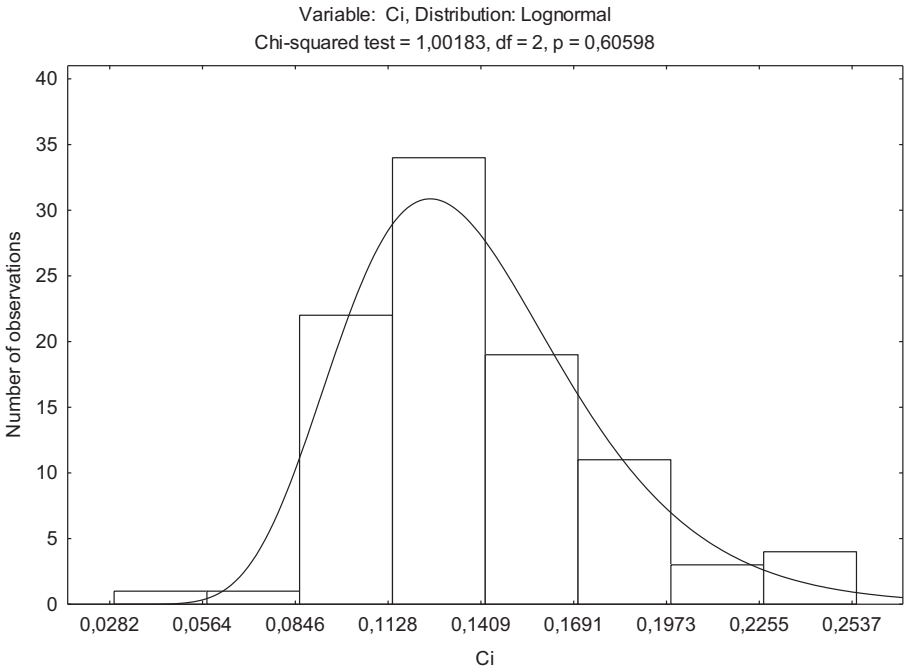


Fig. 2. Histogram for the empirical and expected distributions of lead concentrations in all eluate fractions ($\alpha = 0.05$; $\chi^2_\alpha = 5.991465$)

Table 2. Distribution parameters and selected lead concentration statistics

Lognormal concentration distribution parameters c_i		Selected statistics [mg/dm^3]			
μ	σ	Expected value E[X]	Standard deviation SD[X]	Median [X]	Modal value [X]
-2.0044	0.2661	0.1396	0.03781	0.1347	0.1255

The results of concentration determination for lead, zinc, copper, cadmium and chromium, in all eluate fractions, are given in Attachment 1. The obtained concentrations reached values below and above the determination limit. This means that there was a case of the non-random missing data generation mechanism (MNAR). The occurrence of a non-observed result depended on the non-observed section of the studied variable and the external factor – measuring device accuracy.

A binomial distribution was used for the calculations. The work involved calculating, for each metal and according to Formula 1, the frequency of concentrations below the determination limit p^*. The confidence interval limits for p^* (Table 3) were calculated according to Formula 2 [15].

$$p^* = \frac{Y}{n} \tag{1}$$

$$\frac{2Y + u_\alpha^2 - K}{2(n + u_\alpha^2)} < p < \frac{2Y + u_\alpha^2 + K}{2(n + u_\alpha^2)} \tag{2}$$

$$K = u_\alpha \sqrt{u_\alpha^2 + 4Y(1 - p^*)}$$

where:

p^* – fraction of concentrations below the determination limit [-], n – sequence size [-], Y – number of concentrations below the determination limit [-], u_α – critical value of the normal distribution for significance level α [-].

Table 3 Sizes and frequencies of analysed heavy metal concentration below determination limit

Heavy metal	Zn	Cu	Cd	Cr
Size n* of concentration determinations [-]	95	95	95	95
Number of results below the determination limit [-]	63	95	49	8
Fraction of results below the determination limit p^* [-]	0.663	1.000	0.516	0.084
Confidence level intervals for p^* (α=0.05) [-]	0.563÷0.750	0.961÷1.000	0.417÷0.614	0.043÷0.157
* – size n was applied in Formula 3				

The concentrations for each of the metals were arranged in an ascending order, matching each value c_i with an empirical distribution function value F^*, calculated based on Formula 3, recommended in the case of asymmetric distributions [12].

$$F^*(c_{i(i)}) = \frac{i - 0{,}3}{n + 0{,}4} \tag{3}$$

where:

i – number of value $c_{i(i)}$ in the series [-].

