Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

To correct or not to correct?: Assessing the multiple comparisons problem for association rule mining of environmental DNA (eDNA) detection survey datasets

Nikolett Toth [†], M. Luiza Antonie[†], Jarrett D. Phillips[†,*]
[†] University of Guelph

Abstract

Unsupervised machine learning is a valuable exploratory tool for the small, noisy, and incomplete datasets characteristic of ecological and environmental research, where data limitations often render supervised approaches impractical. Here, we apply association rule mining (ARM) to a brook trout (*Salvelinus fontinalis*) environmental DNA (eDNA) dataset with two objectives: (1) demonstrating ARM as a screening tool to identify environmental correlates and guide targeted metadata collection, and (2) evaluating Bonferroni and Benjamini–Hochberg corrections for managing Type I error rates during significance-based pruning. Our results show that, while both methods retained high-quality rules, the Bonferroni procedure eliminated several ecologically interesting associations that survived Benjamini–Hochberg correction. For small environmental datasets in an exploratory context, we thus favour Benjamini–Hochberg as a more appropriate correction strategy, notwithstanding the need for further external validation.

Keywords: Abiotic metadata variables, Association rule mining, Brook trout, eDNA, Ecology, Spatiotemporal sampling

1. Introduction

Ecological and environmental research is often constrained by small, noisy, and incomplete data sets, rendering supervised machine learning applications impractical. In such settings, unsupervised methods like association rule mining (ARM) [1] represent valuable exploratory tools, identifying associations between environmental variables to inform targeted data collection and thus support future supervised modeling. ARM is particularly appealing in this context as its human-readable rules are actionable for domain experts tasked with optimizing data collection but who lack machine learning expertise.

A promising yet (to our knowledge) unexplored application of ARM is environmental DNA (eDNA) research. eDNA is genetic material (*e.g.*, skin cells, mucous) shed by organisms into water, soil, sediment, and air [2]. By collecting and sequencing this material, ecologists can identify the species present at a particular site, allowing unobtrusive monitoring of at-risk and invasive populations. However, broader adoption of eDNA as a standard biomonitoring tool has been limited by unreliable detection: DNA degrades over time at rates sensitive to environmental conditions, and the factors most responsible for degradation remain poorly characterized [3]. This issue is compounded by the absence of standardized metadata collection protocols, resulting in unrepresentative datasets [4]. ARM is well-positioned to address these limitations by identifying which environmental conditions are most reliably associated with eDNA detection, thus informing targeted metadata collection.

One challenge inherent to ARM is the large number of candidate rules generated (even for small datasets) many of which are likely spurious. Statistical significance testing can help prune rules; however, testing large rulesets without correction results in inflated Type I error rates (*i.e.* falsely identifying nonexistent associations as significant). Multiple testing

* jphill01@uoguelph.ca

This article is © 2026 by author(s) as listed above. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author(s) identified above.

corrections such as the Bonferroni [5] and Benjamini–Hochberg [6] methods can address this, but their utility for ARM in data-limited ecological settings remains largely unexplored.

Here, we apply ARM to a brook trout (*Salvelinus fontinalis*) eDNA dataset with two objectives: (1) to demonstrate its utility as an exploratory screening tool for small ecological datasets and (2) to evaluate the effect of the Bonferroni and Benjamini–Hochberg multiple testing adjustment procedures on the quality and statistical reliability of rules.

2. Methods and Data

2.1. Association Rule Mining

ARM generates implication rules of the form $X \implies Y$, where X and Y denote non-empty, disjoint itemsets from transactional data. In this context, X is termed the antecedent or left hand side (LHS), while Y is termed the consequent or right hand side (RHS). Rule quality is commonly evaluated using three key measures of interestingness: support (a measure of joint probability ($P(X,\ Y)$), *i.e.*, the frequency with which X and Y co-occur), confidence (the conditional probability $P(Y|X)$), and lift (the ratio of observed to expected co-occurrence under independence, with values greater than one indicating positive association). The Apriori algorithm [7] is widely used to efficiently mine association rules by iteratively generating candidate itemsets and pruning those that fail to meet minimum support and confidence thresholds.

