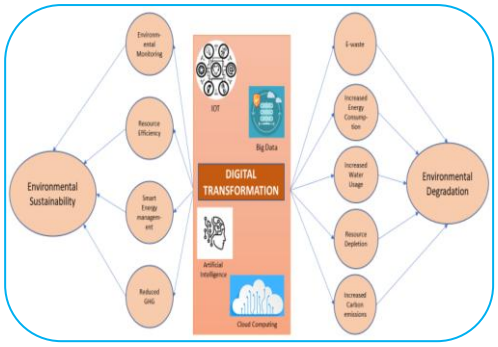
REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
VOLUME - 11 | ISSUE - 6 | MARCH - 2022

PROBLEMS AND PERSPECTIVES OF DATA TRANSFORMATION IN ENVIRONMENTAL STUDIES

Mr. Patil Sahebagouda S.
Department of Zoology, Sangameshwar College, Solapur Autonomous.
Email id : sspatildprc@gmail.com

ABSTRACT :



Data transformation is a crucial technique in ecological research for managing outliers and reducing skewness in datasets. Common methods include log transformation (log x, log(x+1)), square-root transformation, and arcsine transformation, each suited for specific data types. Log transformation is widely used for converting large values while preserving their relative scale, though it requires modifications such as log(x+1) transformation to handle zero and fractional values. Square-root transformation is effective for count data but unsuitable for negative values, whereas arcsine transformation is ideal for proportional data analysis. Among these, log transformation is the most versatile, particularly in regression analyses, ensuring homoscedasticity and linearity in ecological studies. The log(x+1) method is frequently applied to species count data to manage zero values, enhancing statistical accuracy and reliability.

KEYWORDS: *Data transformation , square-root transformation, and arcsine transformation.*

INTRODUCTION:

As climate change and environmental deterioration continue to escalate, the demand for efficient analysis of environmental data is becoming more critical Gupta et al.,2021). The process of transforming environmental data entails converting unprocessed data into organized formats that are appropriate for analysis and informed decision-making (Dibekulu, 2020). However, various obstacles impede successful transformation, highlighting the need for enhanced frameworks and technologies.

Transforming environmental data is crucial for examining and understanding ecological patterns, climate change, and sustainability indicators. Nonetheless, it poses several challenges, such as inconsistencies in data, difficulties with integration, and concerns regarding accuracy (Wang, X., et al., 2010). This study investigates the primary issues associated with environmental data transformation and provides insights on enhancing methods, standardizing data, and leveraging technological progress to improve data accessibility and trustworthiness.

The following are the key issues in Environmental Data Transformation

Data Consistency and Variability:

Environmental data is gathered from a variety of sources, such as satellites, sensors, and ground measurements. The discrepancies in data formats, measurement units, and collection methods create challenges for integration and comparative analysis (Elise F., 2021).

Data Integrity and Precision Problems: Inaccuracies in data may stem from sensor failures, human mistakes, or environmental disturbances. Maintaining high-quality data is essential for dependable analysis and informed decision-making (Sylvie Koziel., 2021).

Absence of Standardization: The lack of standardized data formats and protocols results in inefficiencies in sharing and integrating data across different organizations and research institutions (Gal, Michal & Rubinfeld, Daniel., 2018).

Scalability Issues: As environmental data expands in volume and intricacy, traditional methods of transformation become inadequate, leading to processing and analysis challenges (Uthayasankar Sivarajah,2017).

Security and Privacy Challenges : With growing digitalization, safeguarding sensitive environmental information from unauthorized access and cyber threats has emerged as a major concern (Yuchong Li, 2021).

MATERIAL AND METHODS

The collected data is binary data measured in different scales like Temperature measured in Degree Celsius, Concentration measured in milligrams per litter, Electric conductivity measured in mS/m, Turbidity in NTU etc. Direct application without processing the data or without standardizing the data may lead to wrong interpretation and lead to wrong conclusions as the raw data may or may not normally distributed and their standard deviations are not homogenous, therefore data transformation is essential before application of statistical tests to avoid the wrong estimation and results.

RESULT AND DISCUSSION

The data is collected from water samples of Kurnur dam in 2014 where the water sample was analysed for various physical, chemical and biological factors as shown in the Table No. 1 to 3.

Table No. 1 Physical Data with variable units.