Due to the fact that it is the easiest to determine a distribution function for a random variable with a Lognormal distribution using normal distribution tables ($F_X(x) = F_Y(\ln x)$), the values $F^*(c_{i(i)})$ and fractions p^* were assigned with appropriate critical values of a standardized normal random variable w.

In order to avoid inconvenience associated with negative values w (for probabilities below 0.5), the authors introduced a variable w' – so-called. *probit* [12], acc. to Formula 4.

$$w' = w + 5. \tag{4}$$

The magnification value is selected arbitrarily, as long as the values w' are positive.

Next, each metal was subjected to the analysis of regression between the controlled random variable w' and the distributing random variable $\ln c_i$. Linear regression coefficients $\ln c_i = \mathrm{A} + \mathrm{B} \cdot w'$ were estimated using the least square method (LSM). Critical values of a standardized normal random variable w and w' were used to determine parameters μ ($\mu = \ln c_i = \mathrm{A} + \mathrm{B} \cdot w'$ for $w = 0$) and σ (difference between μ and $\ln c_i = \mathrm{A} + \mathrm{B} \cdot w'$ for $w = -1$) for the distributions of the variable $\ln c_i$ (normal distribution). This was used as a base to calculate selected variable distribution parameters c_i (Lognormal distribution). The analysis results are shown in Tables 4 and 5.

Table 4 Summary of analysis results for the regression between variables w' and $\ln c_i$

Heavy metal	Zn	Cu	Cd	Cr
Size of pair set (w'; $\ln c_i$)* [-]	33	1	47	88
A	-8.7387	–	-8.4286	-4.9620
B	0.8045	–	0.7914	0.3953
R^2	0.9621	–	0.9521	0.9866
* – given p^*				

Attachment 1, Tables Z1 – 2 to Z1 – 5 summarize metal (zinc, copper, cadmium and chromium) concentrations in ascending order with values of the empirical distribution function, critical value of standardized normal distribution w and probit w', as well as the values of the random variable c_i' corresponding to the theoretical distribution function values.

The proposed approach (concentration treated as a censored random variable) partially returns the lost information due to insufficient accuracy of the measuring device in the form of a frequency

variable p^*, which enables determining sample distribution parameters. The described scheme cannot be applied when the metal concentration in all tested eluates is below the determination limit, as it was with copper (Table 3). In such a case, the frequency $p^* = 1$ implies only a single pair of variables $w' - \ln c_i$ (Table 4).

4. COMPARISON OF CONCENTRATION VARIABLE DISTRIBUTION PARAMETERS FOR DIFFERENT CALCULATION METHODS

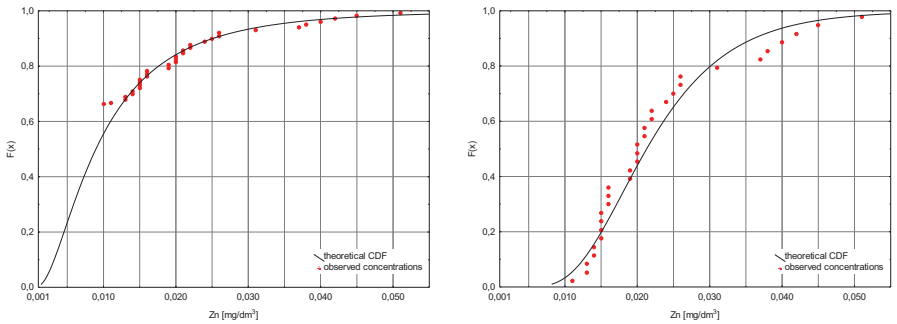
Table 5 contains selected statistics calculated using a censored variable scheme, as well as – for comparison – results obtained when taking into account the following, classic approaches to dealing with missing data (unidimensional analysis):

