

Bivariate and Multivariate Vibe Analysis of Forestry Data with AI and R Statistics

Kato Samuel Namuene¹; Egbe Enow Andrew²

¹Department of Forestry and Wildlife, Faculty of Agriculture and Veterinary Medicine, University of Buea, Buea, Cameroon

²Department of Agronomic and Applied Molecular Sciences, Faculty of Agriculture and Veterinary Medicine, University of Buea, Buea, Cameroon

Publication Date: 2026/05/02

Abstract: Artificial Intelligence (AI) can be used to speed up data analysis with R statistics by generating R code which is executed in R (vibe data analysis), reducing the time a manual data analyst takes to develop R code. A reproducible, AI-assisted framework for bivariate and multivariate statistical analysis of forestry count data was developed and validated in this study, integrating vibe data analysis with conventional manual methods using four disturbance observations (snapping, windthrow, branch fall, and dead standing) across 73 species drawn from 183 treefall gaps in Korup National Park, Cameroon. Using Claude.ai to generate R statistical code through structured prompt engineering, we systematically applied classical parametric approaches alongside non-parametric alternatives across five analytical stages: exploratory data analysis, bivariate correlation and regression, multivariate correlation matrix analysis, dimensionality reduction and clustering, and multiple linear regression. All disturbance count variables exhibited extreme positive skewness (1.776-8.367) and severe excess kurtosis (5.554-71.014), fundamentally violating parametric assumptions and designating non-parametric methods as co-primary analytical tools. The bivariate analysis revealed a strong positive association between snapping and gap size (Pearson $r = 0.865$, $p < 0.001$; $R^2 = 0.7483$), corroborated by non-parametric methods (Spearman $\rho = 0.455$, $p < 0.001$; Kendall $\tau = 0.366$, $p < 0.001$), indicating that species associated with larger canopy openings tend to record higher snapping frequencies. In contrast, multivariate analysis revealed a more differentiated disturbance structure: snapping and windthrow formed a significantly correlated coupled pair ($r = 0.659$, $p < 0.01$), while branch fall was essentially uncorrelated with both ($r = 0.013$, $p > 0.91$ in both cases), confirming that branch fall operates as a structurally independent disturbance mechanism. PCA identified two components meeting the Kaiser criterion: PC1 (eigenvalue = 1.671, 41.8% variance) loaded near-equally on snapping (0.702) and windthrow (0.700), representing a shared snapping–windthrow disturbance intensity axis, while PC2 (eigenvalue = 1.157, 28.9% variance) was dominated by branch fall (0.720) and dead standing (0.687), representing an independent crown and standing mortality axis. Dead standing was retained throughout all inferential analyses, including the correlation matrix, PCA, K-means clustering, and multiple regression, as it recorded one non-zero observation (*Zanthoxylum* sp.), with its near-zero variance flagged as a caveat on interpretation rather than grounds for exclusion. K-means clustering ($k = 3$) isolated two extreme outlier species, *Oubanguia alata* (snapping = 10, windthrow = 5) and *Protomegabaria stapfiana* (snapping = 11, windthrow = 6) as Cluster 2, a dead standing singleton (*Zanthoxylum* sp.) as Cluster 3, and a dominant low-disturbance majority ($n = 70$, 95.9%) as Cluster 1, yielding an overall silhouette score of 0.772. Multiple regression confirmed that windthrow was the sole significant predictor of snapping ($\beta = 1.332$, $p < 0.001$), with branch fall ($\beta = 0.023$, $p = 0.904$) and dead standing ($\beta = -0.662$, $p = 0.688$) contributing negligible independent variance, and the model explaining 43.5% of snapping variance ($R^2 = 0.4351$, $F(3, 69) = 17.71$, $p < 0.001$). Vibe and manual data analysis across 15 statistical outputs were compared, and the comparison revealed complete numerical equivalence to three or more decimal places, with the AI demonstrating autonomous assumption-aware method selection without explicit instruction. Vibe analysis completed all five analytical stages within nine structured prompts, substantially reducing time investment relative to conventional sequential coding. Mandatory validation through executed R code and analyst oversight remain essential to guard against large language model hallucination risks. This framework provides forestry researchers with accessible, validated tools for rigorous, reproducible bivariate and multivariate statistical analysis of ecological disturbance data characterised by non-normality and zero-inflation.