2.2. Significance Testing and Correction

While significance testing can further refine candidate rules beyond standard interestingness thresholds, testing multiple rules simultaneously introduces a new problem: inflated Type I error rates. Here, we compare two correction approaches. The Bonferroni correction controls the probability of observing at least one false positive association, known as the family-wise error rate (FWER), by testing each rule at $p = \frac{\alpha}{n}$ (where p is the adjusted significance threshold, α is the desired significance level, and n is the number of tests). The Benjamini-Hochberg correction instead controls the false discovery rate (FDR), the expected proportion of false positives among rejected hypotheses, by ranking p-values and applying adaptive thresholds that increase with rank. Bonferroni provides stricter Type I error control, while Benjamini-Hochberg offers greater statistical power. Though previously studied in ARM [8], these corrections are understudied for ARM in data-limited ecological settings.

2.3. Brook Trout Dataset and Analysis Pipeline

Data are derived from Nolan *et al.* [9], who compared the detection of native brook trout via eDNA sampling and electrofishing (a traditional technique that involves passing an electric current through water to temporarily stun fish, allowing them to be counted and identified). The dataset includes brook trout eDNA concentrations and electrofishing counts, along with eight physicochemical metadata variables, for a total of 126 transactions with no missing values.

Since ARM requires categorical data, continuous variables were discretized prior to mining. Most variables were divided into `high` and `low` categories using biologically meaningful thresholds from the literature. eDNA concentrations were discretized at the assay's limit of detection (LOD), the minimum concentration reliably detectable by the quantification method [10]. Electrofishing counts were binarized into `present` and `absent`. Remaining variables were split at the median since data were considerably skewed. The full list of variables with discretization schemes is provided in **Appendix A**.

Rules were mined in R [11] using the `arules` [12, 13] package via the `apriori()` function. Minimum confidence and support thresholds were set to $\frac{1}{n}$, where n is the number of transactions ($n = 126$; $\frac{1}{126} = 0.794\%$), to ensure that a given rule is supported by at least one transaction. Permissive thresholds were chosen to preserve rare associations and retain a broad candidate ruleset for exploratory analysis. Our analysis targeted the consequents {`eDNAConc = high`} and {`eFishCatch = present`}, as these are most relevant to species detection and occupancy monitoring.

Results were initially pruned using `is.redundant()` from `arules`, which removes rules that provide no additional information beyond a more general rule with equal or higher confidence [14]. Finally, `is.significant()` from `arules` was applied to the non-redundant set to retain rules significant at $\alpha = 0.05$ using one-sided Fisher's Exact Tests, with p-values adjusted using Bonferroni and Benjamini–Hochberg corrections via R's `p.adjust()` function. The full implementation is available in the associated GitHub repository [15].

3. Results

3.1. Rule Filtering and Counts

Across the 126 transactions and the two considered consequents, a total of 8226 association rules were initially mined. The total rule counts for both consequents after each filtering step are shown in **Table 1**. Overall, Bonferroni retained the fewest rules across both consequents, followed by Benjamini–Hochberg, and then the uncorrected approach.

Consequent	Mined	Non-Redundant	Uncorrected	Benjamini-Hochberg	Bonferroni
{`eDNAConc = high`}	3814	75	41	37	3
{`eFishCatch = present`}	4412	17	14	14	9

Table 1. Counts of association rules by consequent at each filtering stage: the unfiltered ruleset (Mined), the subset remaining after removing redundant rules (Non-Redundant), and the subset of rules deemed statistically significant at $\alpha = 0.05$ under no multiple-testing adjustment (Uncorrected), the Benjamini–Hochberg procedure and the Bonferroni correction.

3.2. Rules of Interest

Complete rulesets for both consequents at each filtering stage (raw, non-redundant, uncorrected, Benjamini–Hochberg, and Bonferroni), along with all associated interestingness measures and correlation statistics (including Spearman coefficients and adjusted p-values), are available in the accompanying GitHub repository [15]. **Table 2** highlights some statistically significant rules of interest from the retained sets.

Antecedent	Consequent	Confidence	Support	Lift	Retained Under
{`WaterTemp = high, VolumeFiltered = high`}	{`eDNAConc = high`}	1.00	0.095	1.68	BH
{`DissolvedOxygen = low, VolumeFiltered = low`}	{`eDNAConc = high`}	0.89	0.127	1.49	BH
{`eDNAConc = high`}	{`eFishCatch = present`}	0.91	0.540	1.12	All

Table 2. Selected association rules of interest. "Retained Under" indicates which correction schemes deemed the rule significant: All (Bonferroni, Benjamini–Hochberg, and uncorrected), BH (Benjamini–Hochberg and uncorrected only), or Uncorr (uncorrected only).

