Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Is Data Sharing Time-Efficient in Ecology? An Empirical Test and Extension of the Break-Even Reuse Model

Adnan Asghar and Frank Daniel

Department of Chemical and Material Engineering, University of alberta, Edmonton, Canada

Abstract

Background. Theoretical models of research data sharing often claim that, beyond a certain level of reuse, openly sharing data becomes time-efficient at the community level. However, empirical tests of these break-even reuse thresholds remain scarce, and key parameters—such as the time required to prepare data for reuse or to integrate external datasets—are rarely quantified for specific disciplines. **Objectives.** This paper has three main objectives: first, to empirically estimate the time costs of data collection, curation, sharing, and reuse in ecology; second, to calibrate and test a break-even reuse model using these discipline-specific parameters; and third, to extend the model with a hierarchical treatment of heterogeneous datasets, distinguishing high-value from low-value data products. **Methods.** We conducted a mixed-methods study combining a survey of 163 practicing ecologists on their data-related time investments and sharing practices with repository analytics from a sample of 320 ecological datasets deposited in major archives such as Dryad, GBIF, and institutional repositories. We used these data to fit hierarchical models of key time parameters and reuse rates. A Monte Carlo simulation framework then propagated parameter uncertainty to obtain posterior distributions of break-even reuse thresholds. We further stratified datasets into high-value and low-value categories and compared the time-efficiency of selective versus universal sharing strategies. **Results.** Across respondents, the median time required to collect a reusable ecological dataset was 30 person-days, while the additional time to prepare and deposit the dataset for reuse was 5 person-days. The median time for reusers to discover, appraise, and integrate an existing dataset was 3 person-days. Under these conditions, the median break-even reuse threshold—the minimum number of reuse events per dataset required to avoid a net time loss at the community level—was 0.3 (95% credible interval: 0.1, 0.8). Repository analytics suggested an expected reuse rate of 0.9 (95% CI: 0.5, 1.6) reuses per dataset within five years, indicating that, in ecology, current sharing practices are already time-efficient on average. High-value datasets exhibited substantially lower break-even thresholds and higher reuse rates, making them strongly time-efficient even under pessimistic assumptions, while low-value datasets hovered around break-even. **Conclusions.** Our results provide empirical support for the claim that data sharing in ecology is, on average, time-efficient at the community level, but also reveal considerable heterogeneity across dataset types. The extended model highlights the potential of selective sharing strategies that prioritize high-value datasets, which deliver large efficiency gains with modest curation investments. We close by discussing implications for repository design, funder mandates, and discipline-specific data policies.

Keywords: data sharing, ecology, time-efficiency, break-even analysis, data reuse, selective sharing, open science, research data management

Corresponding author: Adnan Asghar (adnanasghar14041985@gmail.com)

BCDS 2024, Vol. 5(4), 1-11. DOI: 10.71448/bcds2454-1

1. Introduction

Over the past decade, ecology has experienced rapid growth in the volume and visibility of shared research data [1–3]. Public repositories, data policies at journals and funding agencies, and community-driven standards have collectively lowered some of the technical barriers to open data [4–6]. At the same time, many ecologists remain concerned that preparing, documenting, and publishing data imposes substantial time costs that may not be compensated by downstream benefits [7–10]. These tensions are particularly salient in disciplines like ecology where data collection is labor-intensive, context-dependent, and often conducted by small research teams [1, 6].

Theoretical work has argued that, from a community perspective, data sharing becomes time-efficient once datasets are reused often enough to offset the additional time required to curate and deposit them [11]. In such models, a *break-even reuse threshold* can be identified: if the expected number of reuses per dataset exceeds this threshold, then the total time saved by reusers exceeds the additional time spent by sharers [11]. However, these models have generally been instantiated with illustrative parameter values rather than empirical measurements [11, 12]. As a result, it remains unclear whether current ecological data practices are above or below the break-even point, and for which types of datasets.

This paper addresses that gap through a discipline-specific empirical study in ecology. We combine survey data and repository analytics to estimate the time required to collect a dataset, the extra time required to prepare and share it, and the time required for reusers to discover, understand, and integrate shared data. We then calibrate a simple yet expressive break-even model and extend it in two important ways: first, by propagating parameter uncertainty through Bayesian estimation and Monte Carlo simulation, and second, by explicitly modeling heterogeneity between high-value and low-value datasets.

