

Do not $\log(x+1)$ transform: the bias and alternatives

Vasco Vieira

`vasco.vieira@tecnico.ulisboa.pt`

MARE-NOVA

Cátia Bartilotti

Portuguese Sea and Atmosphere Institute

Jorge Lobo-Arteaga

Portuguese Sea and Atmosphere Institute

Rafael Santos

Portuguese Sea and Atmosphere Institute

David Leitão-Silva

MARE-NOVA

Arthur Veronez

Centre for Functional Ecology

Joana Neves

MARE-NOVA

Ana Brito

MARE-FCUL

Rui Cereja

MARE-FCUL

Diogo Paulo

Centro de Ciências do Mar do Algarve

Francisco Leitão

Centro de Ciências do Mar do Algarve

Article

Keywords: transformation bias, ANOVA, linearization, homoscedasticity, homogeneity of variances, heteroscedasticity

Posted Date: August 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10385620/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

For numerical and statistical analysis, logarithmic transformation is commonly applied to reduce or stabilize variances and to linearize nonlinear relationships, improving the compliance with the assumptions of linear models. However, the log-transform cannot be applied directly to datasets containing zeros as $\log(0)=-\infty$. The most common workaround has been the $\log(x + 1)$ transformation. This transformation belongs to the broader $\log(x + C)$ family of transformations, which has been demonstrated to introduce substantial bias into data analysis. Yet, it remains widely used across the environmental sciences community, often without recognising the biases and misleading inferences it may produce. Here, we demonstrate the bias introduced by the $\log(x + C)$ transform using artificially generated data as well as case studies from the literature. The bias was always large, even when the constant C was small relative to x . Although several alternative transformations can be appropriate under different circumstances, they performed poorly when the data followed an exponential distribution. For exponential models, no satisfactory alternative transformation was identified. In such cases, it is better to implement alternative numerical methods such as GLM.

1. Introduction

In scientific studies, including environmental, ecological and biological among others, the original data often needs to be log-transformed. This typically happens in four situations:

- (1) when finding correlations among variables, the $\log(x)$ transform linearizes non-linear relationships, thus enabling the application of linear analysis tools, of much simpler application than their non-linear counterparts.
- (2) when finding timeseries trends, the $\log(x)$ transform linearizes non-linear responses, thus facilitating the application of time series analysis tools.
- (3) when applying ANOVA or similar analysis over data sets of uneven variances (heteroscedasticity), the $\log(x)$ transform may homogenize their variances (homoscedasticity).
- (4) when working with data (or errors) with skewed distributions (like often happens with abundance data), the $\log(x)$ transform can reshape the distribution into symmetrical Gaussian.

The $\log(x)$ transform has been exposed for introducing serious bias in the data analysis, mainly because back-transforming predictions from log-space produces biased estimates in the original scale (Wilson et al., 1990; Newman 1993; Zeng & Tang, 2011; Vieira et al., 2022a). This is the Retransformation Bias due to Jensen's inequality $E[\log(X)] \neq \log(E[X])$. Another source of bias concerns the additive vs multiplicative nature of the errors (Mascaro et al., 2011). Debate followed with increasing replacement of the $\log(x)$ transformation by modelling alternatives - such as GLM - considering non-linear error structures (St-Pierre et al., 2018).

The $\log(x)$ transform is incompatible with the presence of zeros in the data, given that $\log(0) = -\infty$. For such situations the $\log(x + 1)$ transformation has become commonly used. Its origins date back to the early 20th Century, with no specific work or author associated. This transformation is a particular case of the broader $\log(x + C)$ family, with C being a given constant, for which $C = 0.5$ has been proposed as a better alternative (Yamamura 1999). Further studies have continued to investigate the choice of C , with no universal consensus (Motakis et al., 2006).

Adding to the $\log(x)$ transform bias, the $\log(x + C)$ transform family has its own suite of bias, all relating with the addition of an arbitrary constant C to the raw data. This addition was found strongly affecting means, variances, ordinations, correlations, regressions and rare-species weighting, with such biases becoming more severe when the data contain many zeros or data are sparse (Legendre & Gallagher 2001; Martín-Fernández et al., 2003; Lun 2018; Booeshaghi & Pachter 2021; Bellégo et al., 2022).

Despite increasing criticism and evidence of its limitations, the application of the $\log(x + 1)$ transform is still generalized in the environmental and ecological sciences. In this work, we give simple, straightforward examples of the bias introduced by the $\log(x + C)$ transform, considering both $C = 1$ and $C < x$. Afterwards, we compare with commonly used alternative transformations, namely, the square root, the cubic root, the Box-Cox and the Inverse Hyperbolic Sine. We start with artificial data demonstrating the problems (Sections 2.1 to 2.3) and proceed to real-world data (Sections 3.1 to 3.2). In section 4 we present the conclusions.

2. Demonstration with artificially-generated data

In demonstrating numerical methods, the first step is always testing its properties with perfectly-controlled artificially-generated data. Following this rationale, we generated several $\{x_1, x_2\}$ bivariate data sets, each for a specific case. Within all of them, x_2 was a strict function of x_1 i.e., $x_2 = f(x_1)$, and no error was associated to any of them. We start by demonstrating the $\log(x)$ transform bias with the linear regression case and then proceed to the case of homogenizing variances for ANOVA designs.

2.1 The linearization problem

The family of exponential functions represents phenomena that grow or decay at a rate proportional to their current value. However, estimating the a and b coefficients in $x_2 = a \cdot x_1^b$ is more convenient using its linearized form, $\log(x_2) = \log(a) + b \cdot \log(x_1)$, since this formulation has a closed-form analytical solution given by the minimum sum of squared residuals. Conversely, if $\log(x_2) = \alpha + \beta \cdot \log(x_1)$, then $x_2 = \exp(\alpha) \cdot x_1^\beta$. A fundamental problem arises when the x_1 or x_2 arrays contain zeros. The commonly proposed solution of replacing x_1 and/or x_2 by their $\log(x + C)$ transform, usually with $C = 1$, introduces severe bias in the estimation of the a and b coefficients. To demonstrate this issue, we generated a bivariate $\{x_1, x_2\}$ dataset. The x_1 variable consisted of a progression from 10^{-4} to 10^0 in increments of $10^{0.1}$, with an additional observation at $x_1 = 0$. The x_2 variable was defined as $x_2 = 0.5 \cdot x_1^2$. Hence, for the case $x_1 = 0$

also $x_2 = 0$. Because no error was added to either variable, the regression of x_2 on x_1 , excluding the (0,0) pair, yielded a perfect fit with coefficients $a = 0.5$ and $b = 2$ (Fig. 1). The difficulty was that the (0,0) pair prevented application of the logarithmic transform to the full $\{x_1, x_2\}$ dataset. To address this issue, several proposed alternative transforms were tested (Table 1), including the $\log(x + C)$ transform with $C = 1$, $C = 10^{-4}$, and $C = 10^{-8}$. The resulting fits were then plotted back into the original $\{x_1, x_2\}$ bivariate space using the corresponding reverse transformations (Table 1).

Table 1
– **Data transforms tested in this work.** Data transforms used for model linearization and variance homogenization; their forward (Forward) and back-transform (Backward) formulae.

Transform name	Forward	Backward
Log	$\dot{x} = \log(x + C)$	$x = e^{\dot{x}} - C$
	$\log(y + C) = a + b \cdot \log(x + C)$	$y = e^{a \cdot (x + C)^b} - C$
Square Root	$\dot{x} = \sqrt{x}$	$x = \dot{x}^2$
	$\sqrt{y} = a + b \cdot \sqrt{x}$	$y = (a + b \cdot \sqrt{x})^2$
Cubic Root	$\dot{x} = \sqrt[3]{x}$	$x = \dot{x}^3$
	$\sqrt[3]{y} = a + b \cdot \sqrt[3]{x}$	$y = (a + b \cdot \sqrt[3]{x})^3$
Box-Cox	$\dot{x} = (x^{\lambda} - 1)/\lambda$	$x = \sqrt[\lambda]{\dot{x}\lambda + 1}$
	$(y^{\lambda} - 1)/\lambda = a + b(x^{\lambda} - 1)/\lambda$	$y = \sqrt[\lambda]{a/\lambda + b(x^{\lambda} - 1) + 1}$
Inverse Hyperbolic Sine	$\dot{x} = \frac{e^x - e^{-x}}{2}$	$x = \log(\dot{x} + \sqrt{\dot{x}^2 + 1})$

The $\log(x + C)$ transform was consistently problematic, yielding highly biased results regardless of the value of C used (Fig. 1). When $C = 1$, because C was much larger than x_1 and x_2 , the regression became far more responsive - or, in the terminology of numerical analysis, more sensitive - to the arbitrary choice of $C = 1$ than to the actual values of x . Notably, for the smallest values of x_1 and x_2 , located in the lower-left corner of the plots, $C = 1$ was 10^4 times larger than the true values of x_1 and x_2 . At first, one might be tempted to think that this could be solved by decreasing C to a value substantially smaller than the

smallest non-zero x . However, by doing so, the $x = 0$ values were transformed into values extremely distant from the remaining x values, which also brought a substantial bias to the estimation of the coefficients (Fig. 1). This effect becomes more pronounced as the number of zeros in the data increases.