Dates	Temp °C	pH	EC mS/m.	Turbidity NTU	sechi cm	TDS mg/l	TSS mg/l	TS mg/l
19-01-2014	18	6.3	6.9	4.27	96.31	379.42	10.39	308.03
16-02-2014	20	6.9	7.16	5.36	94.66	354.81	11.22	366.03
16-03-2014	26	6.7	8.69	6.11	87.02	398.11	12.30	410.41
13-04-2014	27	6.9	7.90	6.51	83.6	354.81	11.22	366.03
11-05-2014	29	6.2	7.53	6.4	84.5	338.84	11.22	350.06
08-06-2014	31	6.5	8.89	7.29	77.8	398.11	12.30	410.41
06-07-2014	28	6.9	9.90	7.47	76.59	446.68	13.18	459.86
03-08-2014	23	8.3	12.12	8.3	71.6	562.34	16.22	578.56
31-08-2014	26	8.5	12.04	12.78	54.4	575.44	17.38	592.82
28-09-2014	19	8.9	13.40	11.86	43.18	616.60	23.99	640.59
26-10-2014	19	8.2	11.00	11.47	58.29	549.54	16.22	565.76
23-11-2014	19	7.8	9.94	7.06	79.39	478.63	15.14	493.77
21-12-2014	17	7.6	8.77	7.45	76.7	446.68	14.13	460.81

Table No.2 Chemical data measured in mg/ L

Dates	DO	BOD	COD	Ca	Mg	TH	NO_3-	NO_2-	P	S	Cl
Units	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L
19-01-2014	9.4	2.39	13.51	20.27	11.23	31.50	2.13	0.27	0.09	12.86	28.31
16-02-2014	8.9	2.34	13.49	20.34	11.37	31.71	1.89	0.21	0.10	13.75	30.60
16-03-2014	7.05	3.03	13.18	21.55	11.84	33.39	1.71	0.21	0.12	15.3	40.20
13-04-2014	6.64	3.32	15.14	20.34	11.37	31.71	1.45	0.16	0.14	16.09	37.26
11-05-2014	6.1	3.98	18.20	19.88	11.18	31.06	1.45	0.16	0.17	15.35	35.41
08-06-2014	5.51	3.59	23.99	21.55	11.84	33.39	1.38	0.17	0.21	17.51	42.04
06-07-2014	6.39	3.84	25.12	22.82	12.33	35.15	1.56	0.21	0.13	18.3	46.41
03-08-2014	8.14	4.57	23.44	25.61	13.38	38.99	1.67	0.27	0.09	19.3	55.31
31-08-2014	7.21	6.86	57.54	25.91	13.49	39.40	2.33	0.40	0.19	18.69	53.37
28-09-2014	9.41	6.06	52.48	26.82	13.82	40.64	2.73	0.65	0.08	22.28	61.68
26-10-2014	9.44	5.52	45.71	25.32	13.27	38.59	2.18	0.35	0.07	17.05	46.58
23-11-2014	9.84	4.52	38.90	23.63	12.64	36.27	2.16	0.33	0.06	13.81	43.67
21-12-2014	10.51	3.46	32.36	22.82	12.33	35.15	2.19	0.31	0.13	12.59	36.48

Table. 3 Biological data in Organisms/L

Zooplankton	Rotifera	Copepoda	Cladocera	Ostracoda	Protozoa
19-01-2014	92	45	43	31	13
16-02-2014	103	38	36	30	12
16-03-2014	112	29	28	25	10
13-04-2014	117	18	21	21	9
11-05-2014	132	21	25	25	12
08-06-2014	118	23	23	27	10
06-07-2014	125	42	29	43	12
03-08-2014	105	21	24	48	20
31-08-2014	112	26	21	38	21
28-09-2014	103	34	26	34	19
26-10-2014	89	48	38	35	16
23-11-2014	82	51	51	28	13
21-12-2014	92	45	43	31	13

The primary goal of data transformation is to address outliers (Zurr et al., 2007) and to eliminate skewness in the data (Altman DG et al., 1996). Various methods of data transformation exist, such as log transformation (log of x), log(x+1) transformation, square-root transformation, Arcsine transformation, and others. Multiple sources were examined to identify the types of data transformation that are appropriate for ecological research, particularly regarding count data in biology, which detailed the log transformation, square root transformation, and Generalized Linear Model (GML) (Anne P. St-Pierre et al., 2018).

All calculations should be performed using the transformed data, and any presentation of the data should reflect the back-transformed values (Swinscow TD et al., 2003). There are two forms of log transformation: transforming data to a logarithm with base 10 or using the natural log (base e). Log transformation influences the values of the slope and intercept in regression analysis, yet it converts large numbers into smaller values while preserving data magnitude. A key issue with log

transformation is the existence of negative values, fractional values, and zero values (Robert B. O'Hara et al., 2010). To address this issue, values that are zero or fractional are incremented by 1 or another common value, a process referred to as log(x+1) transformation (Anne P. St-Pierre et al., 2018). The square root transformation is unsuitable for data with negative numbers but is often used with count data (Bartlett, 1936).