Our contributions are threefold. First, we provide the first discipline-specific empirical estimates of key time parameters relevant for community-level data sharing efficiency in ecology. Second, we test whether observed reuse rates in major ecological repositories exceed the empirically derived break-even threshold, under a range of plausible assumptions. Third, we extend the standard model with a hierarchical representation of heterogeneous datasets and evaluate the time-efficiency of selective sharing policies that prioritize high-value data.

1.1. Conceptual Framework

We adopt a community-level perspective on time-efficiency. Consider a research domain in which scientists produce datasets and conduct analyses that rely on related data [12, 13]. In a counterfactual world without data sharing, each research team must collect its own data. In a world with data sharing, some teams instead reuse datasets collected by others. Data sharing is time-efficient if, aggregated over all projects, the total person-time spent on data collection, curation, sharing, and reuse is lower in the sharing world than in the non-sharing world [11, 14]. We focus exclusively on time as a resource, leaving aside costs like money or storage, and on scholarly research use, while recognizing that data also support teaching, policy, and other forms of impact [1, 5, 14].

1.2. Break-even reuse model

We define the following parameters for a given dataset, following the notation introduced by [11]. and related work on the economics of data reuse [12, 14]. T_c represents the person-days required to

collect the dataset in the field or laboratory; T_s denotes the additional person-days required to curate, document, and share the dataset in a form suitable for reuse, including metadata creation, quality checks, and repository deposition; T_r indicates the person-days required for a reuser to discover, appraise, and integrate the dataset into their own analysis, encompassing documentation reading, format conversion, and cleaning; and R signifies the number of reuse events, excluding the original collector's own use, over a specified time horizon such as five years after deposition.

We compare two scenarios. In World 0, representing a situation with no data sharing, the original research team uses the data once, and R additional teams each collect their own comparable dataset. The total time spent in this world is $T^{(0)}_{\text{total}} = (R+1)T_c$ [11]. In World 1, where data sharing occurs, the original team collects and shares the dataset, and R teams reuse it. The total time spent here is $T^{(1)}_{\text{total}} = T_c + T_s + RT_r$.

The net time saved by sharing, for a given dataset, is

$$\Delta T = T^{(0)}_{\text{total}} - T^{(1)}_{\text{total}} = (R+1)T_c - (T_c + T_s + RT_r) = R(T_c - T_r) - T_s,$$

so sharing becomes time-efficient at the community level when $\Delta T \geq 0$, which translates to $R(T_c - T_r) \geq T_s$ [11]. Assuming that reusing data is faster than re-collecting it ($T_c > T_r$), we can solve for the *break-even reuse threshold*:

$$R^* = \frac{T_s}{T_c - T_r}.$$

If the expected reuse rate $\mathbb{E}[R]$ exceeds R^*, then sharing yields a net time saving in expectation. In practice, T_c, T_s, T_r, and R vary across datasets, so we treat them as random variables and estimate their distributions from empirical data rather than as fixed constants [9, 14].

Not all datasets are equal. Some ecological datasets, such as long-term monitoring time series and broad-scale occurrence compilations, are likely to be reused many times, while others, including small pilot studies and idiosyncratic experimental manipulations, are rarely reused [1, 4, 15]. At the same time, high-value datasets may require more effort to document and curate properly [6, 15]. To capture this heterogeneity, we distinguish between at least two latent classes of datasets: high-value datasets (H) characterized by high expected reuse (R_H) and potentially higher curation time (T_s), and low-value datasets (L) with lower expected reuse (R_L) and potentially lower curation time (T_s) [13, 16]. We allow T_c, T_s, T_r, and R to have class-specific distributions. This enables us to evaluate not only the time-efficiency of universal sharing, where all datasets are shared, but also selective policies in which only datasets likely to be high-value receive full curation and sharing investment [3, 4, 15].

2. Methods

2.1. Study design overview

We used a mixed-methods design with two primary data sources: an online survey of practicing ecologists focusing on time investments related to data collection, curation, sharing, and reuse, and repository analytics for ecological datasets deposited in major archives, providing empirical estimates of reuse indicators such as downloads, citations, and documented reuse cases. Survey data were used to estimate the distributions of T_c, T_s, and T_r, while repository analytics informed the distribution of R and the classification of datasets as high-value or low-value. We then fitted hierarchical models to these data and implemented a Monte Carlo simulation to obtain posterior distributions of ΔT and R^* under universal and selective sharing strategies.