The transform alternatives (in Table 1) did not provide any improvement (Fig. 2). Nevertheless, we note that these were applied to an artificial dataset specifically constructed from an exponential function. With real-world data, the exponential functions are not necessarily the best descriptor of the process or dynamics under study. Therefore, the first step in any study should be to test which family of functions best linearizes the data, in which case the $\log(x + C)$ transform should be excluded from consideration. Whenever the data follows the exponential model while containing zeros that are necessary for model calibration, there is no reliable linearization. In such cases, the study must rely on a nonlinear method for parameter estimation.

2.2 The heteroscedasticity problem with non-linear data

In some studies, it is necessary to determine whether categorical groups differ with respect to a common dependent variable, x . This is the classical ANOVA design, for which a key assumption is the homogeneity of within-group variances (homoscedasticity). Even when non-parametric alternatives are used (e.g., permutation tests), this constrain still holds. In practice, however, data often exhibit heterogeneous variances (heteroscedasticity) - most commonly as variances proportional to the group means, thereby undermining the direct application of ANOVA-designed tests.

Here, we first addressed the case of non-linear x data exhibiting heteroscedasticity among categorical classes. To this end, we generated a dataset comprising variable x divided into five groups (A, B, C, D, and E), each containing 30 sampling units and having group means (μ) of 0.0001, 0.001, 0.01, 0.1, and 1, respectively. All groups followed a log-normal distribution with a standard deviation of 0.4 for $\log(x)$. As a result, the groups were completely distinct, with almost no overlap among them (Fig. 3a). In this case, the $\log_{10}x$ transform successfully homogenized the variances among groups (Fig. 3a). This is an effective solution when the data contain no zeros, its main limitation being the bias associated to back-transforming coefficients estimated by numerical methods. The $\log(x + 1)$ transform, on the other hand, was proposed to address the problem of applying a logarithmic transformation to data containing zeros. However, this transformation has the drawbacks illustrated by its application to the artificially generated dataset (Fig. 3b): The addition of the constant $C = 1$ introduced two biases, namely, (i) it increased, rather than reduced, heteroscedasticity, and (ii) it compressed the means (μ) of groups for which $x < C$ toward $\mu \approx \log(C)$. These biases become more pronounced as x decreases relative to C . In the current example, for groups A through E, the constant $C = 1$ was 10,000 $\times$, 1,000 $\times$, 100 $\times$, 10 $\times$, and 1 $\times$ the magnitude of x , respectively.

The question then arose as to whether these biases could be mitigated by choosing a sufficiently small value of C . To investigate this, the $\log(x + C)$ transform was applied to the artificial x data using $C = 1$

(Fig. 4a), $C = 10^{-4}$ (Fig. 4b), and $C = 10^{-8}$ (Fig. 4c). Indeed, the previously described biases tended to disappear as C decreased to infinitesimal. Furthermore, the distribution of (x) within each group became increasingly symmetric (Gaussian), satisfying another fundamental assumption of ANOVA. However, this was only possible because the data contained no zeros. But in such cases, there is no justification for adding a constant C instead of simply applying the $\log(x)$ transform. To evaluate the performance of these transformations in the presence of zeros, ten observations with zero values were added to each group of the original dataset. Consequently, each group now contained 40 observations, 10 of which equaled zero. When the $\log(x + C)$ transform was applied to this modified dataset using $C = 1$ (Fig. 4d), $C = 10^{-4}$ (Fig. 4e), and $C = 10^{-8}$ (Fig. 4f), the resulting biases were dramatic: the groups became markedly heteroscedastic and their within-group distributions became outstandingly skewed. In conclusion, the $\log(x + C)$ transform brings large bias to ANOVA-type of tests, whatever the value of C , and should be avoided.

The question then became whether the alternative transformations were able to homogenize variances, even in the dataset containing zero counts. The short answer is no. To streamline this presentation avoiding an overwhelming amount of figure panels, we do not show the results for all transformations applied to both datasets. Instead, we focus on the Box–Cox transformation, as it provided the most informative results. After testing and tuning its λ parameter to 0.01, the Box–Cox transformation applied to the dataset with zeros successfully homogenized variances within groups (Fig. 5a), despite the dataset having been generated from an exponential function. This finding strongly supports the Box–Cox transformation as a versatile approach for addressing heteroscedasticity. However, when applied to the same dataset as above, including zero counts in 25% of the observations, its performance was poor as (i) it failed to homogenize variances while (ii) the within-group variances became severely skewed, violating the assumption of symmetrical Gaussian-type distributions (Fig. 5b). Finally, in the original artificial dataset, to simulate the larger likelihood of zero counts appearing associated to the group showing smaller x , we tested replacing 3 of the counts from group A by zero. Consequently, on the account of group A, this dataset became highly heteroscedastic (Fig. 5c).

In conclusion, none of the transformations tested was able to homogenize variances in non-linear datasets containing observations valued zero. This does not mean that any transformation is condemned to fail. For each dataset should be tested its specific case, which may not necessarily follow an exponential model. Hence, alternative transformations may work well. Nevertheless, it should be known in advance that zeros in the data likely undermine any transformation. In our tests, the Box-Cox transform tuning its λ was the most successful approach.

2.3 The heteroscedasticity problem with linear data

The data transformations were tested on ANOVA-designed linear data with variances proportional to group means. To this end, we generated a dataset comprising variable x divided into five groups (A, B, C, D, and E), each containing 50 sampling units and having group means (μ) of 1, 2, 3, 4, and 5, respectively. All groups followed a normal distribution, $N(\mu, 0.1 \cdot \mu)$ i.e., with a mean μ and a variance of $0.1 \cdot \mu$. As a

result, the groups were clearly distinct, albeit some overlap among them (Fig. 6a). In order to objectively compare the transformations' ability to homogenize variances, Bartlett's, Levene's, and Brown–Forsythe's tests were applied (Table 2). In these tests, the null hypothesis is homoscedasticity. Therefore, a p-value < 0.05 leads to rejection of the null hypothesis in favour of the alternative hypothesis of heteroscedasticity, which was clearly the case for the original x data.

Table 2
– **Tests on the homogeneity of variances in the data without zeros.** Bartlett's, Levene's, and Brown–Forsythe's tests applied to the original dataset and its transforms, namely the $\log(x)$, $\log(x + 1)$ and Box-Cox. Significant values, representing heteroscedasticity, typed in **bold**.

		original	$\log(x)$	$\log(x + 1)$	Box-Cox
Bartlet	value	118.7	0.91	15.5	0.939
	p	0	0.923	0.004	0.919
Levene	value	18.9	0.08	3.13	0.089
	p	0	0.989	0.016	0.986
Brown – Forsythe	value	18.5	0.08	2.95	0.091
	p	0	0.989	0.021	0.985

The $\log(x)$, $\log(x + 1)$ and Box–Cox (Fig. 6b-d, respectively) transformations all reduced the heteroscedasticity of the original x data containing no zeros (Table 2). Nevertheless, the $\log(x + 1)$ transformation remained heteroscedastic. By contrast, the $\log(x)$ and Box–Cox, transformations achieved homoscedasticity.

To assess and compare the performance of these transformations when applied to data containing zeros, five observations in group A were replaced with zero (Fig. 6e). In this case, all transformations kept, and could even worsen, the heteroscedasticity already present in the untransformed x data (Fig. 6g-h; Table 3). In the Box–Cox transformation λ was changed to 0.5, as with $\lambda = 0.01$ it yielded a highly biased result.

Table 3

– **Tests on the homogeneity of variances in the data with zeros in group A.** Bartlett's, Levene's, and Brown–Forsythe's tests applied to the original dataset and its transforms, namely the $\log(x)$, $\log(x + 1)$ and Box-Cox. Significant values, representing heteroscedasticity, typed in **bold**.

		original	$\log(x + 1)$	Box-Cox
Bartlett	value	47.4	121	169
	p	0	0	0
Levene	value	8.57	6.67	8.84
	p	0	0	0
Brown – Forsythe	value	8.62	2.57	2.71
	p	0	0.039	0.031

In conclusion, the $\log(x + 1)$ transform is never a valid choice to homogenize the data. On the other hand, alternative transformations as the $\log(x)$ and Box-Cox may or not be valid, depending on the existence of zeros in the data, and on how much these outlay the remaining data.

3. Application to real-world case-studies

3.1 Seagrass *Posidonia oceanica* Flowering Intensity vs Marine Heat Wave

This dataset was taken from a pan-Mediterranean study by Astruch et al. (2026) that investigated the summer 2022 marine heatwave triggering a mass flowering of the seagrass *Posidonia oceanica* (Linnaeus) Delile, 1813 of unprecedented scale. Increased reproductive effort by seagrasses is commonly considered an indicator of their stress. In Astruch et al. (2026), as in most studies on the subject, reproductive effort was quantified using Flowering Intensity (FI), defined as the average number of inflorescences per shoot. The heatwave was represented by the cumulative sum of daily sea surface temperature anomalies (MHW_{icum} , °C) during the time the marine heatwave lasted. Because the heatwave lasted all summer (June, July and August), some days showing anomalies up to 4.6°C, the MHW_{icum} attained values as high as 300°C. The increase in FI with increasing MHW_{icum} was non-linear and seemingly exponential (Fig. 7). To estimate the corresponding curve in the original FI– MHW_{icum} space, linear regressions were fitted to the data in the FI-transformed – MHW_{icum} bivariate space (Fig. 7). The fitted lines were then back-transformed into the original space, yielding curved relationships. An important consideration in the fitting procedure is that the use of seagrass reproductive effort as a proxy for stress is complicated by collateral factors. For example, seagrasses rely on energetic reserves to support the metabolic costs of reproduction, and these reserves require considerable time to

accumulate. Consequently, individuals that flowered during the previous summer are less likely to flower again in the current season. Such factors introduce substantial variability that obscures the underlying FI–MHW_{icum} relationship. One approach to reduce the influence of this noise is to apply quantile regression (Zhang et al. 2005; Geraci 2019) to the extreme-most observations in the data (Vieira et al. 2018b, 2021, 2022b, 2024, 2026; Creed et al. 2019). In the present case, quantile regressions fitted to the highest FI values identified the maximum FI response attainable at a given MHW_{icum} level when nothing else interferes. Observations falling substantially below this upper boundary were interpreted as reflecting the influence of collateral factors rather than the direct effect of marine heatwave intensity.