3.3. Statistical Significance and Rule Quality

To assess the effect of multiple testing correction on rule quality, we examined the relationship between p-values and interestingness measures using Spearman rank correlations

on the non-redundant rulesets, with p-values adjusted using Benjamini–Hochberg. For the ruleset predicting {`eDNAConc = high`} ($n = 75$), confidence exhibited strong negative correlations with p-values across all correction strategies (Spearman ρ ranging from -0.600 to -0.798, all significant), indicating that significance-based pruning preferentially retained high-confidence rules regardless of correction method (**Figure 1**). Support showed no significant correlation. Rule length showed a weak negative correlation with p-values (Spearman ρ ranging from -0.287 to -0.420, all significant), though the relationship is visually diffuse (**Figure 1**) and thus may not represent a practically meaningful trend. No significant correlations were found for {`eFishCatch = present`} rules, likely due to the small non-redundant ruleset ($n = 17$). Scatterplots for support and lift are provided in **Appendix A**, while those for the consequent {`eFishCatch = present`} are available in the GitHub repository [15].

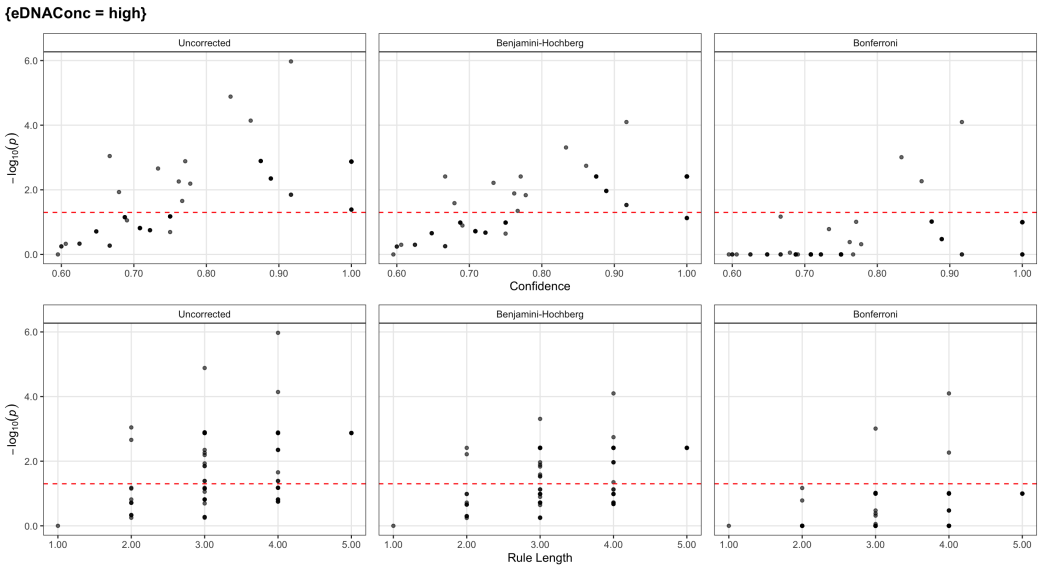

Figure 1. Scatterplots of association rule confidence (top row) and length (bottom row) against $-\log_{10}(p)$ for the consequent {`eDNAConc = high`}. Each point is a single, non-redundant rule. Columns correspond to uncorrected p-values, Benjamini-Hochberg-adjusted p-values, and Bonferroni-adjusted p-values computed using one-sided Fisher's exact tests. The dashed red line indicates the log-transformed significance threshold ($\alpha = 0.05$; $-\log_{10}(0.05) = 1.30$).

4. Discussion

4.1. Rule Volume and Filtering

Even with only 126 transactions, 8226 candidate rules were generated, far exceeding what a domain expert could meaningfully interpret. Redundancy pruning accounted for the largest reduction in rule volume, resulting in just 92 non-redundant rules. However, redundancy pruning removes only structurally overlapping rules and does not address statistical validity: at $\alpha = 0.05$, approximately four false positives would still be expected among the 75 non-redundant {`eDNAConc = high`} rules by chance alone, motivating significance testing as a secondary refinement. The need for further refinement is compounded by the deliberately permissive thresholds of $\frac{1}{n}$ used in this analysis. While these thresholds maximize exploratory breadth, the tradeoff is that they also retain weak and potentially unstable rules, placing the burden of quality control on post hoc significance testing. As Liu *et al.* note, stricter minima would reduce this but risk discarding genuine associations [8].