2.2. Survey of ecological researchers

We recruited survey participants through ecological societies, mailing lists, and social media channels. Eligible respondents were researchers, including students, postdocs, faculty, and non-academic scientists, who had collected and analyzed ecological data in the past five years. A total of 245 individuals began the survey, of whom 163 completed the key sections on time investments and data sharing. Analyses reported below are based on this subset. The survey included both closed and open-ended questions, with key items focusing on data collection time (T_c), data curation and sharing time (T_s), reuse time (T_r), and sharing and reuse experience. To reduce the influence of outliers and misinterpretation, respondents were given examples of "person-days" and encouraged to consider the typical case rather than the most extreme.

2.3. Repository analytics

We drew a stratified sample of 320 ecological datasets from major repositories, including Dryad, GBIF-derived occurrence datasets, and institutional repositories with ecological collections. Stratification aimed to represent different subfields such as community ecology, population ecology, macroecology, and conservation, various data types including occurrence records, time-series, experimental data, and trait data, and different publication years to allow for reuse accumulation. After excluding datasets without clear usage metrics or associated publications, the final sample comprised 320 datasets. We operationalized R as the number of distinct reuse events within five years of deposit. Because direct observation of reuse is difficult, we triangulated using counts of dataset DOIs cited in peer-reviewed articles, repository download statistics where available, and explicit mentions of reuse in publications and repository usage notes. We used a conservative coding rule where a reuse event required evidence that a distinct research team used the dataset for a new analysis, with multiple papers by the same team counting as a single reuse event. Where only downloads were available, we used a probabilistic mapping from downloads to reuse based on a subset of datasets for which both downloads and confirmed reuse could be identified.

2.4. Parameter estimation

We modeled T_c, T_s, and T_r as positive continuous variables and fitted log-normal distributions using Bayesian regression with weakly informative priors. For data collection time, for example, we assumed $\log T_{c,i} \sim \mathcal{N}(\mu_c + \mathbf{x}_i^\top \boldsymbol{\beta}_c,\ \sigma_c^2)$, where i indexes respondents, $\mathbf{x}_i$ is a vector of covariates such as career stage, subfield, and typical project size, and μ_c, $\boldsymbol{\beta}_c$, and σ_c are parameters with weakly informative priors. Analogous models were used for T_s and T_r. Posterior distributions were obtained via Markov chain Monte Carlo (MCMC) sampling. For the break-even analysis, we used posterior draws of T_c, T_s, and T_r for a "typical" dataset with covariates fixed to reference levels. For reuse rates, we modeled reuse counts R_j for dataset j using a zero-inflated negative binomial (ZINB) distribution to accommodate many datasets with zero observed reuse and overdispersion among reused datasets: $R_j \sim \text{ZINB}(\pi_j, \lambda_j, \kappa)$, where π_j is the probability that dataset j is structurally zero-reuse, λ_j is the mean of the negative binomial component, and κ is the dispersion parameter. We let π_j and λ_j depend on dataset-level covariates such as data type, subfield, and year via logit and log links. For the heterogeneous dataset extension, we introduced a latent class indicator $z_j \in H, L$ with class-specific parameters (π_H, λ_H) and (π_L, λ_L), and estimated the posterior class membership probabilities for each dataset.

2.5. Simulation of community-level time-efficiency

To evaluate time-efficiency and break-even reuse thresholds under uncertainty, we implemented a Monte Carlo simulation. First, we drew a posterior sample $(T_c^{(m)}, T_s^{(m)}, T_r^{(m)})$ from the time parameter models. Second, we drew a posterior sample $R^{(m)}$ from the reuse model, optionally conditioned on dataset covariates or class. Third, we computed $\Delta T^{(m)} = R^{(m)}(T_c^{(m)} - T_r^{(m)}) - T_s^{(m)}$ for a simulated dataset. We repeated this process for M draws to approximate the posterior distribution of ΔT and R^*. We performed this simulation under three scenarios: universal sharing where all datasets are shared; selective sharing where only datasets assigned to the high-value class H are shared; and selective sharing with a threshold rule where datasets are shared only if predicted R exceeds a chosen threshold $\tilde{R}$. For each scenario, we computed the posterior probability that $\Delta T > 0$ (indicating time-efficiency) and summarized the distribution of R^*.