For the application of quantile regressions, the FI was transformed using $\log(x)$, $\log(x + 1)$, square root, Box-Cox, and inverse hyperbolic sine (IHS) transformations (Fig. 7). The Box-Cox transform was tuned with $\lambda = 0.02$. The $\log(x + 1)$ and IHS transformations led to poor quantile regressions, as they were unable to capture the nonlinear dynamics of the original FI–MHW_{icum} relationship. In contrast, the $\log(x)$ and Box-Cox transforms produced good quantile regressions, capturing this relationship quite well. The striking difference between the quantile regressions applied to the $\log(x)$ and $\log(x + 1)$ transforms highlights the adverse effect of adding a constant (in this case $C = 1$) to the $\log(x)$ transform.

3.2 Estuarine Environmental Quality

This dataset was derived from a study by Vieira et al. (2024) focused on the ecological status of the Tagus Estuary and Sado Estuary, Portugal, which are dominated by the seagrass *Zostera noltei* Hornemann, 1832. Here, we tested for correlations among concentrations of nutrients in the water and sediment; seagrass traits, including above-ground biomass (AB), below-ground biomass (BB), the AB:BB ratio, and Efficiency of Space Occupation (ESO, represented by the d_{grass} index); benthic macrofauna species evenness (Simpson's D and Shannon–Wiener H indices); and the abundance of the marine snail *Peringia ulvae* (Pennant, 1777) (formerly *Hydrobia ulvae*), which is commonly associated with good environmental conditions.

The concentrations of inorganic nitrogen (N) and phosphorus (P), as well as their N:P stoichiometric ratio in water and sediments, are fundamental determinants of nutrient-limited seagrass growth. However, these variables are highly dynamic and depend on nutrient sources and sinks, as well as sediment carbonate chemistry. Orthophosphate (PO_4) in particular is readily adsorbed onto carbonate-rich sediments. Concentrations of Dissolved Inorganic Nitrogen (DIN) and PO_4 were higher in the Tagus than in the Sado, and were generally higher in sediments than in the water column (Fig. 8). Nevertheless, the data clearly align within a narrow N:P band in the PO_4 –DIN plane, which appears to be characteristic of estuaries in this region. The central tendency and the upper and lower boundaries of this band were estimated using Quantile Regression (QR) applied to square-root and $\log(x + 1)$ transformations of the data, followed by back-transformation to the original PO_4 –DIN plane. The square-root transformation (SQRT) preserved the structure of the original data. The estimates of the band's central tendency and boundaries based on the SQRT fitted accurately the original data. In contrast, the $\log(x + 1)$ transformation distorted the data distribution, introducing substantial bias into the estimates of the

band's central tendency and boundaries. This is noticeable in their curved shapes, and particularly in their inflection downwards at both extremes.

The AB:BB ratio and Efficiency of Space Occupation (ESO), represented by the d_{grass} index, have been shown to be reliable indicators of seagrass health. Higher AB:BB ratios and higher ESO values (corresponding to lower d_{grass} values) indicate better seagrass condition. Nurtured by higher nitrogen (N) and phosphorus (P) concentrations, *Zostera noltei* meadows in the Tagus were healthier than those in the Sado (Fig. 8). Further supporting their usefulness as indicators of seagrass health, the AB:BB ratio and d_{grass} index were strongly associated, with meadows from both the Tagus and Sado estuaries distributed along a well-defined band in the AB:BB– d_{grass} plane. The upper and lower boundaries of this band were accurately estimated using the $\log(x)$ transformation. In contrast, estimates obtained from the $\log(x + 1)$ transformation were severely biased. This is particularly noticeable from the inflection downwards at the lower AB:BB and higher d_{grass} extreme.

AB:BB > 1 indicates growing seagrass meadows, whereas AB:BB < 1 indicates declining seagrass meadows (Vieira et al., 2024). However, the direct use of the AB:BB ratio is inappropriate in ANOVA designs. This is because AB:BB is a ratio, and metrics derived from multiplicative processes or geometric progressions are not well suited to the additive/arithmetic assumptions underlying ANOVA (Vieira et al., 2018a). In such cases, the variances from groups with observations valued < 1 get compressed whereas the variances from groups with observations valued > 1 get inflated. Consider the balanced case, where 1:1 = 1. Now consider the unbalanced case 2:1 = 2, which differs from the balanced case by 2 – 1 = 1. In contrast, the unbalanced case 1:2 = 0.5 differs from the balanced case by 0.5 – 1 = – 0.5. Thus, equivalent proportional deviations above and below unity are not represented symmetrically on the original ratio scale. The solution is to linearize the ratio using a logarithmic transformation, because $|\log(x)| = |\log(1/x)|$. For example, $\log_{10}(10) = 1$ and $\log_{10}(1/10) = -1$. The comparison between the Tagus and Sado estuaries provides a clear example of this issue. Because AB:BB was consistently > 1 in the Tagus but predominantly < 1 in the Sado, their variances were highly heteroscedastic. Applying a $\log(x)$ transformation resolved this problem, resulting in homoscedastic variances (Fig. 8). In contrast, the $\log(x + 1)$ transformation introduced a systematic bias because the arbitrary addition of 1 had a much greater effect on the smaller AB:BB values observed in the Sado than on the larger values observed in the Tagus (Fig. 8).

P. ulvae (formerly *Hydrobia ulvae*) is a marine gastropod associated with seagrass meadows and is commonly linked to brackish waters and healthy environmental conditions (Cardoso et al., 2005; Araújo et al., 2012; Ondiviela et al., 2018; Vieira et al., 2024). Reflecting the much greater freshwater input and healthier environmental conditions in the Tagus estuary, *P. ulvae* was by far the dominant species. This resulted in a significant increase in abundance but a decrease in species evenness (Fig. 9). In contrast, *P. ulvae* was almost absent from the Sado estuary, leading to a significant decrease in abundance but an increase in species evenness. Overall, *P. ulvae* abundance was strongly negatively correlated with species evenness. To capture this non-linear relationship, both the Simpson's D and Shannon–Wiener H

indices of species evenness were fitted as functions of *P. ulvae* abundance (Fig. 9). Variables were first transformed using square-root, Box–Cox, or $\log(x + 1)$ transformations, after which a second-degree polynomial was fitted. The fitted models were then back-transformed to the original scale. The square-root and Box–Cox transformations produced very similar results and, when back-transformed, accurately captured the relationship between *P. ulvae* abundance and species evenness. In contrast, models fitted to the $\log(x + 1)$ -transformed data exhibited substantial bias after back-transformation; which is particularly evident from the way this fit, represented by the red line in Fig. 9, departs the central tendency towards extremely high Simpson D and extremely low Shannon-Weiner H. This bias arose because the addition of 1 affected different portions of the variable ranges unevenly, even causing the concavity of the fitted curve in the $\log(x + 1)$ transformed scale to reverse direction. relying on the $\log(x + 1)$ transform in this scenario would lead to a completely flawed ecological interpretation regarding the rate of biodiversity change. This critical failure perfectly exemplifies the study's overarching conclusion: the $\log(x + 1)$ transform fundamentally compromises analytical integrity and should be entirely avoided in environmental modelling.

4. Discussion

The log-transform conflicts with the existence of zeros in the data because $\log(0) = -\infty$. The first proposed workaround was to apply the $\log(x + 1)$ transformation instead, this having become the most commonly used solution. However, criticisms emerged thereafter. One relates with the bias introduced by back-transforming from log-space (Wilson et al., 1990; Newman 1993; Zeng & Tang, 2011; Vieira et al., 2022a). This applies to any logarithmic transformation, including the $\log(x)$ and the $\log(x + C)$. Another debate followed on the adequate value of C (Yamamura 1999; Motakis et al., 2006). However, our results show that this is not the way to go as the $\log(x + C)$ transformation, whatever the C value and including the commonly used $\log(x + 1)$ transformation, introduces systematic bias that can profoundly affect data analysis. When the constant C is large relative to the values of x , the analysis becomes predominantly responsive to C . On the other hand, very small values of C replacing the zero counts lead these to significantly depart the remaining data towards infinitesimal values. On both cases, the analysis loses track of the true signal in the data. This bias associated with the $\log(x + C)$ transformation can affect regression analyses, ANOVA designs, and attempts to stabilize or homogenize variances. Therefore, the $\log(x + C)$ transformation, whatever the C value arbitrarily chosen, should be avoided whenever possible and used only as a last resort, following appropriate evaluation and testing.

