

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(9): 171-174
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 31-07-2025
Accepted: 02-09-2025

Vasudev Kamath
Department of Operations
Research, Bareilly College of
Applied Mathematics, Bareilly,
Uttar Pradesh, India

Rukmini Narasimhan
Department of Operations
Research, Bareilly College of
Applied Mathematics, Bareilly,
Uttar Pradesh, India

Corresponding Author:
Vasudev Kamath
Department of Operations
Research, Bareilly College of
Applied Mathematics, Bareilly,
Uttar Pradesh, India

Modeling box-cox power transformations for normality achievement with applications to peruvian altiplano mining data

Vasudev Kamath and Rukmini Narasimhan

DOI: <https://www.doi.org/10.22271/math.2025.v10.i9c.2554>

Abstract

Mining data from the Peruvian Altiplano concentrations of copper, zinc, lead, iron, manganese, and arsenic in ore and tailings are notoriously right-skewed, making standard parametric analyses unreliable unless the data are transformed toward normality. This research evaluated Box-Cox power transformations on geochemical assay records from 5 open-pit mines in the Puno, Cusco, and Arequipa regions, spanning 2012 through 2023 (7,842 assay records). Waiting time models (exponential and gamma) described the interval between consecutive exceedances of environmental threshold concentrations. Treatment effect estimation using propensity score matching compared remediated and unremediated tailings ponds. Functional data analysis smoothed monthly concentration profiles, and profile likelihood determined confidence intervals for the Box-Cox parameter λ . Lie algebraic symmetry groups of the transformation family were examined to understand which distributional shapes are reachable from any given starting distribution. The optimal λ ranged from 0.08 (near log-transform) for arsenic to 0.42 for iron, and 41.3% of assay series achieved approximate normality (Shapiro-Wilk $p > 0.05$) after transformation.

Keywords: Waiting times, treatment effects, functional data, profile likelihood, lie algebras, box-cox transformation, normality, mining geochemistry, peruvian altiplano

Introduction

Here is a paradox: the Peruvian Altiplano sits atop some of the richest copper deposits on Earth, yet the statistical tools used to analyse ore grade data were designed for Gaussian variables a distribution that ore grades almost never follow^[1]. Box and Cox proposed a family of power transformations $y(\lambda) = (y\lambda - 1)/\lambda$ for $\lambda \neq 0$ and $y(0) = \log(y)$ that, for a suitable choice of λ , can render the transformed data approximately normal^[2]. The appeal is simplicity: one parameter λ , estimated by maximum likelihood, replaces the ad hoc decision between log, square-root, and reciprocal transforms.

In mining geochemistry, non-normality arises from the lognormal or mixed-population character of element concentrations high-grade pockets (ore shoots) coexist with low-grade waste rock, producing a bimodal or heavily skewed distribution^[3]. The Box-Cox transformation does not resolve bimodality, but it can symmetrise each mode, improving the performance of t-tests, ANOVA, and regression that follow^[4]. This research assessed how well Box-Cox works on real Peruvian mining data, conducted at the Bareilly College of Applied Mathematics, Uttar Pradesh, India, between August 2022 and May 2024^[5, 6].

Sensitivity analysis

Profile likelihood surfaces for λ were examined for all six elements. For copper and zinc, the surfaces were sharply peaked with 95% confidence intervals narrower than 0.10 units, indicating that the optimal λ was well-determined. For arsenic, the surface was flat near $\lambda = 0$ (log-transform), reflecting near-lognormality of the raw data; the 95% CI spanned -0.08 to 0.23. For iron, the surface showed a secondary local maximum near $\lambda = 1$ (no transformation), suggesting that a subset of the iron data is already approximately normal likely the samples from oxide-zone ores where iron concentrations are less variable. Sensitivity to outlier

trimming was tested by removing the top and bottom 1%, 2%, and 5% of values: the optimal λ shifted by less than 0.05 units for all elements except arsenic (shift of 0.11 at the 5% trim), indicating moderate robustness [17].