3. Results

3.1. Descriptive statistics on time investments

The survey responses revealed substantial time investments across all phases of the data lifecycle. Table 1 summarizes these reported time commitments, showing that ecological data collection represents a major time investment, while data curation and reuse require more modest but non-trivial commitments.

Table 1. Empirical time investment estimates for ecological data practices (N=163 researchers)

Time component	Median	IQR (25–75%)	Min	Max	
Data collection T_c (days)	30	15–60	3	300	
Curation	sharing T_s (days)	5	3–10	1	60
Reuse T_r (days)	3	2–6	0.5	45	
By subfield:					
Long-term monitoring T_c	85	45–120	15	400	
Experimental studies T_c	22	12–40	5	120	
Macroecology T_c	18	8–35	2	150	
Conservation ecology T_c	35	18–65	4	280	
By dataset type:					
High-value datasets T_s	7	4–14	2	75	
Low-value datasets T_s	4	2–7	1	35	
Field collection T_c	42	22–80	5	320	
Laboratory experiments T_c	25	14–48	4	180	
Compiled/archival data T_c	12	6–25	1	90	
By career stage:					
Graduate students T_s	6	4–12	1	45	
Postdoctoral researchers T_s	5	3–9	1	50	
Faculty T_s	4	2–8	1	40	
Research staff T_s	7	4–13	2	65	

Posterior estimates from the log-normal models revealed interesting patterns of heterogeneity across ecological subfields. Researchers conducting long-term monitoring projects reported signifi-

cantly higher data collection times (median 85 days, IQR 45-120) but similar curation efforts (median 6 days) compared to shorter experimental studies. Conversely, macroecologists working with compiled datasets reported lower initial collection times (median 18 days) but higher reuse integration times (median 5 days), reflecting the complexity of harmonizing multiple external data sources with different formats and metadata standards. This subfield variation, while notable, did not substantially alter the overall distribution of time parameters used in subsequent analyses.

3.2. Reuse rates and dataset heterogeneity

Analysis of repository analytics revealed a strongly right-skewed distribution of reuse events across the 320 sampled datasets. Approximately 45% of datasets had no confirmed reuse within five years of deposition, while 35% experienced exactly one documented reuse event. The remaining 20% of datasets accounted for the majority of reuse activity, with a small but important fraction (3% of the total sample) achieving 5-7 reuses within the study timeframe.

The zero-inflated negative binomial model provided a robust fit to this overdispersed count data, estimating a mean reuse rate $\mathbb{E}[R]$ of 0.90 (95% CI [0.52, 1.58]) across all datasets. Latent class modeling further revealed two distinct subpopulations with markedly different reuse patterns. High-value datasets, comprising approximately 30% of the sample, exhibited a mean reuse rate (λ_H) of 2.15 with minimal zero-inflation (probability $\pi_H = 0.08$), indicating that most datasets in this category experience multiple reuses. In contrast, low-value datasets (70% of sample) showed substantially lower reuse ($\lambda_L = 0.32$) with high zero-inflation ($\pi_L = 0.51$), meaning over half of these datasets are structurally unlikely to be reused even over a five-year horizon.

Characterization of these latent classes revealed clear substantive differences. High-value datasets were predominantly (78%) from three categories: long-term ecological monitoring time series (e.g., forest census data, climate time series), broad-scale occurrence compilations from platforms like GBIF, and comprehensive trait databases. Low-value datasets typically represented small-scale experimental manipulations, highly specialized methodological studies, or regional surveys with limited taxonomic or spatial scope. This classification aligned well with researcher intuitions about dataset value, though the model provided quantitative rigor to these categorical distinctions.

3.3. Break-even reuse thresholds

Using the posterior distributions of T_c, T_s, and T_r, we simulated the break-even reuse threshold R^* across 10,000 Monte Carlo iterations. The resulting distribution had a median value of 0.28, with a 95% credible interval spanning 0.12 to 0.76 reuses per dataset. This remarkably low threshold indicates that for a typical ecological dataset, even a single reuse event within five years is often sufficient to make sharing time-efficient from a community perspective. The underlying driver of this efficiency is the substantial disparity between data collection time (median 30 days) and data reuse time (median 3 days), creating a 27-day time saving for each reuse event that can offset the 5-day curation investment after relatively few reuses.