Despite the inadequacy of the $\log(x + 1)$ transform, straightforward alternatives are often lacking, particularly when the data follow an exponential distribution or when the relationship between x and y is exponential. In such cases, the Box–Cox was found the most flexible alternative transformation, as it can be tailored to the degree of nonlinearity present in the data through adjustment of its λ parameter. Still, for the particular case of exponential data, modelling alternatives such as GLM (Nelder & Wedderburn 1972)) are likely better solutions than performing any data transformation (St-Pierre et al., 2018).

Declarations

Data Availability Statement

The artificially-generated data was created following the protocol presented in the body of the text. The *Zostera noltii* dataset analysed in this study was originally generated and described by Vieira et al. (2024) in Communications Earth and Environment (<https://www.nature.com/articles/s43247-024-01758-0>).

The *Posidonia oceanica* dataset analysed in this study was originally generated and described by Astruch et al. (2026) in Communications Earth and Environment (<https://www.nature.com/articles/s43247-026-03631-8>). This study represents a secondary, independent re-analysis of those published data for demonstrating the biases introduced by the $\log(x+C)$ transformation, and no new biological samples were collected.

Code Availability statement

All analyses were made in Matlab ®. This software is provided online by Mathworks.

Competing Interests statement

The authors declare no financial nor non-financial competing interests.

Authors contributions

Vasco Vieira: funding, conceptualization, field work, laboratory work, data curation, data analysis, interpretation, writing the manuscript, reviewing the manuscript.

Cátia Bartilotti: funding, reviewing the manuscript.

Jorge Lobo-Arteaga: funding, reviewing the manuscript.

Rafael Santos: reviewing the manuscript.

David Leitão-Silva: reviewing the manuscript.

Arthur Veronez: reviewing the manuscript.

Joana M. Neves: reviewing the manuscript.

Ana Brito: reviewing the manuscript.

Rui Cereja: reviewing the manuscript.

Diogo S. Paulo: funding, reviewing the manuscript.

Francisco Leitão: funding, reviewing the manuscript.

Funding statement

This work received Portuguese national funds from FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portugal) - through contracts UID/04326/2025 (DOI: 10.54499/UID/04326/2025), UID/PRR/04326/2025 (DOI: 10.54499/UID/PRR/04326/2025) and LA/P/0101/2020 (DOI:10.54499/LA/P/0101/2020), and from the operational programmes CRESC Algarve 2020 and COMPETE 2020 through contract EMBRC.PT ALG-01-0145-FEDER-022121.

This work received Portuguese national funds from FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portugal) - through the projects UID/4292/2025 (DOI: 10.54499/UID/04292/2025) and UID/PRR/4292/2025 (DOI: 10.54499/UID/PRR/04292/2025) granted to MARE - Marine and Environmental Sciences Centre, and the project LA/P/0069/2020 (DOI: 10.54499/LA/P/0069/2020) granted to the Associate Laboratory ARNET - Aquatic Research Network.

Joana M. Neves was supported by a PhD scholarship granted by FCT (UI/BD/150954/2021).

Arthur Veronez was supported by an individual fellowship granted by FCT(PRT/BD/153040/2021).

David L.-Silva was supported by an individual fellowship granted by FCT (2023.03020.BDANA)

References

1. Araújo, C. V. M., Blasco, J. & Moreno-Garrido, I. Measuring the avoidance behaviour shown by the snail *Hydrobia ulvae* exposed to sediment with a known contamination gradient. *Ecotoxicology* **21**, 750–758. <https://doi.org/10.1007/s10646-011-0835-6> (2012).
2. Astruch, P. et al. Mass flowering of the seagrass *Posidonia oceanica* after 2022 record-breaking marine heatwaves, a Pan-Mediterranean study. *Commun. Earth Environ.* <https://doi.org/10.1038/s43247-026-03631-8> (2026).
3. Bellégo, C., Benatia, D. & Pape, L. Dealing with logs and zeros in regression models. arXiv preprint arXiv:2203.11820. (2022).
4. Boeshaghi, A. S. & Pachter, L. Normalization of single-cell RNA-seq counts by $\log(x + 1)$ or $\log(1 + x)$. *Bioinformatics* **9** (15), 2223–2224. [10.1093/bioinformatics/btab085](https://doi.org/10.1093/bioinformatics/btab085) (2021).
5. Cardoso, P. G., Brandão, A., Pardal, M. A., Raffaelli, D. & Marques, J. C. Resilience of *Hydrobia ulvae* populations to anthropogenic and natural disturbances. *Mar. Ecol. Prog Ser.* **289**, 191–199 (2005).
6. Creed, J., Vieira, V. M. N. C. S., Norton, T. A. & Caetano, D. A metaanalysis shows that seaweeds surpass plants, setting life-onEarth's limit for biomass packing. *BMC Ecol.* **19**, 1–11 (2019).
7. Legendre, P. & Gallagher, E. Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**, 271–280. [10.1007/s004420100716](https://doi.org/10.1007/s004420100716) (2001).
8. Lun, A. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. (2018). [10.1101/404962](https://doi.org/10.1101/404962)

9. Martín-Fernández, J. A., Barceló-Vidal, C. & Pawlowsky-Glahn, V. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Math. Geol.* **35**, 253–278. <https://doi.org/10.1023/A:1023866030544> (2003).
10. Mascaro, J., Litton, C. M., Hughes, R. F., Uowolo, A. & Schnitzer, S. S.A. Minimizing bias in biomass allometry: Model selection and log transformation of data. *Biotropica*. (2011). <https://doi.org/10.1111/j.1744-7429.2011.00798.x>
11. Motakis, E. S., Nason, G. P., Fryzlewicz, P. & Rutter, G. A. Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach, *Bioinformatics*, **22**(20), 2547–2553, <https://doi.org/10.1093/bioinformatics/btl412> (2006).
12. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *J. Royal Stat. Soc. Ser. A.* **135** (3), 370–384 (1972).
13. Newman, M. C. Regression analysis of log-transformed data: Statistical bias and its correction. *Environ. Toxicol. Chem.* **12** (6), 1129–1133. <https://doi.org/10.1002/etc.5620120618> (1993).
14. Ondiviela, B., Fernández, L., Puente, A., García-Castrillo, G. & Juanes, J. A. Characterization of a resilient seagrass meadow during a decline period. *Sci. Mar.* **82**, 67–75 (2018).
15. St-Pierre, A. P., Shikon, V. & Schneider, D. C. Count data in biology—Data transformation or model reformation? *Ecol. Evol.* **8**, 3077–3085. <https://doi.org/10.1002/ece3.3807> (2018).
16. Vieira, V. M. N. C. S., Engelen, A. H., Huanel, O. & Guillemín, M. L. Differentiation of haploid and diploid fertilities in *Gracilaria chilensis* affect ploidy ratio. *BMC Evol. Biol.* **18**, 183. [10.1186/s12862-018-1287-x](https://doi.org/10.1186/s12862-018-1287-x) (2018a).
17. Vieira, V. M. N. C. S., Lopes, I. E. & Creed, J. C. The biomass–density relationship in seagrasses and its use as an ecological indicator. *BMC Ecol.* **18**, 44 (2018b).
18. Vieira, V. M. N. C. S., Engelen, A. H., Huanel, O. & Guillemín, M. L. Differential frond growth in the isomorphic haploid-diploid red seaweed *Agarophyton chilense* by long-term in situ monitoring. *J. Phycol.* **57**, 592–605 (2021).
19. Vieira, V. M. N. C. S., Engelen, A. H., Huanel, O. R. & Guillemín, M. L. An Individual-Based Model of the Red Alga *Agarophyton chilense* Unravels the Complex Demography of Its Intertidal Stands. *Front. Ecol. Evol.* **10**, 797350. [10.3389/fevo.2022.797350](https://doi.org/10.3389/fevo.2022.797350) (2022a).
20. Vieira, V. M. N. C. S. et al. Seagrasses benefit from mild anthropogenic nutrient additions. *Front. Mar. Sci.* **9**, 960249 (2022b).
21. Vieira, V. M. N. C. S. et al. Seagrass space occupation efficiency is key for their role as ecosystem engineers and ecological indicators. *Commun. Earth Environ.* **5**, 592 (2024).
22. Vieira, V. M. N. C. S. et al. Interplay among self-thinning, efficiency of space occupation and biodiversity in terrestrial plant communities. *Commun. Earth Environ.* **7**, 431. <https://doi.org/10.1038/s43247-026-03583-z> (2026).
23. Wilson, B. G., Adams, B. J. & Karney, B. W. Bias in log-transformed frequency distributions. *J. Hydrol.* **118** (1–4), 19–37. [https://doi.org/10.1016/0022-1694\(90\)90248-V](https://doi.org/10.1016/0022-1694(90)90248-V) (1990).

24. Yamamura, K. Transformation using $(x + 0.5)$ to stabilize the variance of populations. *Popul. Ecol.* **41**, 229–234. <https://doi.org/10.1007/s101440050026> (1999).
25. Zhang, L., Bi, H., Gove, J. H. & Heath, L. S. A comparison of alternative methods for estimating the self-thinning boundary line. *Can. J. Res.* **35**, 1507–1514 (2005).
26. Geraci, M. Modelling and estimation of nonlinear quantile regression with clustered data. *Comput. Stat. Data Anal.* **136**, 30–46 (2019).
27. Zeng, W. & Tang, S. Bias correction in logarithmic regression and comparison with weighted regression for nonlinear models. *Nat. Prec.* **24**, 137–143. [10.1038/npre.2011.6708.1](https://doi.org/10.1038/npre.2011.6708.1) (2011).