4.2. **ARM as an Exploratory Screening Tool**

ARM has previously been applied in biological contexts including bioinformatics [16], transcription factor analysis [17], and microbiome research [18]; to our knowledge, its application to eDNA data is novel. We expand on the rules highlighted in Section 3.2 below, examining their ecological implications as a proof-of-concept of ARM's ability to identify actionable associations in a data-limited eDNA context.

The recovery of the rule {`eDNAConc = high`} ⇒ {`eFishCatch = present`} under all three correction schemes provides informal validation of the pipeline, confirming that ARM recovers expected biological relationships. Beyond this, several multi-variable associations were identified that would not be readily apparent from traditional univariate statistical analyses (**Appendix A**). The rule {`WaterTemp = high, VolumeFiltered = high`} ⇒ {`eDNAConc = high`} suggests that collecting larger sample volumes may compensate for thermally accelerated eDNA degradation, consistent with recommendations to increase sampling effort under high-degradation conditions [19]. Another ecologically interesting association was identified in the rule {`DissolvedOxygen = low, VolumeFiltered = low`} ⇒ {`eDNAConc = high`}. Previous research has demonstrated that eDNA degradation rates decline with increasing biochemical oxygen demand, chlorophyll concentration, and total eDNA [20], suggesting that high biological activity environments may paradoxically preserve rather than degrade eDNA. Consistent with this, the rule may reflect environments with high biological activity, where increased oxygen consumption drives down dissolved oxygen levels while simultaneously elevating eDNA shedding, producing a detectable signal even from small sample volumes.

Note that the above interpretations should be treated with caution, as each rule is a hypothesis from a small, noisy observational dataset rather than a validated finding. External validation would be required to confirm whether retained associations reflect genuine ecological signals rather than false discoveries. Nevertheless, they represent a starting point for domain experts to prioritize which metadata variables are worth collecting, laying the groundwork for higher-quality datasets and thus future supervised models.

4.3. **Multiple Testing Corrections**

The two correction approaches produced markedly different rulesets, reflecting their distinct priorities in error control. Benjamini–Hochberg was permissive, preserving most uncorrected rules, while Bonferroni resulted in more drastic reductions, retaining only three rules for {`eDNAConc = high`}. Notably, many of the ecologically interesting associations from Section 3.2 were eliminated under Bonferroni but preserved under Benjamini–Hochberg. Regardless of correction method, confidence was strongly negatively correlated with p-values across all schemes (**Figure 1**), indicating that significance-based pruning preferentially retains high-quality rules. For small environmental datasets in an exploratory context, we therefore favour Benjamini–Hochberg as the more appropriate correction strategy. This recommendation is limited by the absence of external validation. It is also worth noting that correction method is not the only outcome-defining decision in this pipeline: discretization thresholds may similarly shape which associations are recovered. Sensitivity analyses across alternative cutoffs represent an important direction for future work.

Acknowledgements

We would like to thank Daniel (Dan) Gillis, Kathleen (Kat) Nolan, Robert (Bob) Hanner, as well as all present and past members of the Hanner Lab for their valuable input throughout this project. Further, we wish to thank the two analymous reviewers whose input significantly strengthened this manuscript.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. "Mining Association Rules between Sets of Items in Large Databases". In: *SIGMOD Rec.* 22.2 (1993), 207–216.

[2] G. Ficetola, C. Miaud, F. Pompanon, and P. Taberlet. "Species detection using environmental DNA from water samples". In: *Biology Letters* 4 (2008), pp. 423–425.

[3] J. Harrison, J. Sunday, and S. Rogers. "Predicting the fate of eDNA in the environment and implications for studying biodiversity". In: *Proceedings of the Royal Society B* 286 (2019), p. 20191409.

[4] A. Nicholson, D. McIsaac, C. MacDonald, P. Gec, B. Mason, W. Rein, J. Wrobel, M. de Boer, Y. Milián-García, and R. Hanner. "An analysis of metadata reporting in freshwater environmental DNA research calls for the development of best practice guidelines". In: *Environmental DNA* 2.3 (2020), pp. 343–349.

[5] C. E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Firenze: Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[6] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 0035-9246.

[7] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules in Large Databases". In: *Proceedings of the 20th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc., 1994, 487–499.

[8] G. Liu, H. Zhang, and L. Wong. "Controlling false positives in association rule mining". In: *Proceedings of the VLDB Endowment* 5.2 (2011), 145–156.