a) Omission of observations with missing data (truncated distribution) – calculations conducted with a complete omission of cases (listwise deletion), in which the concentration was below the determination limit (rejection from the analysis). Other stages as in the example of estimating censored distribution parameters. In the case of unidimensional analysis, the approach is the same as removing pair observations (pairwise deletion).
b) Single constant imputation – missing data replaced (compensated) with a value equal to half the range from zero to the concentration determination limit (Table 1). Next, the Lognormal distribution parameters were determined, as in the example of estimating censored distribution parameters.
c) Stochastic regressive imputation – missing data were supplemented with the concentration value calculated based on regression analysis (homoscedasticity condition assumed satisfied) similarly to the example of estimating censored distribution parameters, with the difference that in regression model only the cases of concentrations above the determination limit were used (truncated distribution). Then the random component was determined (normal distribution with a mean of zero and variance estimated by a regression model for the cases of concentrations above the determination limit). The MT19937 [16] algorithm used when generating the random value. Then the random component was added to each of the values of $\ln c_i$ estimated by regression model. Such prepared data was subject to another regression analysis, which was used as a base to determine the Lognormal distribution parameters (similarly to the example of estimating censored distribution parameters).

Table 5 Comparison of computational approaches

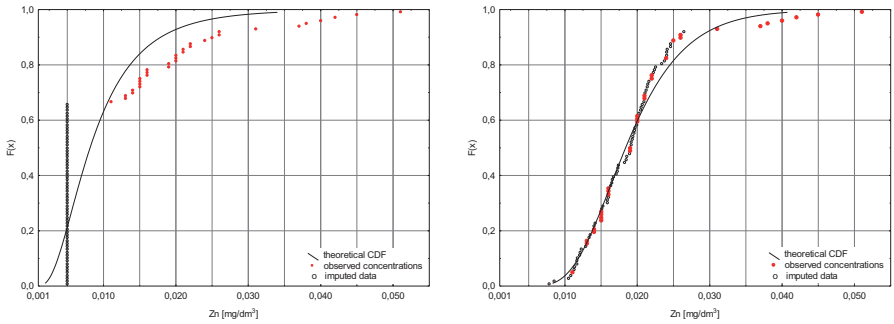
Heavy metal	Zn [mg/dm^3]			Cd [mg/dm^3]			Cr [mg/dm^3]		
	Me[X]*	SD[X]	Distribution parameters	Me [X]	SD[X]	Distribution parameters	Me[X]	SD[X]	Distribution parameters
Distribution censoring	0.0090	0.0118	μ = -4.7160 σ = 0.8045 R^2** = 0.9621	0.0114	0.0146	μ = -4.4717 σ = 0.7914 R^2 =0.9521	0.0505	0.0225	μ = -2.9853 σ = 0.3953 R^2 = 0.9866
Distribution cutting	0.0213	0.0010	μ = -3.8495 σ = 0.4117 R^2 = 0.9473	0.0218	0.0120	μ = -3.8255 σ = 0.4655 R^2 = 0.9449	0.0539	0.0198	μ = -2.9206 σ = 0.3377 R^2 = 0.9579
Single constant imputation	0.0081	0.0067	μ = -4.8103 σ = 0.6145 R^2 = 0.6908	0.0102	0.0113	μ = -4.5852 σ = 0.7355 R^2 = 0.8034	0.0484	0.0270	μ = -3.0284 σ = 0.4718 R^2 = 0.9256
Stochastic regressive imputation	0.0184	0.0069	μ = -3.9962 σ = 0.3420 R^2 = 0.9680	0.0180	0.0088	μ = -4.0190 σ = 0.4287 R^2 = 0.9628	0.0512	0.0210	μ = -2.9725 σ = 0.3705 R^2 = 0.9796
* The Lognormal distribution (with positive skewness) density function is asymmetric. Hence, a median is a better measure for the central tendency within such a distribution than the arithmetic mean [12].									
** Matching of the theoretical distribution function regression model with the results.									

Figure 3 compares the theoretical distribution function of zinc concentration distribution with empirical data for the applied computational approaches.