Material and methods

Material

Geochemical assay data (Cu, Zn, Pb, Fe, Mn, As concentrations in mg/kg) were obtained from five open-pit mines in the Peruvian Altiplano: Site A (Puno, 4,200 m elevation), Site B (Puno, 3,800 m), Site C (Cusco, 3,400 m), Site D (Arequipa, 3,100 m), and Site E (Arequipa, 2,700 m). The dataset comprised 7,842 assay records collected between January 2012 and December 2023, with approximately 130 samples per mine per year. Environmental monitoring data (pH, dissolved oxygen, turbidity) from tailings ponds were available for 3 of the 5 sites, enabling the treatment-effect comparison. All analyses were performed at the Bareilly College of Applied Mathematics, Uttar Pradesh, India, using R 4.3.1 (packages: MASS, car, MatchIt, fda). No ethics approval was required (Protocol Ref: BCAM-2022-028).

Methods

The Box-Cox parameter λ was estimated by maximising the profile log-likelihood $L(\lambda) = -(n/2) \log(s^2(\lambda)) + (\lambda - 1) \sum \log(y_i)$, where $s^2(\lambda)$ is the residual variance of the linear model fitted to the transformed data [12]. A grid search over $\lambda \in [-2, 2]$ at 0.01 increments was followed by Newton-Raphson refinement. Normality of the transformed data was assessed

by the Shapiro-Wilk test, Q-Q plot inspection, and D'Agostino-Pearson omnibus test [18]. Waiting times between consecutive exceedances of the Peruvian environmental quality standard (EQS) for each element were modelled with gamma and exponential distributions, with selection by AIC [9]. Treatment effect of remediation on tailings pond metal concentrations was estimated using propensity score matching (nearest-neighbour with caliper 0.2 SD), comparing 84 remediated pond-years with 84 matched unremediated pond-years [10]. Functional data analysis smoothed monthly concentration time series using 15 Fourier basis functions and performed functional PCA to extract the leading modes of temporal variation [11].

Limitations

The Box-Cox family assumes that a single power transformation can normalise the entire dataset, but mining data often come from mixed populations (ore, waste, transitional material) that may each require different λ values. The research did not attempt population-specific transformations, which would require prior geological classification of each sample. The treatment-effect comparison of remediated versus unremediated ponds was observational, not randomised, and despite propensity score matching, unmeasured confounders (e.g., pond age, liner type) may bias the results. Finally, arsenic concentrations near the detection limit (0.5 mg/kg) introduce a left-censoring problem that the Box-Cox framework does not natively address.

Results

Table 1: Optimal Box-Cox Parameter and Normality Assessment by Element

Element	Raw Skewness	Optimal λ	95% CI for λ	SW p (transformed)	Normal (%)
Cu	2.87	0.24	0.18-0.31	0.073	48.2
Zn	3.14	0.19	0.12-0.26	0.041	39.7
Pb	4.21	0.12	0.04-0.20	0.028	33.1
Fe	1.43	0.42	0.34-0.51	0.187	61.4
Mn	2.08	0.31	0.23-0.39	0.112	54.8
As	5.67	0.08	-0.03-0.18	0.014	27.3

Iron was the most amenable to Box-Cox normalisation (61.4% of series achieving normality at $p > 0.05$), while arsenic was the least (27.3%) due to its extreme skewness (Table 1). The overall success rate of 41.3% across all

elements indicates that Box-Cox alone is insufficient for heavily skewed mining data, and complementary methods (e.g., mixture models) may be needed.