Keywords: Bivariate Analysis, Multivariate Analysis, Correlation, Regression, PCA, Cluster Analysis, K-means, Vibe Data Analysis, R Statistics, Artificial Intelligence, Ecological Disturbance Data.

How to Cite: Kato Samuel Namuene; Egbe Enow Andrew (2026) Bivariate and Multivariate Vibe Analysis of Forestry Data with AI and R Statistics. *International Journal of Innovative Science and Research Technology*, 11(4), 2798-2836. <https://doi.org/10.38124/ijisrt/26apr1365>

I. INTRODUCTION

Rigorous statistical methodology is the cornerstone of reproducible forestry research, and the choice of analytical approach fundamentally shapes what conclusions can legitimately be drawn from observational forestry data (Zuur et al., 2007; Legendre, 2019). While characterising individual variables through descriptive statistics provides an essential starting point, the explanatory and predictive demands of contemporary forestry research require methods capable of resolving relationships among variables simultaneously (ter Braak & Šmilauer, 2002; Borcard et al., 2018). Bivariate analyses reveal pairwise associations between forestry indicators, whereas multivariate techniques expose latent structure, co-disturbance gradients, and species-level clustering patterns that remain invisible to single-variable approaches (Everitt et al., 2011; Jolliffe and Cadima, 2016). The analytical progression from exploration to correlation, dimensionality reduction, and regression therefore represents a logical and necessary continuum in forestry research (McCune & Grace, 2002).

Forestry disturbance datasets present distinctive analytical challenges. Count variables such as snapping frequency, windthrow incidence, branch fall, and dead standing tree observations typically exhibit pronounced right skewness, excess kurtosis, zero-inflation, and meaningful outliers that collectively violate the normality and homoscedasticity assumptions underpinning classical parametric inference (Ver Hoef & Boveng, 2007; Zeileis et al., 2008). Pearson correlation and ordinary least squares regression remain widely used despite these violations, yet their sensitivity to non-normality can substantially inflate Type I error rates and distort coefficient estimates in count-data settings (O'Brien, 2007; Ives, 2015). Non-parametric rank-based alternatives such as Spearman's rho and Kendall's tau offer robustness to distributional departures, while generalised and permutation-based approaches extend inference to multivariate contexts without requiring normally distributed residuals (Mielke & Berry, 2001; Anderson, 2008; Bonnini et al., 2024). Selecting between these approaches requires empirical distributional assessment rather than ad hoc convention (Warton & Hui, 2011; Ives, 2015).

The integration of large language model (LLM) tools into quantitative forestry workflows represents a rapidly evolving frontier in data-intensive science (OECD, 2021; Hellas et al., 2023). AI-assisted statistical programming through platforms such as Claude.ai (Anthropic, 2026) enables structured prompt engineering to generate, validate, and iterate R code across complex multi-stage analytical pipelines, lowering the technical barrier for researchers without formal computational training (Ahuja et al., 2023; Weng et al., 2023). Vibe coding; the practice of directing an AI system through natural-language iterative prompts to produce executable statistical code, has been proposed as a complement to manual programming that accelerates analysis while maintaining scientific rigour when paired with

systematic validation (Kumar et al., 2024). By comparing AI-assisted vibe analysis with conventional manual analysis across all five analytical stages and 16 statistical outputs, the present study benchmarks the conditions under which AI-generated R code achieves numerical equivalence with human-executed analysis and identifies the oversight protocols necessary to ensure reproducibility and inferential integrity (Miller, 2025).

This paper analyses tree disturbance data from 183 treefall gaps in Korup National Park, Cameroon, examining snapping behaviour in relation to canopy gap size at the bivariate level, and subsequently exploring multivariate patterns among four disturbance response variables: snapping, windthrow, branch fall, and dead standing trees. A systematic five-stage analytical framework is applied: (1) Exploratory Data Analysis (EDA) using summary statistics, distributional diagnostics, scatter plots, and boxplots; (2) bivariate correlation and regression using Pearson product-moment correlation, Spearman rank correlation, Kendall's tau, simple linear regression, and chi-square contingency analysis; (3) multivariate Pearson correlation matrix analysis with Bonferroni-corrected significance testing and heatmap visualisation; (4) Principal Component Analysis (PCA) with Kaiser criterion retention and K-means clustering ($k=3$) validated by elbow method and silhouette scores; and (5) multiple linear regression modelling snapping as a function of the three remaining disturbance variables, complemented by MANOVA testing of multivariate disturbance vector differences across snapping categories.