Censored distribution

Truncated distribution



Single constant imputation Stochastic regressive imputation

Fig. 3 Comparison of empirical and theoretical distribution functions for the approaches in question

The analysis of values listed in Table 5 and Figure 3 indicates that the case of rejecting results characterized by a concentration below the determination level (truncated distribution), thus, omitting such data in the regression model, leads to the loss of a certain amount of information on the continuous variable, the greater the more cases omitted. This results in the inference process omitting the non-random cause of missing data occurrence, hence, impacts the form of the variable distribution. In the case in question, this procedure artificially inflates the median, while simultaneously deflating the variance, which can clearly lead to incorrect conclusions. However, matching the observed results with the theoretical (truncated) distribution (fitting technique) is significantly good ($R^2 = 0.9473$).

Imputation is an alternative to removing missing data. The easiest way would be to replace non-observed variable sections with an average value [5], however, in this case this would not make sense, since a part of the lost information would be replaced with data obtained from a truncated distribution, which significantly inflates the central measure of concentration. In the light of the above, imputation was performed using a value equal to half the interval from zero to the determination limit (different for each metal). The outcome of this operation was (Table 5) a lower value of the median and standard deviation relative to censored distribution, which could itself indicate the approach correctness, but for the analysis of Figure 3 and the statistic R^2 (Table 5), which points to a much worse fit of the results to the theoretical distribution. The method, which seemingly maintains more information (does not reject missing data cases), in practice also leads to biased estimated parameters.

Stochastic regressive imputation is a more complex form of replacing missing data. In the case of the discussed example, owing to the adoption of this approach, the outcome was the best fit of "empirical" data to the theoretical distribution function (Table 5), however, this results from the deterministic (to

a certain extent) nature of determining the central measure, which also leads to lowering of the variance, hence, more results below the determination limit. Smaller differences between the statistics, relative to distribution censoring, were observed along with a decrease in the number of results below the determination limit.

The stochastic regressive imputation, which is considered to be one of the most effective classic methods of dealing with missing data [3], is characterized by a greater workload relative to distribution censoring, as well as it is still based on truncated distribution, which modifies the variable distribution form (e.g. variance reduction).

The analysis results confirm the missing data generating mechanism of the MNAR type.

5. SUMMARY AND CONCLUSIONS

The presented classic approaches to dealing with missing data (distribution truncation, single constant imputation and stochastic regressive imputation) in studying the leachability of heavy metals from a hardening slurry (MNAR mechanism for generating missing data) do not provide satisfactory results. Dealing with a non-observed section of variables is not intended to replace the missing values (especially in natural or technical sciences), but rather to improve the estimation of population parameters [3], which literally implements the censored random variable scheme. The information lost due to missing data are partially recovered in the form of frequencies p^* (missing data incidence frequencies), which is why it is not necessary to replace missing data with imputed values. This procedure does not modify the variable distribution form, which is why it ensures the objective and efficient (no variance reduction) estimation of distribution parameters within studies affected by certain restrictions of technical or physical nature, with a relatively low workload. In order to utilize it, it is necessary to know the frequency of missing data incidence and their cause, whereas the knowledge of the population distribution type, though not required, will have a beneficial impact on the estimation quality. Furthermore, this method is simple to implement and does not require specialized software. It can be briefly described in four steps:

1. Analysing the cause behind the missing data and calculating their incidence frequency (p^*).
2. Appropriate arrangement of data (also p^*) and assigning them to empirical distribution function values (taking into account the type of the variable theoretical distribution).
3. Assigning Critical values of a standardized normal random variable values to empirical distribution function values.

4. Regression analysis and calculating censored variable distribution parameters.

The datasets generated during and/or analysed during the current study are available in the Attachment 1.

ACKNOWLEDGEMENTS

Funding: This research was supported by the Strategic Research Project of the Warsaw University of Technology "Circular Economy" and the Dean of the Faculty of Building Services, Hydro and Environmental Engineering of the Warsaw University of Technology [grant number 504/02644 titled Leaching of heavy metals from hardening slurries based on fly ash from thermal treatment of sewage sludge].