Arcsine transformation, also known as Arcsine square root transformation or angular transformation, involves taking the square root of values labeled as Arcsine numbers multiplied by two; in some instances, it may not be multiplied by two to prevent shifting the arcsine scale from zero to pi and instead restrict it to pi/2. Arcsine numbers span from 0 to 1 or from zero to pi, and the conversion is based on value proportions. This transformation is appropriate for values in the range of 0 to 1 and is commonly utilized in proportion studies, making it inappropriate for measurement variables; however, it is applied in multivariate analyses where ordination or cluster analysis is necessary within ecological studies. This method is especially effective for species analysis (Sokal et al., 1995).

Log transformation is generally favored as it encompasses all real numbers rather than confining itself to a specific range. The log transformation can extend the log scale from positive infinity to negative infinity. Its intuitive nature aids in interpreting slopes and logistical regressions while also establishing a relationship between means and variances in binomial data (Warton et al., 2011). The log(x+1) transformation is the simplest and most suitable for datasets containing zero or fractional values. In log transformation, ecological count values are increased by one common number, which minimally affects the transformation process, allowing for the treatment of zero or fractional values. Directly applying log transformation to fractional values could yield negative results alongside some positive ones, complicating comparisons (McDonald, J.H., 2014).

In our research, we utilized the log (x) transformation for physicochemical factors that were measured on varying scales and in different units. For biological data, such as the species count of zooplankton, we applied the log transformation of log (x+1) to convert any zero values into a complete number. Similar transformation methods have been used in other ecological studies (Negreiros, Natalia Felix et al., 2010; Nikolaos Th et al., 2009). This approach also ensures homoscedasticity and linearity between the dependent and independent variables (Nosrati et al., 2015; Xiong, W et al, 2016; Caitlin A.E. et al., 2018).

Approaches to Improve Environmental Data Transformation

Adoption of Uniform Data Formats: Creating and implementing standardized protocols like NetCDF, CSV, and XML can enable smooth data integration and sharing among researchers and organizations (Briney, Kristin et al., 2020).

Utilization of AI and Machine Learning: Machine learning techniques can assist in identifying anomalies, enhancing data precision, and automating the transformation process, thereby minimizing human involvement and errors (Sarker, I.H. 2021).

Cloud-Based Data Storage and Processing: Cloud computing provides efficient storage, processing, and immediate access to vast amounts of environmental data, addressing the issues related to scalability(Ibrahim Abaker et al., 2015).

Enhancing Cybersecurity Protocols: Improving encryption, access controls, and authentication systems can safeguard environmental data against security breaches and unauthorized access (Borky JM., 2018).

Collaboration Among Stakeholders: Governments, researchers, and private entities must work together to establish unified frameworks for data collection, sharing, and analysis, ensuring consistency and accuracy across different datasets (James Scheibner et al.,2020).

CONCLUSION

Data transformation is essential for addressing outliers and eliminating skewness in ecological datasets. Various transformation methods, including log transformation (log x, log(x+1)), square-root

transformation, and arcsine transformation, are used to improve data interpretation and statistical modeling.

Log transformation, which can use either base 10 or the natural logarithm (base e), is widely applied to convert large numbers into manageable values while maintaining their relative magnitude. However, it poses challenges when dealing with negative, fractional, or zero values, which can be addressed through log(x+1) transformation. Square-root transformation is effective for count data but unsuitable for negative values, while arcsine transformation is primarily used for proportional data in ecological studies.

Among these methods, log transformation is preferred for its broad applicability and intuitive interpretation in regression analyses. In ecological research, log(x+1) transformation is commonly used for species count data to handle zero values effectively. This method ensures homoscedasticity and linearity, facilitating accurate statistical analyses.