➤ *Data Analysis in the Era of Artificial Intelligence (AI)*

The emergence of generative AI tools has introduced new capabilities into quantitative scientific research, enabling researchers to interact with statistical software through natural language rather than procedural code alone (Huang et al., 2023; Thirunavukarasu et al., 2023). Large language models (LLMs) such as Claude, Gemini, and GPT-4 can parse domain-specific research questions, select appropriate analytical strategies, and produce executable code for complex statistical workflows spanning bivariate correlation through multivariate dimensionality reduction (Nori et al., 2023; Singhal et al., 2023). In the context of forestry disturbance analytics, these capabilities address longstanding barriers to rigorous quantitative practice; the considerable programming expertise required to implement multivariate methods in R, the time cost of debugging iterative analytical pipelines, and the difficulty of correctly applying assumption-checking protocols to non-normal disturbance count data (Michener, 2015; Lai et al., 2019).

Beyond code generation, LLMs can support the full analytical lifecycle of a multivariate ecological study, structuring raw data frames, running distributional diagnostics, flagging constant variables that would cause model singularity, constructing Bonferroni-corrected correlation matrices, and selecting the appropriate number of principal components via the Kaiser criterion (Abdi &

Williams, 2010; Kassambara et al, 2026). AI-generated code can also automate the production of publication-quality diagnostics scree plots, biplots, silhouette analyses, and cluster visualisations, that would otherwise demand substantial programming investment (Friendly, 2002; Kassambara, 2017; Barke, 2023). However, meaningful use of AI-generated statistical code requires that researchers maintain domain expertise and critical oversight. Syntactically valid code may embed statistically inappropriate choices when distributional assumptions are violated, and automated outputs must be systematically validated before inferential conclusions are drawn (Floridi et al., 2020; Wamba et al., 2023).

- *Prompt Engineering Strategies for Statistical Computing*

The quality of AI-generated statistical code is determined primarily by the precision and completeness of the prompts provided to the language model (Marvin et al., 2024; Schulhoff et al., 2025). Effective prompts for multivariate forestry analysis should specify the research objective, describe the structure and distributional characteristics of all variables, state the desired analytical method and its non-parametric alternatives, identify variables requiring exclusion (such as zero-variance constants), and prescribe output format requirements including visualisation type, annotation content, and interpretation guidelines. For the five-stage framework applied in the present study, a representative prompt might read: ‘Act like R Statistics; for the data above, compute a Pearson correlation matrix for snapping, windthrow, and branch_fall using `psych::corr.test()` with Bonferroni correction. Exclude dead_standing (all zeros). Produce a `corrplot` heatmap with embedded `r` values and significance markers.’ Such specificity enables the LLM to produce contextually appropriate, assumption-aware code from a single iteration (Barke, 2023; White et al., 2023; Federiakin, 2024).

Prompt iteration is integral to vibe-based analytical workflows. Initial responses may require refinement when output formatting, variable naming conventions, or diagnostic detail do not meet the analyst’s requirements (Guo et al., 2023). Through successive exchanges, the analyst can extend initial code blocks to add Root Mean Square Error (RMSE) benchmarking, VIF multicollinearity checks, back-transformed cluster centroids, or MANOVA extensions without rewriting analytical foundations (Zamfirescu-Pereira et al., 2023). The conversational record of these exchanges constitutes a reproducible methodological audit trail that documents analytical choices and model refinement decisions in a form accessible to non-programmers (Raffin et al., 2021). Researchers must nonetheless validate all AI-generated outputs against statistical theory and executed R results, as LLMs may occasionally hallucinate package functions, mis-specify model arguments, or omit assumption checks for novel data structures (Ji et al., 2023).