Figures

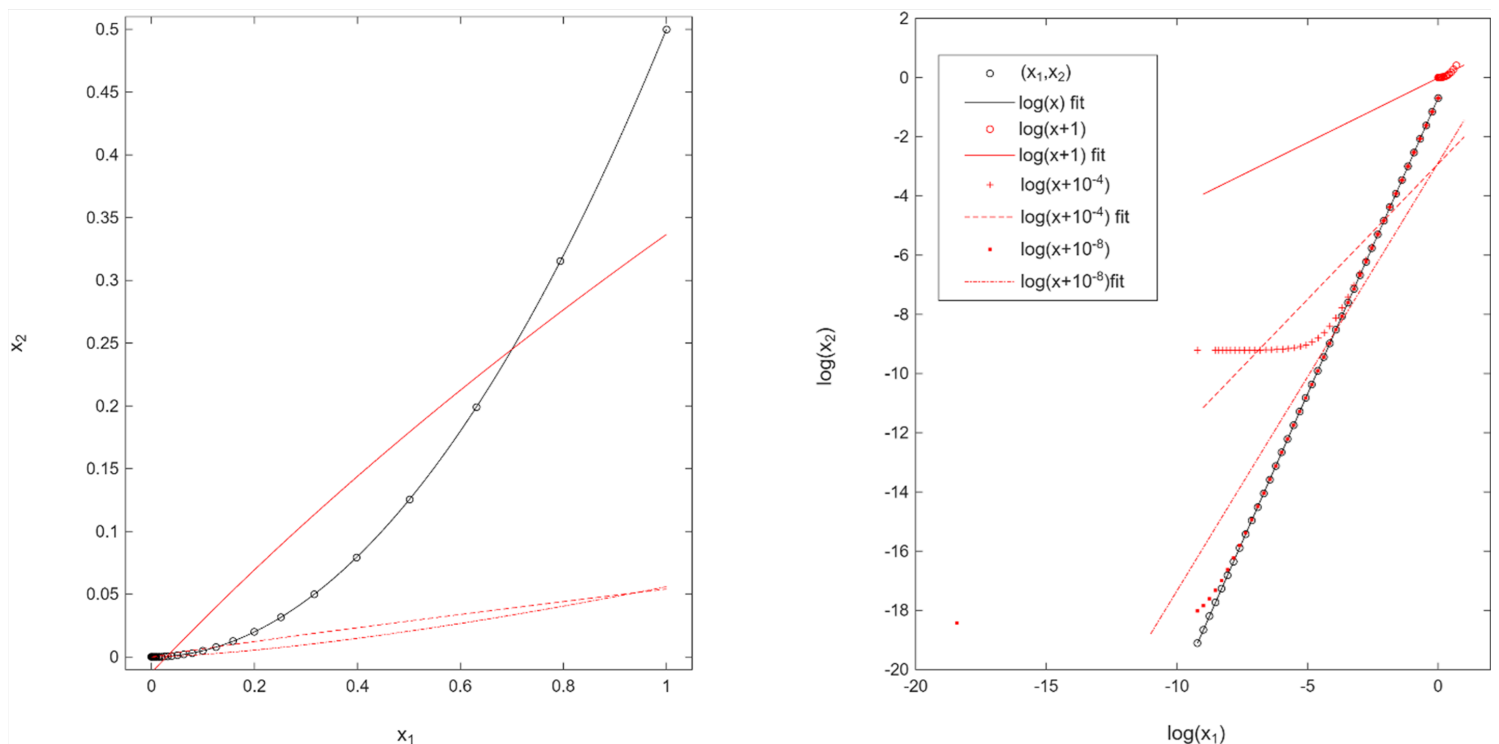


Figure 1

Application of the Log(x+C) transform. The $\log(x+C)$ transform was applied with $C=1$, $C=10^{-4}$ and $C=10^{-8}$, to the $\{x_1, x_2\}$ artificial data set generated from an exponential function. Results plotted in linear axis (left) and logarithmic axis (right). Left panel has the axes in linear scales whereas the right panel has the axes in logarithmic scales.

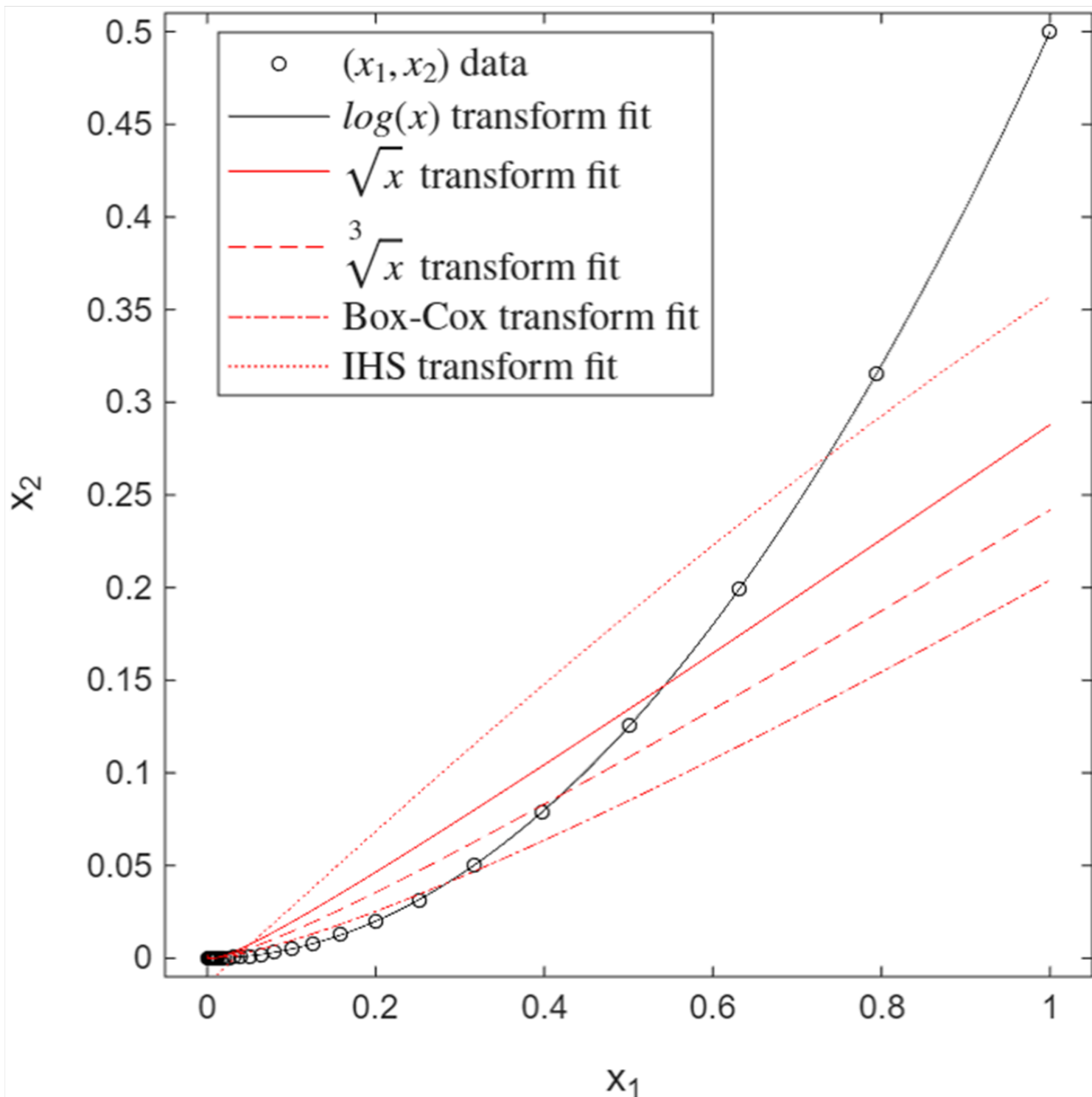


Figure 2

Application of transforms alternative to the $\log(x+C)$. The Square Root, Cubic Root, Boc-Cox and Inverse Hyperbolic Sine (IHS) transforms were applied to the $\{x_1, x_2\}$ artificial data set constructed from an exponential function.