[9] K. Nolan, T Loeza-Quintana, H. Little, J. McLeod, B. Ranger, D. Bourque, and R. Hanner. "Detection of brook trout in spatiotemporally separate locations using validated eDNA technology". In: *Journal of Environmental Studies and Sciences* 13 (2023), pp. 66–82.

[10] K. E. Klymus, C. M. Merkes, M. J. Allison, C. S. Goldberg, C. C. Helbing, M. E. Hunter, C. A. Jackson, R. F. Lance, A. M. Mangan, E. M. Monroe, A. J. Piaggio, J. P. Stokdyk, C. S. Wilson, and C. A. Richter. "Reporting the limits of detection and quantification for environmental DNA assays". In: *Environmental DNA* 2.3 (2020), pp. 271–282.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

[12] M. Hahsler, B. Gruen, and K. Hornik. "arules – A Computational Environment for Mining Association Rules and Frequent Item Sets". In: *Journal of Statistical Software* 14.15 (2005), pp. 1–25.

[13] M. Hahsler, C. Buchta, B. Gruen, and K. Hornik. *arules: Mining Association Rules and Frequent Itemsets*. Version 1.7-6. 2023. URL: https://CRAN.R-project.org/package=arules.

[14] M. Hahsler and I. Johnson. *arulesCBA: Classification Based on Association Rules*. R package version 1.2.5. 2022. URL: https://CRAN.R-project.org/package=arulesCBA.

[15] Toth, N. and Phillips, J.D. *Canadian-AI-2026-eDNA-Association-Rules*. https://github.com/nikolett0203/Canadian-AI-2026-eDNA-Association-Rules. 2026.

[16] S. Naulaerts, P. Meysman, W. Bittremieux, T. N. Vu, W. Vanden Berghe, B. Goethals, and K. Laukens. "A primer to frequent itemset mining for bioinformatics". eng. In: *Briefings in bioinformatics* 16.2 (2015), pp. 216–231. ISSN: 1467-5463.

[17] G. Ceddia, L. N. Martino, A. Parodi, P. Secchi, S. Campaner, and M. Masseroli. "Association rule mining to identify transcription factor interactions in genomic regions". eng. In: *Bioinformatics* 36.4 (2020), pp. 1007–1013. ISSN: 1367-4803.

[18] A. Giulia, S. Anna, B. Antonia, P. Dario, and C. Maurizio. "Extending Association Rule Mining to Microbiome Pattern Analysis: Tools and Guidelines to Support Real Applications". In: *Frontiers in Bioinformatics* 1 (2022). ISSN: 2673-7647.

[19] M. P. Piggott. "Evaluating the effects of laboratory protocols on eDNA detection probability for an endangered freshwater fish". In: *Ecology and Evolution* 6.9 (2016), pp. 2739–2750. ISSN: 2045-7758.

[20] M. A. Barnes, C. R. Turner, C. L. Jerde, M. A. Renshaw, W. L. Chadderton, and D. M. Lodge. "Environmental Conditions Influence eDNA Persistence in Aquatic Systems". In: *Environmental science technology* 48.3 (2014), pp. 1819–1827. ISSN: 0013-936X.