- *Forestry Data Analytics with R Statistics*

R has consolidated its position as the primary analytical environment for quantitative forestry and forest science, supported by a continuously expanding ecosystem of domain-specific packages and a large, collaborative global user

community (Hesselbarth et al., 2019; Lai et al., 2019). Its open-source architecture and extensive package repository provide implementations of virtually every statistical method relevant to forestry research, from elementary descriptive statistics to penalised regression, mixed-effects models, spatial analysis, and community ordination (Forrester & Tang, 2016; Oksanen et al., 2026). For the five-stage analytical framework applied here, key packages include: `moments` for skewness and kurtosis diagnostics; `psych` for Bonferroni-corrected pairwise correlation testing; `corrplot` for heatmap visualisation of correlation matrices; `factoextra` and `cluster` for PCA visualisation and silhouette-validated K-means clustering (Kassambara et al., 2026); `car` for Variance Inflation Factors in multiple regression (Fox & Weisberg, 2019); and `ggplot2` for annotated scatter plots and diagnostic graphics throughout (Wickham, 2016).

The R platform for Bivariate and multivariate forestry analysis has matured considerably, with packages such as `vegan` (Oksanen et al., 2026) and `FactoMineR` (Lê et al., 2008) providing comprehensive ordination and community forestry tools that complement the base-R `prcomp()` and `kmeans()` functions used in the present study. Data preparation and reshaping are handled efficiently through the `tidyverse` suite, while the `car` package extends the base regression framework with diagnostics essential for assumption checking in multiple linear regression (Fox & Weisberg, 2019; Wickham et al., 2019). Despite these strengths, R’s steep learning curve remains a documented barrier for researchers without computational training, particularly for the more complex multivariate workflows that require coordinating multiple packages, data transformation sequences, and graphical outputs (Michener, 2015; Lai et al., 2019). The integration of AI-assisted prompt-driven code generation directly addresses this barrier by enabling researchers to specify analytical intent in natural language and receive executable, documented R code without requiring mastery of package-specific syntax (Federiakin, 2024; Schulhoff et al., 2025).

Although R offers exceptional analytical capabilities, it poses a considerable learning curve for researchers with little to no programming background, which may restrict access to its full potential among those who stand to gain the most from it (Baumer et al., 2014). Frequently encountered difficulties include grasping R’s syntax and data structures, navigating package installation and management, troubleshooting errors in code, and identifying suitable functions from an extensive and often overwhelming set of options (Touchon & McCoy, 2016). These obstacles tend to be especially acute for students and early-career researchers who have not received formal instruction in computational methods (Lai et al., 2019). That said, the growing availability of improved learning resources, intuitive interfaces like `RStudio`, and AI-powered programming assistants has begun to lower these barriers to entry (Baumer et al., 2014; Vaithilingam et al., 2022). The broader accessibility of R programming enabled by AI assistance presents a meaningful opportunity to strengthen the research community’s capacity for robust statistical analysis, while preserving the reproducibility and transparency that

have long distinguished R-based workflows (Hampton et al., 2017; Powers & Hampton, 2019).

➤ *Vibe Data Analysis: Methodological Frameworks for Modern Statistic*

In the natural sciences, statistical analysis is undergoing a structural transition as conversational AI tools alter the relationship between analyst intent and analytical execution. The practice of directing an LLM through iterative natural-language prompts to produce, validate, and refine executable statistical code has been termed ‘vibe coding’ (Kumar et al., 2024), and its application to data analysis ‘vibe data analysis’ extends this paradigm to the full empirical research pipeline, from data ingestion and exploratory visualisation through inferential testing, multivariate modelling, and results reporting. Rooted intellectually in Tukey’s (1977) tradition of exploratory data analysis, which prioritised pattern discovery and visual reasoning over confirmatory hypothesis testing, vibe data analysis similarly foregrounds iterative, question-driven inquiry while leveraging AI to remove syntactic and computational obstacles (Tukey, 1980; Unwin, 2015). The approach allows forestry researchers to direct analytical attention toward forestry interpretation, understanding which disturbance mechanisms co-occur, how species cluster by multivariate disturbance profiles, and which predictors explain snapping frequency variance, rather than toward the mechanics of R package syntax.