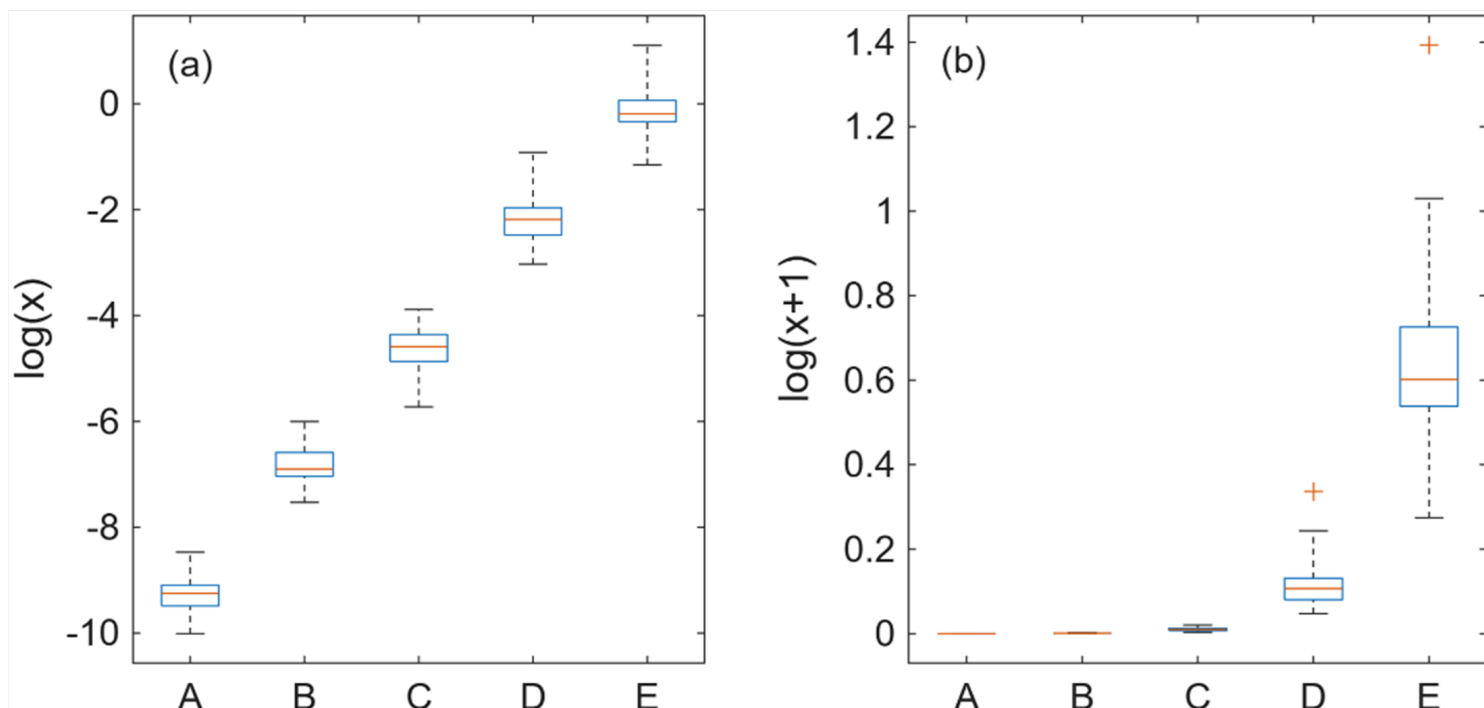


Figure 3

Log transform of the artificially generated data for ANOVA designs. Original variable x was artificially generated with no zeros and depended on groups A, B, C, D and E. Then, it was subject to the $\log_{10}(x)$ transformation (a) and the $\log_{10}(x+1)$ transformation (b). The + represent outliers.

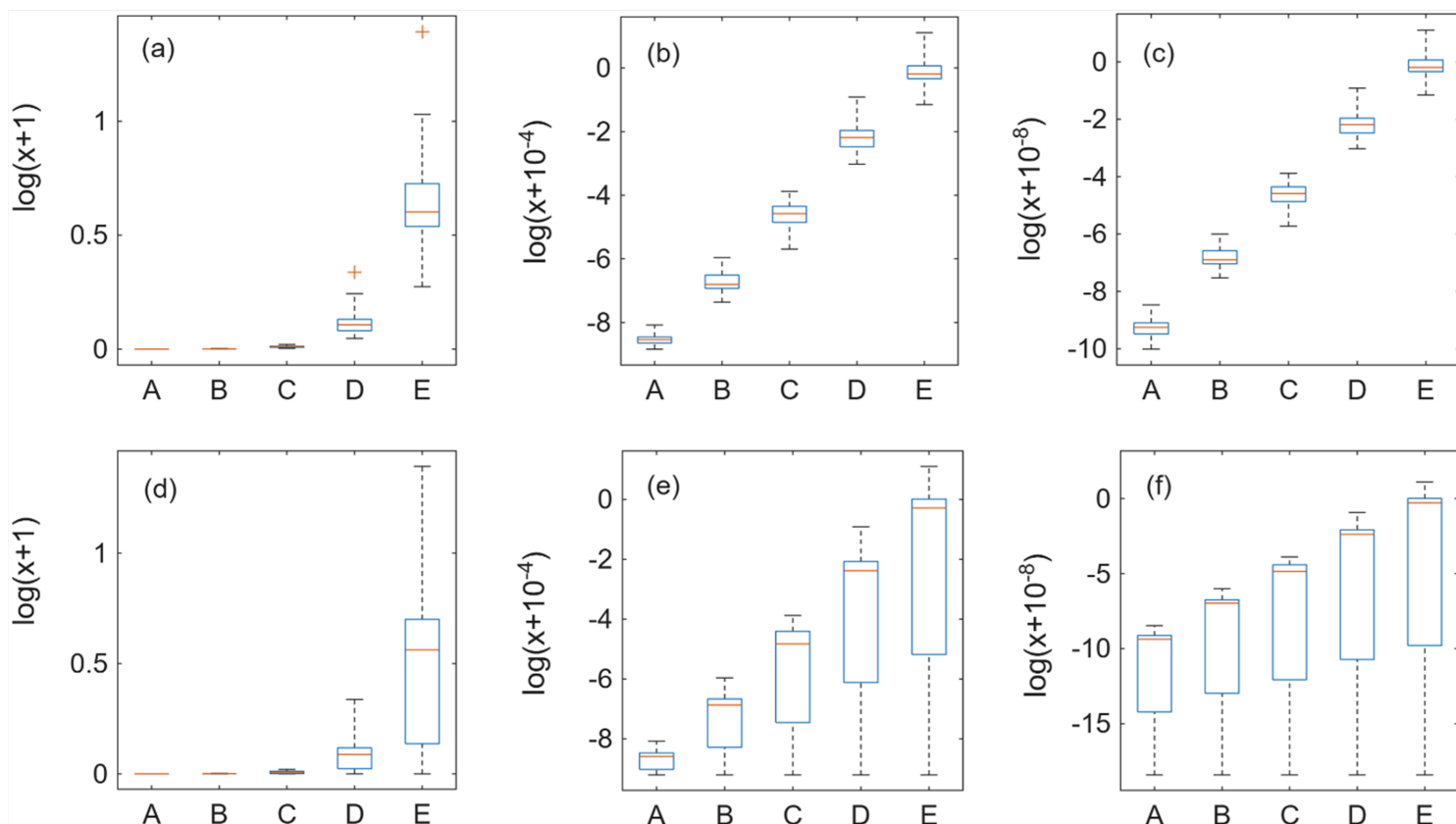


Figure 4

Effects of C level on the $\log(x+C)$ transform for ANOVA designs. Variable x was artificially generated with no zeros and dependent on groups A, B, C, D and E. Original x (in panels a to c) was subject to the $\log_{10}(x+C)$ transform with $C=1$ (a) $C=10^{-4}$ (b) and $C=10^{-8}$ (c). Ten observations valued zero were added to the original x in each group (in panels d to f). This new x was subject to the $\log_{10}(x+C)$ transform with $C=1$ (d) $C=10^{-4}$ (e) and $C=10^{-8}$ (f). The + represent outliers.

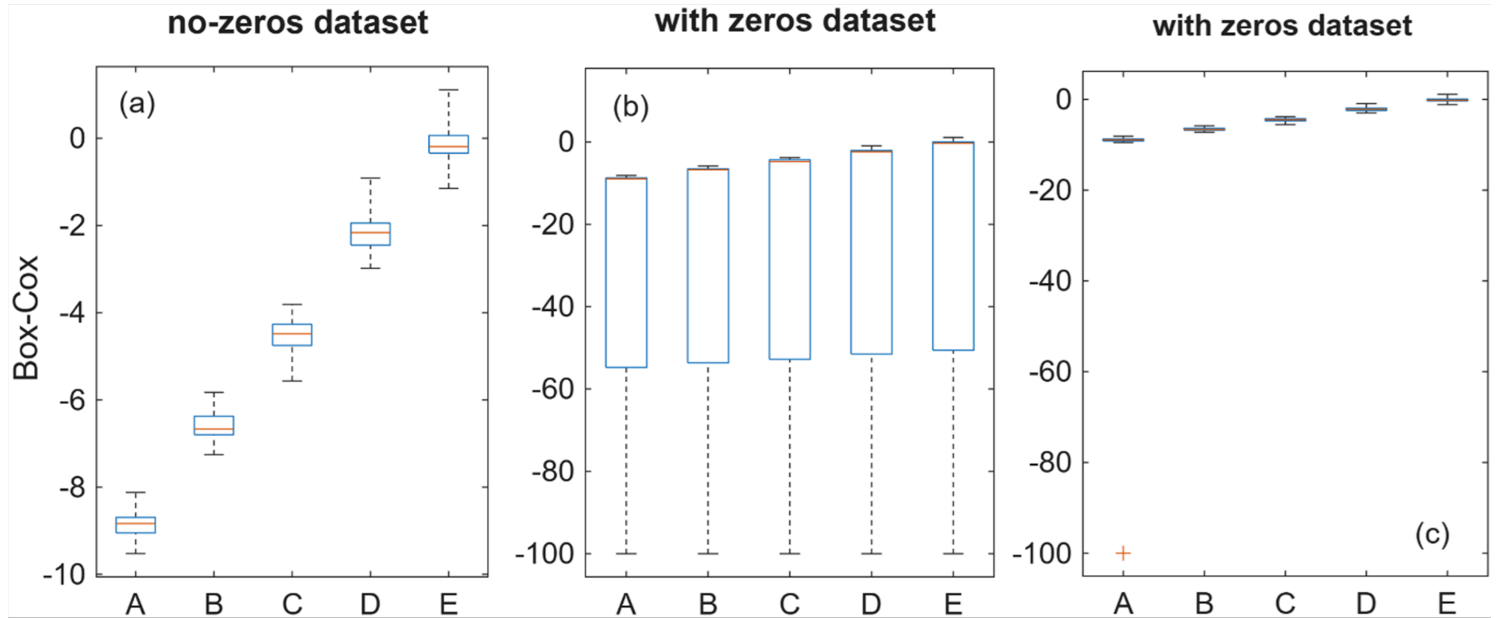


Figure 5

Box-Cox transforms for ANOVA designs. Variable x was artificially generated from an exponential function, with no zeros and dependent on groups A, B, C, D and E (a). Each category had 30 observations. Afterwards, ten observations valued zero were added to the original x in each group (b). Finally, upon the original x data, three observations from group A were changed to zero (c). All datasets were subject to the Box-Cox transform with $\lambda = 0.01$. The + represent the zero-count outliers.