5. Appendix A

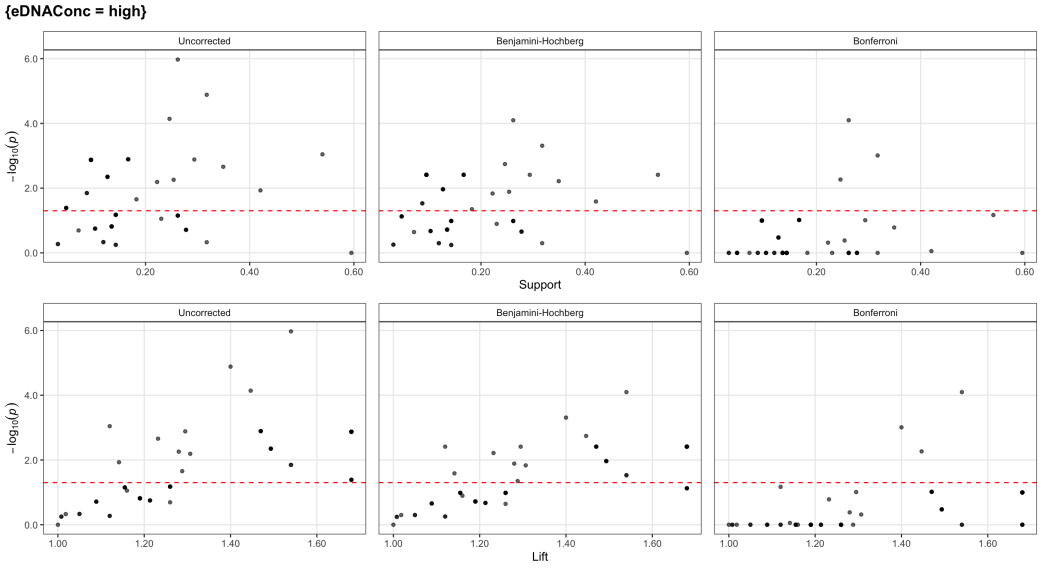

Figure 2. As in Figure 1, but for support (top row) and lift (bottom row).

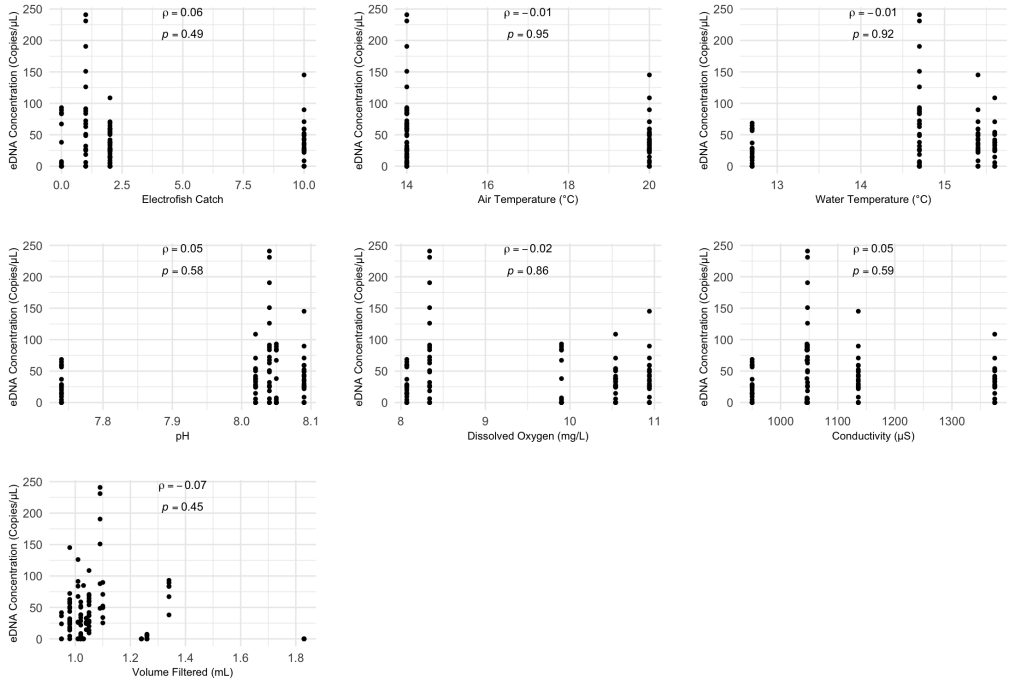

Figure 3. Scatterplots showing the relationship between eDNA concentration and key environmental variables. Each subplot includes the Spearman correlation coefficient (ρ) and associated p-value. All associations are extremely weak and not significantly different from zero. The discreteness of the data reflects repeated measurements at the same values.

Variable	Definition	Unique Values	Discretization	Cutoff
`Backpack`	eDNA sampler employed	2	{`ANDe, OSMOS`}	—
`Site`	site sampled	5	{`1, 2, 3, 4, 5`}	—
`eFishCatch`	site electrofishing count	4	{`absent, present`}	Zero catch is absent, otherwise present
`AirTemp`	air temperature (°C)	2	{`low, high`}	Site mean daily temperature
`WaterTemp`	water temperature (°C)	4	{`low, high`}	Brook trout thermal optimum
`pH`	water pH	5	{`low, high`}	CCME freshwater guideline
`DissolvedOxygen`	water dissolved Oxygen (ppm)	5	{`low, high`}	CCME freshwater guideline
`Conductivity`	water conductivity (μS/cm)	5	{`low, high`}	Dataset median
`VolumeFiltered`	water volume (L)	13	{`low, high`}	Dataset median
`eDNAConc`	eDNA concentration (copies/μL)	83	{`low, high`}	Assay LOD

Table 3. Variable definitions and discretization thresholds used for ARM. *Note:* CCME refers to the Canadian Council of Ministers of the Environment.