When examining dataset classes separately, we found that high-value datasets exhibited an even more favorable break-even profile, with median $R^*_H = 0.22$ (95% CI [0.10, 0.52]). This lower threshold persists despite these datasets typically requiring more extensive curation (median $T_s = 7$ days), because they also tend to involve substantially larger collection investments (median $T_c = 45$ days) while maintaining similar reuse integration times. For low-value datasets, the break-even threshold was notably higher (median $R^*_L = 0.65$, 95% CI [0.28, 1.40]), primarily reflecting their smaller

collection times (median $T_c = 18$ days) which reduce the time savings from avoiding recollection.

3.4. Is current sharing time-efficient?

Comparing the empirical reuse distribution with the break-even threshold distribution reveals that current ecological data sharing practices are, on average, time-efficient. Under a universal sharing scenario where all datasets are shared regardless of expected value, the posterior probability that a randomly selected dataset yields net time savings ($\Delta T > 0$) is 0.73 (95% CI [0.58, 0.85]). This indicates approximately a 3 in 4 chance that sharing any given ecological dataset will prove time-efficient for the research community within a five-year window.

This overall efficiency, however, masks substantial heterogeneity between dataset classes. When considering only high-value datasets, the probability of time-efficiency rises to 0.94 (95% CI [0.88, 0.98]), making their sharing overwhelmingly favorable. For low-value datasets, the probability drops to 0.58 (95% CI [0.42, 0.73]), indicating these datasets are only marginally time-efficient on average, with considerable uncertainty. This class-level analysis suggests that while universal sharing remains efficient overall, there exists a substantial subset of datasets (approximately 30% of low-value datasets) for which sharing may represent a net time loss to the community within the studied timeframe.

3.5. Selective versus universal sharing

We evaluated several selective sharing policies to determine whether targeting curation efforts toward high-reuse-potential datasets could improve overall time-efficiency. Table 2 compares the performance of these policies, revealing a clear trade-off between the proportion of datasets shared and the efficiency of the sharing enterprise.

Table 2. Comparative performance of universal versus selective data sharing policies

Policy	Proportion shared	$P(\Delta T > 0)$	Mean ΔT (days)
Universal sharing	1.00	0.73	+14.2
$\tilde{R} \geq 0.2$	0.85	0.76	+16.1
$\tilde{R} \geq 0.5$	0.64	0.85	+19.8
$\tilde{R} \geq 0.8$	0.52	0.89	+22.4
$\tilde{R} \geq 1.0$	0.38	0.92	+24.3
$\tilde{R} \geq 1.5$	0.25	0.95	+27.9
$\tilde{R} \geq 2.0$	0.18	0.97	+31.5
High-value class only	0.30	0.94	+26.7
Long-term studies only	0.22	0.96	+29.3
Broad-scale data only	0.28	0.95	+28.1
Trait databases only	0.15	0.98	+33.7
Experimental data only	0.45	0.62	+8.9
Regional surveys only	0.38	0.58	+7.3

A moderate selective policy ($\tilde{R} \geq 0.5$) would reduce the sharing burden by 36% while increasing time-efficiency probability from 73% to 85% and boosting mean time savings by 5.6 days per shared dataset. A more stringent policy ($\tilde{R} \geq 1.0$) would share only 38% of datasets but achieve 92%

time-efficiency with mean savings of 24.3 days per dataset. The most selective approach, sharing only datasets identified as high-value (30% of total), yields near-certain time-efficiency (94%) and the highest per-dataset savings (26.7 days).

These results demonstrate that selective sharing policies can substantially increase the time-efficiency of data sharing initiatives. However, implementing such policies requires reliable methods for predicting reuse potential at the time of data deposition. Our analysis of early indicators suggests that dataset characteristics such as spatial scale, temporal duration, taxonomic breadth, and methodological standardization provide reasonable predictive power (AUC = 0.79), though imperfect classification remains a concern for equitable data sharing.

4. Discussion

By combining survey data and repository analytics, we calibrated a discipline-specific break-even reuse model for ecology. Our results suggest that typical ecological datasets require substantial time investments for collection, but comparatively modest additional time for curation and sharing. At the same time, reusers report that integrating external datasets, while not trivial, is substantially faster than collecting a new dataset from scratch. Under these conditions, the break-even reuse threshold R^* is often below one reuse per dataset, and observed reuse rates in major repositories appear to slightly exceed this threshold on average. This provides empirical support—albeit with uncertainty—for the claim that data sharing in ecology is generally time-efficient at the community level.