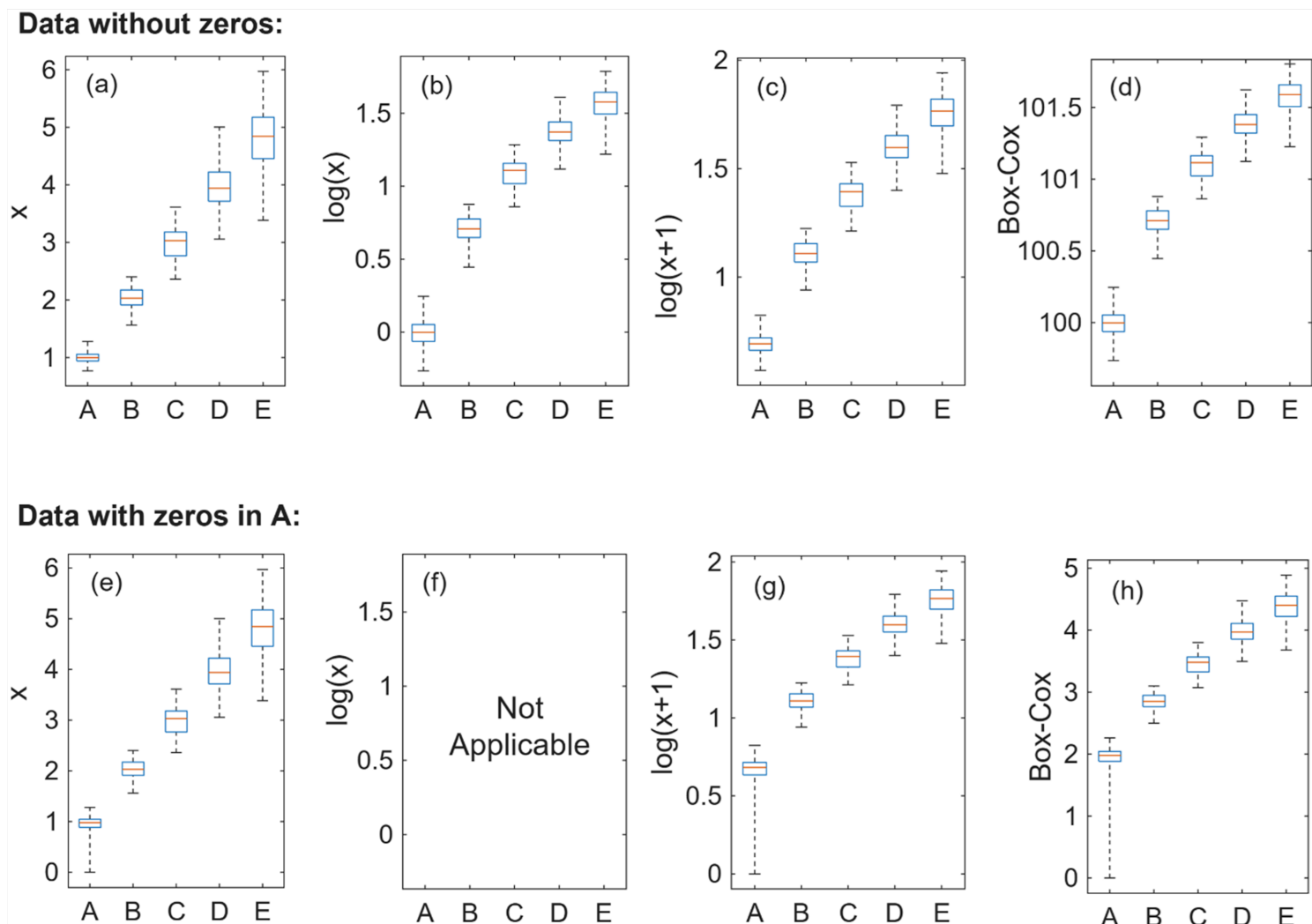


Figure 6

x transforms for ANOVA designs with heteroscedastic linear data. Variable x was artificially generated from a linear function, with no zeros and with mean and variance dependent on groups A, B, C, D and E. Each category had 50 observations (a). Later, five observations from group A were changed to zero (e). Both datasets were subject to the $\log(x)$, $\log(x+1)$ and Box-Cox transform. The $\log(x)$ transform is Not Applicable to data with zeros (f). The Box-Cox transform was applied with $\lambda = 0.01$ (d) and $\lambda = 0.5$ (h). The + represent the zero-count outliers.

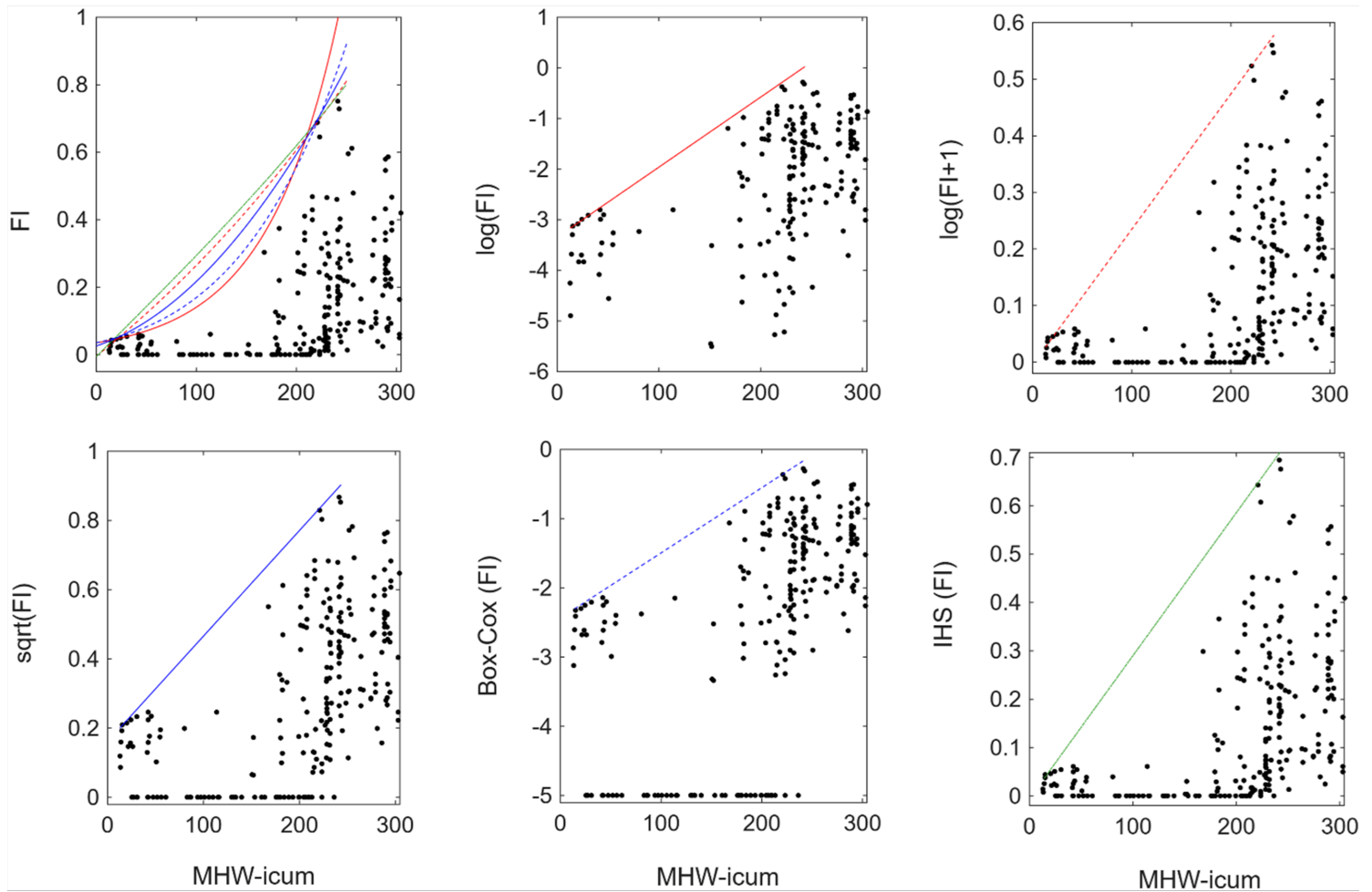


Figure 7

Seagrass *Posidonia oceanica* Flowering Intensity vs Marine Heat Wave. Data from Astruch et al (2026). Flowering Intensity in inflorescence per shoot (FI). Marine Heat Wave cumulative intensity in °C (MHW-icum) is the cumulative sum of the daily Sea Surface Temperature anomaly. FI was subject to the transforms $\log(x)$, $\log(x+1)$, square root (sqrt), and Inverse Hyperbolic Sine (IHS). Straight lines are quantile regressions to the 99% quantile fit to transformed data. Curved lines are the quantile regressions transposed to the original FI – MHW-icum plane.