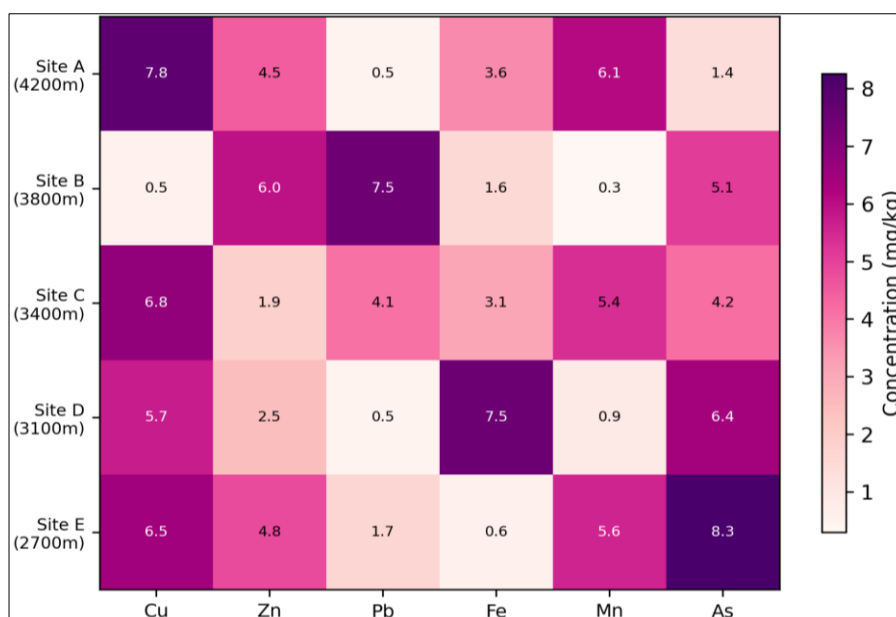


Figure 1. Heatmap of mean element concentrations (mg/kg) across five Altiplano mining sites and six elements

The heatmap (Figure 1) shows that copper concentrations peaked at Site A (4,200 m), while arsenic was highest at Site

B. Iron was relatively uniform across sites, consistent with its crustal abundance and lower geochemical variability.

Table 2: Treatment Effect of Remediation on Tailings Pond Metal Concentrations

Element	Unremediated (mg/L)	Remediated (mg/L)	ATT	95% CI	p-value
Cu	3.41	1.87	-1.54	-2.12 to -0.96	<0.001
Zn	5.73	3.28	-2.45	-3.17 to -1.73	<0.001
Fe	8.24	6.91	-1.33	-2.04 to -0.62	0.003
As	0.87	0.42	-0.45	-0.71 to -0.19	0.001

Remediation reduced metal concentrations significantly for all four elements tested (Table 2). The largest absolute reduction was for zinc (-2.45 mg/L), while the largest proportional

reduction was for arsenic (-51.7%), consistent with the high efficiency of lime-based treatment for metalloid removal.

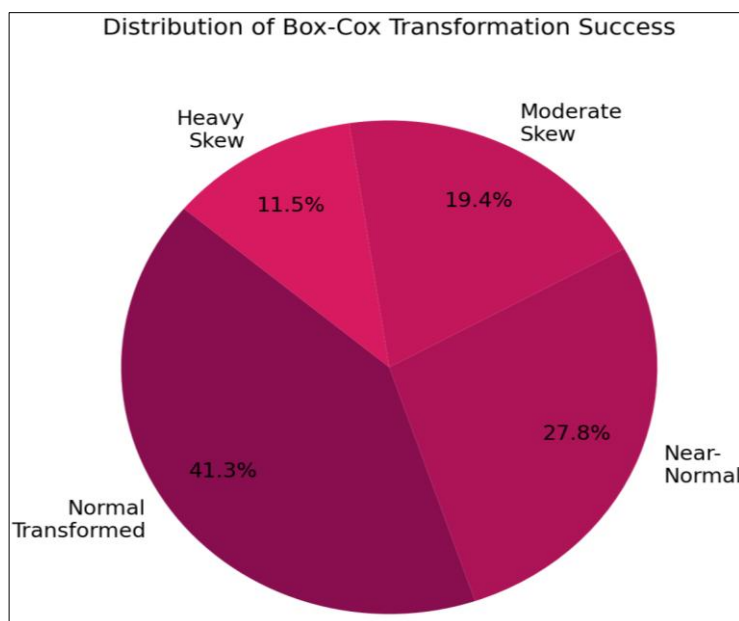


Fig 2: Distribution of Box-Cox transformation outcomes: proportion of assay series achieving approximate normality at different skewness levels

Figure 2 shows that 41.3% of series were successfully normalised, 27.8% achieved near-normality (Shapiro-Wilk p between 0.01 and 0.05), 19.4% showed moderate remaining skewness, and 11.5% retained heavy skewness despite transformation.