REFERENCES

1. Altman DG, Bland JM. Detecting skewness from summary information. BMJ. 1996;313:1200
2. Anne P. St-Pierre, Violaine Shikon, and David C. Schneider , 2018 "Count data in biology—Data transformation or model reformation? Ecol Evol. 2018 Mar; 8(6): 3077–3085. Published online 2018 Feb 16. doi: 10.1002/ece3.3807 PMCID: PMC5869353 PMID: 29607007
3. Bartlett M.S. 1936 "The Square Root Transformation in Analysis of Variance" Supplement to the Journal of the Royal Statistical Society Vol. 3, No. 1 (1936), pp. 68-78 JSTOR DOI: 10.2307/2983678 https://www.jstor.org/stable/2983678
4. Borky JM, Bradley TH. Protecting Information with Cybersecurity. Effective Model-Based Systems Engineering. 2018 Sep 9:345–404. doi: 10.1007/978-3-319-95669-5_10. PMCID: PMC7122347.
5. Briney, Kristin & Coates, Heather & Goben, Abigail. (2020). Foundational Practices of Research Data Management. Research Ideas and Outcomes. 6. 10.3897/rio.6.e56508.
6. Caitlin A.E. McKinstry Robert W.Campbell 2018 "Deep Sea Research Part II: Topical Studies in Oceanography" Volume 147, January 2018, Pages 69-78 Deep Sea Research Part II: Topical Studies in Oceanography Seasonal variation of zooplankton abundance and community structure in Prince William Sound, Alaska, 2009–2016
7. Dibekulu, Dawit. (2020). An Overview of Data Analysis and Interpretations in Research. 1-27. 10.14662/IJARER2020.015.
8. Elise F Zipkin, Erin R Zylstra, Alexander D Wright, Sarah P Saunders, Andrew O Finley, Michael C Dietze, Malcolm S Itter, Morgan W Tingley (2021) "Addressing data integration challenges to link ecological processes across scales"Special Issue:Macrosystems Biology - Challenges and Successes Volume19, Issue1, Pages 30-38, https://doi.org/10.1002/fee.2290.
9. Gal, Michal & Rubinfeld, Daniel. (2018). Data Standardization. SSRN Electronic Journal. 10.2139/ssrn.3326377.
10. Gupta, Suraj & Aga, Diana & Pruden, Amy & Zhang, Liqing & Vikesland, Peter. (2021). Data Analytics for Environmental Science and Engineering Research. Environmental Science & Technology. 55. 10.1021/acs.est.1c01026.
11. Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan, The rise of "big data" on cloud computing: Review and open research issues, Information Systems, Volume 47, 2015, Pages 98-115, ISSN 0306-4379, https://doi.org/10.1016/j.is.2014.07.006.
12. James Scheibner, Marcello Ienca, Sotiria Kechagia, Juan Ramon Troncoso-Pastoriza, Jean Louis Raisaro, Jean-Pierre Hubaux, Jacques Fellay, Effy Vayena, Data protection and ethics requirements for multisite research with health data: a comparative examination of legislative governance frameworks and the role of data protection technologies, Journal of Law and the Biosciences, Volume 7, Issue 1, January-June 2020, lsaa010, https://doi.org/10.1093/jlb/lsaa010
13. McDonald, J.H. (2014) Handbook of Biological Statistics. Sparky House, Maryland.

14. Negreiros, Natalia Felix, Santos-Wisniewski, Maria José dos, Santos, Renata Martins dos, & Rocha, Odete. (2010). The influence of environmental factors on the seasonal dynamics and composition of Rotifera in the Sapucaí River arm of Furnas Reservoir, MG, Brazil. Biota Neotropica, 10(4), 173-182. https://dx.doi.org/10.1590/S1676-06032010000400023
15. Nikolaos Th.SkoulikidisIoannisKaraouzasKonstantinos C.Gritzalis , 2009 "Identifying key environmental variables structuring benthic fauna for establishing a biotic typology for Greek running waters" Limnologica Volume 39, Issue 1, February 2009, Pages 56-66.
16. Nosrati, K. Model. Earth Syst. Environ. 2015 1: 19. https://doi.org/10.1007/s40808-015-0021-
17. Robert B O'HaraD.Johan Kotze ., 2010, " Do not log transform count data" Methods in Ecology and EvolutionVolume 1, Issue 2, 2010 https://doi.org/10.1111/j.2041-210X.2010.00021.x
18. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
19. Sokal, R.R., and F.J. Rohlf. 1995. Biometry. Freeman, New York, 887 p
20. Swinscow TD, Campbell MJ. 2003 " Statistics at square one" . 10th ed. New Delhi: Viva Books Private limited; 2003. (Indian)
21. Sylvie Koziel, Patrik Hilber, Per Westerlund, Ebrahim Shayesteh, Investments in data quality: Evaluating impacts of faulty data on asset management in power systems, Applied Energy,Volume 281,2021,116057,ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2020.116057.
22. Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody, Critical analysis of Big Data challenges and analytical methods, Journal of Business Research, Volume 70, 2017, Pages 263-286, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2016.08.001
23. Wang, X., Huang, LP., Zhang, Y. et al. A Solution of Data Inconsistencies in Data Integration — Designed for Pervasive Computing Environment. J. Comput. Sci. Technol. 25, 499–508 (2010). https://doi.org/10.1007/s11390-010-9340-2
24. Warton, D.I., and F.K.C. Hui. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92:3–10.
25. Xiong, W., Li, J., Chen, Y. et al. Determinants of community structure of zooplankton in heavily polluted river ecosystems. Sci Rep 6, 22043 (2016) doi:10.1038/srep22043
26. Yuchong Li, Qinghui Liu, A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments, Energy Reports, Volume 7, 2021, Pages 8176-8186, ISSN 2352-4847, https://doi.org/10.1016/j.egyr.2021.08.126.
27. Zuur, Alain & Ieno, Elena & Smith, Graham. (2007). Analysing Ecological Data. 10.1007/978-0-387-45972-1.