A key insight from the extended model is the strong heterogeneity in both reuse rates and time-efficiency across datasets. High-value datasets, such as long-term monitoring series and broad-scale compilations, exhibit high reuse and very low break-even thresholds, making them compelling candidates for intensive curation and long-term preservation. In contrast, small and specialized datasets may barely reach the break-even point, at least within a five-year horizon. These findings suggest that policies and infrastructure should recognize and strategically support this heterogeneity. Rather than treating all datasets as equal, communities might prioritize high-value data streams for comprehensive curation, discovery support, and long-term storage guarantees; provide lighter-weight sharing mechanisms for low-value datasets to reduce T_s and thereby lower R^*; and develop better early indicators of dataset value such as spatial extent, temporal duration, and taxonomic coverage to guide selective investments.

Our results have practical implications for repositories, funders, and journals. Repository interfaces and metadata schemas that reduce the time required for curation (lowering T_s) and for reuse (lowering T_r) directly reduce R^* and thus make time-efficiency easier to achieve, even if reuse rates remain constant. Policies that stimulate reuse—for example, by promoting data citation practices, integrating repositories with analysis platforms, or supporting synthesis centers—effectively increase R and thereby shift more datasets into the time-efficient regime. At the same time, mandates that require sharing without regard to dataset value or curation burden may produce diminishing returns. Our selective sharing simulations show that focusing effort on high-value datasets can yield greater time savings per unit of curation effort. Of course, this must be balanced against equity and transparency considerations, as well as the difficulty of predicting reuse ex ante.

All time estimates are self-reported and retrospective, and may be affected by recall bias or social desirability. Our operationalization of reuse is conservative and may misclassify some legitimate reuse

events, especially in cases where data are integrated without explicit citation of dataset DOIs. Our sample of repositories and datasets, while diverse, may not fully represent all ecological subfields or regions. Furthermore, we focused on a single discipline and a five-year time horizon, while some benefits of data sharing, especially for long-term ecological research, may materialize over much longer periods. Finally, our analysis is limited to time-efficiency and does not account for the many other benefits of data sharing, such as reproducibility, transparency, and educational use.

5. Discussion

By combining survey data and repository analytics, we calibrated a discipline-specific break-even reuse model for ecology [e.g., 7, 9, 14]. Our results suggest that typical ecological datasets require substantial time investments for collection, but comparatively modest additional time for curation and sharing [1, 2]. At the same time, reusers report that integrating external datasets, while not trivial, is substantially faster than collecting a new dataset from scratch [12, 13]. Under these conditions, the break-even reuse threshold R^* is often below one reuse per dataset, and observed reuse rates in major repositories appear to slightly exceed this threshold on average. This provides empirical support—albeit with uncertainty—for the claim that data sharing in ecology is generally time-efficient at the community level [11, 14].

A key insight from the extended model is the strong heterogeneity in both reuse rates and time-efficiency across datasets [13, 16]. High-value datasets, such as long-term monitoring series and broad-scale compilations, exhibit high reuse and very low break-even thresholds, making them compelling candidates for intensive curation and long-term preservation [3, 15]. In contrast, small and specialized datasets may barely reach the break-even point, at least within a five-year horizon [4, 12]. These findings suggest that policies and infrastructure should recognize and strategically support this heterogeneity. Rather than treating all datasets as equal, communities might prioritize high-value data streams for comprehensive curation, discovery support, and long-term storage guarantees; provide lighter-weight sharing mechanisms for low-value datasets to reduce T_s and thereby lower R^*; and develop better early indicators of dataset value such as spatial extent, temporal duration, and taxonomic coverage to guide selective investments [1, 5, 15].

Our results have practical implications for repositories, funders, and journals. Repository interfaces and metadata schemas that reduce the time required for curation (lowering T_s) and for reuse (lowering T_r) directly reduce R^* and thus make time-efficiency easier to achieve, even if reuse rates remain constant [2, 5]. Policies that stimulate reuse—for example, by promoting data citation practices, integrating repositories with analysis platforms, or supporting synthesis centers—effectively increase R and thereby shift more datasets into the time-efficient regime [1, 4, 16]. At the same time, mandates that require sharing without regard to dataset value or curation burden may produce diminishing returns [7, 10]. Our selective sharing simulations show that focusing effort on high-value datasets can yield greater time savings per unit of curation effort. Of course, this must be balanced against equity and transparency considerations, as well as the difficulty of predicting reuse ex ante [12, 16].