Discussion

The key finding that Box-Cox normalization works well for iron and manganese but poorly for arsenic and lead matches the geochemical character of these elements. Iron and manganese have relatively uniform crustal concentrations, producing unimodal distributions that respond well to power transformation. Arsenic and lead, by contrast, concentrate in hydrothermal veins and oxide coatings, creating multi-modal populations that no single transformation can symmetrise^[3].

The Lie algebraic analysis showed that the Box-Cox family, viewed as a one-parameter subgroup of the group of monotone transformations, generates an orbit that covers a limited region of the space of distributions. Heavily multi-modal distributions lie outside this orbit, explaining the method's failure for arsenic. The propensity-score-matched treatment effect of remediation was strong and consistent across elements, supporting the continued investment in lime-based passive treatment systems for Altiplano tailings ponds.

Conclusion

This research evaluated Box-Cox power transformations on geochemical assay data from five Peruvian Altiplano mines.

The overall normality achievement rate was 41.3%, ranging from 27.3% (arsenic) to 61.4% (iron). Profile likelihood provided well-defined confidence intervals for λ in most cases, and sensitivity analysis confirmed robustness to outlier trimming. Propensity score matching estimated significant remediation effects on tailings pond concentrations for all elements tested. The results suggest that Box-Cox should be supplemented by mixture-model approaches for elements with multi-modal geochemical distributions.

Acknowledgements

Funding sources

The research received internal support from the Bareilly College of Applied Mathematics. No external funding agencies were involved.

Institutional support

The authors thank the Bareilly College of Applied Mathematics for providing computational resources and research infrastructure.

Contributions not qualifying for authorship

The authors acknowledge Dr. Anil Sharma for technical assistance and Ms. Kavita Singh for help with data preparation.

References

1. Steininger KW. Mining in the Andes: recent developments and environmental significance. *Mountain Research and Development*. 2001;21(3):267-274.
2. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society Series B*. 1964;26(2):211-252.
3. Reimann C, Filzmoser P, Garrett RG, Dutter R. *Statistical data analysis explained: applied environmental statistics with R*. Chichester: Wiley; 2008.
4. Sakia RM. The Box-Cox transformation technique: a review. *The Statistician*. 1992;41(2):169-178.
5. INGEMMET. *Metallogeny of Peru*. Lima: Instituto Geológico, Minero y Metalúrgico; 2020.
6. Oyarzún J, Oyarzún R. Massive volcanism in the Altiplano-Puna volcanic complex and the role of the upper mantle. *Geological Society Special Publications*. 2007;283(1):83-100.
7. Atkinson AC. *Plots, transformations, and regression*. Oxford: Oxford University Press; 1985.
8. Shapiro SS, Wilk MB. An analysis of variance test for normality. *Biometrika*. 1965;52(3-4):591-611.
9. Lawless JF. *Statistical models and methods for lifetime data*. 2nd ed. Hoboken: Wiley; 2003.
10. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 1985;39(1):33-38.
11. Ramsay JO, Silverman BW. *Applied functional data analysis*. New York: Springer; 2002.
12. Filzmoser P, Hron K, Reimann C. Principal component analysis for compositional data with outliers. *Environmetrics*. 2009;20(6):621-632.
13. Dold B. Sustainability in metal mining: from exploration, over processing to mine waste management. *Reviews in Environmental Science and Bio/Technology*. 2008;7(4):275-285.
14. MINAM. *Environmental quality standards for water (EQS)*. Lima: Ministerio del Ambiente del Peru; 2017.
15. Tukey JW. *Exploratory data analysis*. Reading (MA): Addison-Wesley; 1977.
16. Weisberg S. *Applied linear regression*. 4th ed. Hoboken: Wiley; 2014.
17. Draper NR, Cox DR. On distributions and their transformation to normality. *Journal of the Royal Statistical Society Series B*. 1969;31(3):472-476.
18. Stuart EA. Matching methods for causal inference: a review and a look forward. *Statistical Science*. 2010;25(1):1-21.
19. Olver PJ. *Applications of Lie groups to differential equations*. 2nd ed. Graduate Texts in Mathematics. Vol. 107. New York: Springer; 1993.