REFERENCES

1. D. B. Rubin, "Inference and missing data," Biometrika, vol. 63, no. 3, pp. 581–592, 1976.
2. J. W. Graham, "Missing data analysis: Making it work in the real world," Annual review of psychology, vol. 60, pp. 549–576, 2009.
3. J. W. Graham, P. E. Cumsille, A. E. Shevock, "Methods for handling missing data," Handbook of Psychology, Second Edition, vol. 2, pp. 109–138, 2012.
4. R. J. A. Little, D. B. Rubin, Statistical analysis with missing data, vol. 793. John Wiley & Sons, pp. 581–592, 2019.
5. A. Pokropek, "Wybrane statystyczne metody radzenia sobie z brakami danych," Polskie Forum Psychologiczne, vol. 23, no. 2, pp. 291–310, 2018.
6. C. K. Enders, Applied missing data analysis. Guilford press, 2010.
7. A. Hald, "Statistical theory with engineering applications," 1952.
8. E. L. Korn, "Censoring distributions as a measure of follow-up in survival analysis," Statistics in medicine, vol. 5, no. 3, pp. 255–260, 1986.
9. K. Hamada, N. Takayama, "Censored income distributions and the measurement of poverty," Bulletin of the International Statistical Institute, vol. 47, pp. 617–630, 1977.
10. Z. Kledyński, Integracja i współzależność wybranych kryteriów oceny wodoszczelności betonu. Wydawnictwa Politechniki Warszawskiej, 1993.
11. A. C. Cohen, "Simplified estimators for the normal distribution when samples are singly censored or truncated," Technometrics, vol. 1, no. 3, pp. 217–237, 1959.
12. J. Maksymiuk, F. Wohlmuth, Metody statystyczne w inżynierii elektrotechnicznej, Wyd. 2, Po. Warszawa: Wydawnictwo Politechniki Warszawskiej, 1984.
13. J. R. Benjamin, C. A. Cornell, Rachunek prawdopodobieństwa, statystyka matematyczna i teoria decyzji dla inżynierów. Wydawnictwa Naukowo-Techniczne, 1977.
14. Ł. Szarek, "Leachability of heavy metals from hardening slurries with the addition of fly ashes from thermal treatment of municipal sewage sludge," Warsaw Univeristy of Technology, 2019.
15. W. Oktaba, "Elementy statystyki matematycznej i metodyka do wiadczalnictwa," PWN, Warszawa (in Polish), 1980.
16. M. Matsumoto, T. Nishimura, "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 8, no. 1, pp. 3–30, 1998.

LIST OF FIGURES AND TABLES:

ZMIENNA MIESZANA LOSOWA JAKO FORMA RADZENIA SOBIE Z BRAKAMI DANYCH W BADANIU WYMYWALNOŚCI METALI CIĘŻKICH Z ZAWIESINY TWARDNIEJĄCEJ

Słowa kluczowe: *braki danych; cenzurowana zmienna losowa; rozkład cenzurowany; metale ciężkie; wymywalność;*

STRESZCZENIE:

Braki danych w tablicach wyników badań mogą w znaczący sposób wpływać na jakość analizy, szczególnie w naukach technicznych, gdzie mechanizm generujący braki danych często jest nielosowy, a ich występowanie zależy od części nieobserwowanej badanych zmiennych. W takich przypadkach zastosowanie nieodpowiedniej metody radzenia sobie z brakami danych prowadzi do obciążenia estymowanych parametrów rozkładu.

W artykule przedstawiono stosunkowo prostą w implementacji metodę radzenia sobie z brakami danych powstałymi w wyniku mechanizmu MNAR wykorzystującą rozkład cenzurowany. Procedura ta nie modyfikuje postaci rozkładu zmiennej, przez co zapewnia obiektywne i skuteczne estymowanie parametrów rozkładu w badaniach dotkniętych pewnymi ograniczeniami natury technicznej lub fizycznej, przy stosunkowo niskim nakładzie pracy. Ponadto nie wymaga zastosowania specjalistycznego oprogramowania. Warunkiem koniecznym zastosowania metody jest znajomość częstości występowania braków danych oraz ich przyczyny.

Sposób estymacji parametrów rozkładu cenzurowanego zmiennej losowej przedstawiono na przykładzie badania wymywalności wybranych metali ciężkich z zawiesiny twardniejącej. Wyniki analizy porównano z klasycznymi sposobami radzenia sobie z brakami danych: pominięciem obserwacji z brakami danych, imputacją oraz stochastyczną imputacją regresyjną.

Received: 15.07.2020, Revised: 07.12.2020