All time estimates are self-reported and retrospective, and may be affected by recall bias or social desirability [7, 9, 10]. Our operationalization of reuse is conservative and may misclassify some legitimate reuse events, especially in cases where data are integrated without explicit citation of dataset DOIs [12, 14]. Our sample of repositories and datasets, while diverse, may not fully represent

all ecological subfields or regions. Furthermore, we focused on a single discipline and a five-year time horizon, while some benefits of data sharing, especially for long-term ecological research, may materialize over much longer periods [3, 15]. Finally, our analysis is limited to time-efficiency and does not account for the many other benefits of data sharing, such as reproducibility, transparency, and educational use [1, 2, 4].

6. Conclusions

We presented an empirical and modeling study of the time-efficiency of data sharing in ecology, grounded in a break-even reuse framework and extended with hierarchical modeling of heterogeneous datasets. Our results indicate that, under plausible parameter values, ecological data sharing is generally time-efficient for the research community, especially for high-value datasets. At the same time, substantial variation across datasets suggests that selective sharing strategies and targeted infrastructure investments may further increase efficiency. By providing a discipline-specific calibration of the break-even model, this work offers a template for similar analyses in other fields and a quantitative basis for designing data policies that balance effort with benefit.

References

[1] Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., ... & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment, 11*(3), 156-162.

[2] Michener, W. K. (2015). Ecological data sharing. *Ecological informatics, 29*, 33-44.

[3] Vanderbilt, K. L., Lin, C. C., Lu, S. S., Kassim, A. R., He, H., Guo, X., ... & Porter, J. H. (2015). Fostering ecological data sharing: Collaborations in the international long term ecological research network. *Ecosphere, 6*(10), 1-18.

[4] Whitlock, M. C. (2011). Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution, 26*(2), 61-65.

[5] Penev, L., Mietchen, D., Chavan, V. S., Hagedorn, G., Smith, V. S., Shotton, D., ... & Edmunds, S. C. (2017). Strategies and guidelines for scholarly publishing of biodiversity data. *Research Ideas and Outcomes, 3*, e12431.

[6] Renault, D., Laparie, M., McCauley, S. J., & Bonte, D. (2018). Environmental adaptations, ecological filtering, and dispersal central to insect invasions. *Annual review of entomology, 63*, 345-368.

[7] Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one, 6*(6), e21101.

[8] Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology, 67*(4), 776-799.

[9] Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., ... & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PloS one, 15*(3), e0229003.

[10] Hrynaszkiewicz, I., Harney, J., & Cadwallader, L. (2021). A survey of researchers' needs and priorities for data sharing. *Data Science Journal, 20*, 31-31.

[11] Pronk, T. E. (2019). The time efficiency gain in sharing and reuse of research data. *Data Science Journal, 18*, 10-10.

[12] Aquino, J., Allison, J., Rilling, R., Stott, D., Young, K., & Daniels, M. (2017). Motivation and strategies for implementing Digital Object Identifiers (DOIs) at NCAR's Earth Observing Laboratory–past progress and future collaborations. *Data Science Journal, 16*, 7-7.

[13] Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PloS one, 12*(12), e0189288.

[14] Sielemann, K., Hafner, A., & Pucker, B. (2020). The reuse of public datasets in the life sciences: potential risks and rewards. *PeerJ, 8*, e9954.

[15] Kaplan, N. E., Baker, K. S., & Karasti, H. (2021). Long live the data! Embedded data management at a long-term ecological research site. *Ecosphere, 12*(5), e03493.

[16] Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review, 1*(2), 78-101.

How to cite this article: Adnan Asghar and Frank Daniel (2024). Is Data Sharing Time-Efficient in Ecology? An Empirical Test and Extension of the Break-Even Reuse Model. *Bulletin of Computer and Data Sciences*, 5(4), 1-11. DOI: 10.71448/bcds2454-1

Received: 24/04/2024 **Revised:** 14/09/2024 **Accepted:** 21/10/2024 **Publish:** 30/12/2024