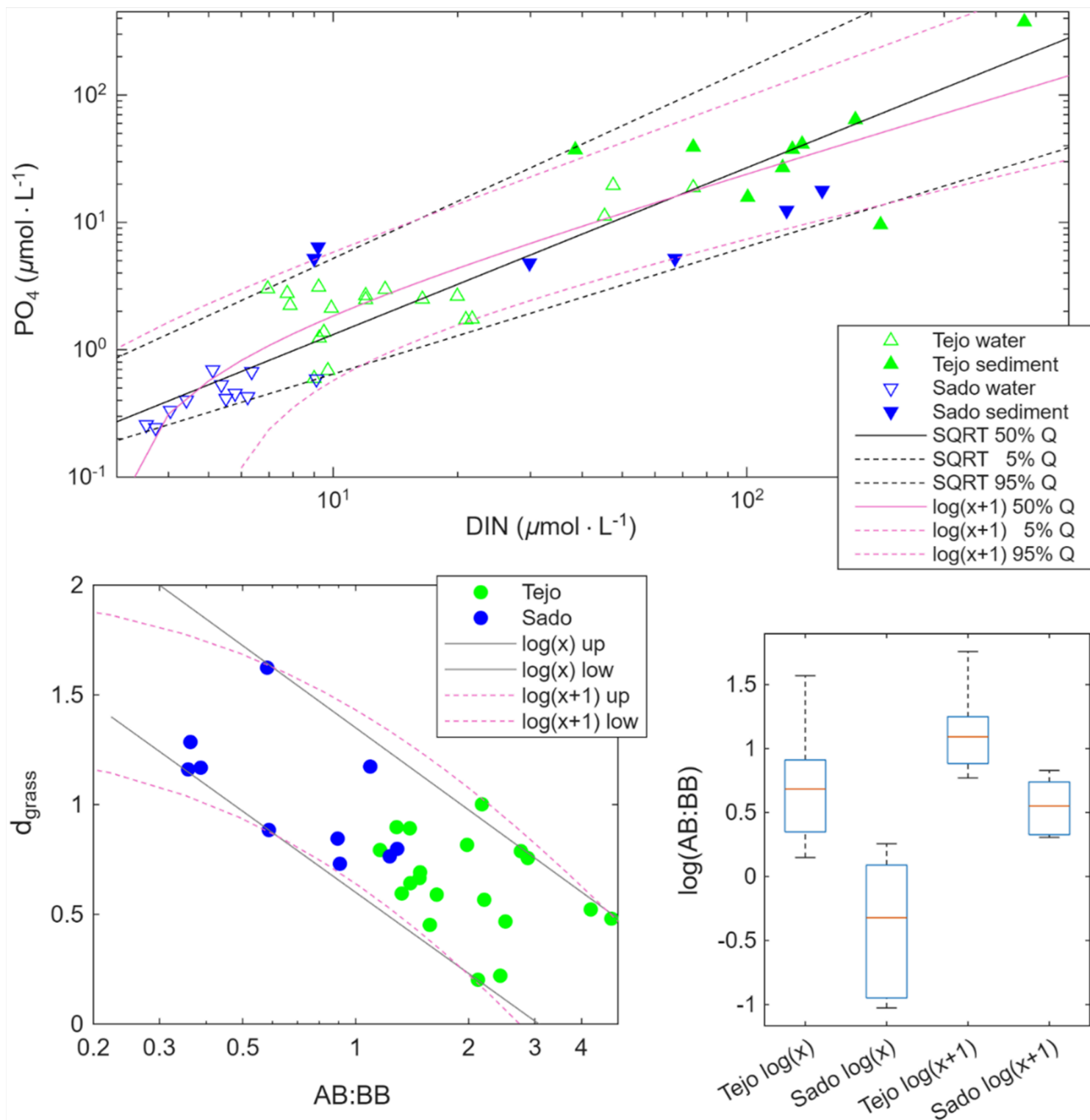


Figure 8

Nutrient and Seagrass properties in the Tagus and Sado estuaries. Concentrations of Dissolved Inorganic Nitrogen (DIN) and orthophosphate (PO_4), in the water and sediment. The seagrass ratio of aboveground biomass (AB) to belowground biomass (BB), and Efficiency of Space Occupation (d_{grass}). Central tendency given by the 50% quantile (50% Q), upper boundary given by the 95% quantile (95% Q),

and lower boundary given by the 5% quantile (5% Q). Quantiles estimated by Quantile Regression applied to the square-root (SQRT) and log(x+1) transforms.

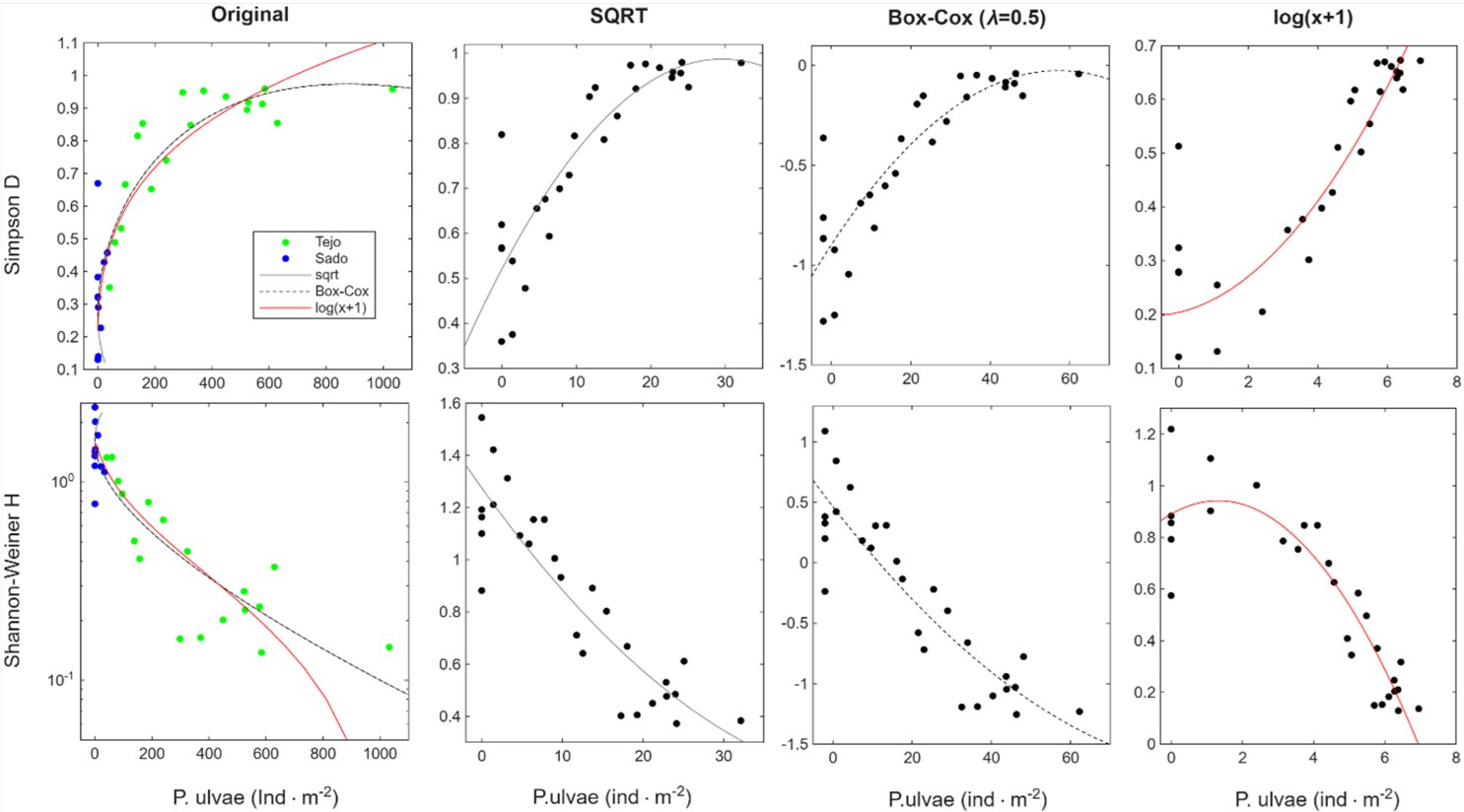


Figure 9

***Peringia ulvae* Abundance vs Biodiversity.** Data from Vieira et al (2024). Biodiversity given by the Simpson D and Shannon-Weiner H indexes for species evenness. Panels have the original data and transformations by the Square Root, Box-Cox and log(x+1). Transformed data are fit to second degree polynomials. Fits are back-transformed to the original data plane.