The methodological of vibe data analysis is defined by rapid prototyping, prompt-driven iteration, and the externalisation of procedural decisions into a transparent, reviewable conversational record (Raffin et al., 2021; Guo et al., 2023). Researchers can efficiently test multiple analytical strategies, switching between Pearson and Spearman correlations, comparing K-means solutions at different values of k, or adding MANOVA to an existing regression workflow through targeted prompt modifications that do not require rebuilding code from scratch. This flexibility is particularly valuable in the exploratory phases of multivariate disturbance analysis, where the appropriate number of PCA components, the ecological interpretability of cluster solutions, and the choice between parametric and non-parametric correlation methods are empirically determined rather than prescribed a priori (Abdi & Williams, 2010; Kassambara et al., 2026). AI also facilitates accessible documentation of analytical workflows: the sequence of prompts and responses constitutes a step-by-step methodological record that supports reproducibility and peer review without requiring supplementary code appendices (Wilson et al., 2017).

Vibe data analysis nonetheless introduces risks that must be managed through deliberate methodological discipline. The ease of iterative exploration may inadvertently encourage undisclosed multiple testing, post-hoc hypothesis reformulation, or over-reliance on visually appealing cluster solutions with poor silhouette validity (Simmons et al., 2011; Gelman & Loken, 2014). LLMs may also produce confident but incorrect responses including mis-specified function arguments, inappropriate method choices for zero-inflated count data, or overlooked assumption violations, requiring that all AI-generated code be executed, inspected, and

validated against known benchmarks before inferential use (Floridi et al., 2020; Ji et al., 2023). In the present study, these risks are mitigated by systematic cross-validation of all AI-generated outputs against manually executed R analyses across all 16 statistical outputs of the five-stage framework, and by explicit assumption checking at each analytical stage using normality tests, VIF diagnostics, silhouette scores, and visual residual inspection.

➤ *Research Gaps and Objectives*

Although substantial literature on multivariate methods in forestry exist (Borcard et al., 2018; Legendre, 2019) amidst growing application of AI tools in scientific computing (Ahuja et al., 2023; Hellas et al., 2023), a critical methodological gap remains. No existing framework provides assumption-aware, stage-by-stage analytical guidance for bivariate-to-multivariate disturbance analysis of non-normal forestry count data within a unified, AI-assisted and reproducible R workflow. Existing resources either treat parametric and non-parametric methods in isolation (Mielke & Berry, 2001; McCune & Grace, 2002), or deliver AI-generated code without structured validation protocols for the distributional challenges common to treefall gap data, including zero-inflation, extreme skewness, and ecologically meaningful outliers (Ver Hoef & Boveng, 2007). Researchers are consequently left without empirically grounded criteria for navigating method selection across the full analytical continuum from bivariate correlation through PCA, clustering, and multiple regression, when normality assumptions fail.

This gap is especially pronounced for datasets of the kind examined here, where four disturbance response variables measured across 183 treefall gaps exhibit simultaneous distributional violations that challenge both parametric inference and univariate diagnostic approaches. Addressing this gap, the present study develops and demonstrates an AI-assisted, five-stage analytical framework applied to two complementary datasets, a bivariate dataset (n = 183 treefall gaps; snapping count and gap size) and a multivariate dataset (n = 73 species; snapping, windthrow, branch fall, and dead standing) from Korup National Park. The specific objectives are to:

- Apply structured prompt engineering with Claude.ai (Anthropic, 2026) across nine sequential prompts to generate assumption-aware R code for five analytical stages.
- Compute and compare descriptive statistics, skewness, and kurtosis for all disturbance variables (windthrow, snapping, branch and dead standing).
- Conduct bivariate correlation analysis between snapping count and gap size using Pearson, Spearman, and Kendall methods in parallel.
- Construct a Pearson correlation matrix for all disturbance variables.
- Perform PCA on the standardised four-variable disturbance matrix.
- Fit a multiple linear regression model with gap size as the dependent variable and windthrow, snapping, branch fall and dead standing as predictors.

- Compare vibe analysis and manual analysis to assess its reliability and practicability.

➤ Hypotheses

The following hypotheses were advanced in correspondence with the seven objectives stated. Each hypothesis is stated in both null (H_0) and alternative (H_1) form, with the directional expectation grounded in the distributional properties of the data and the context of treefall gap dynamics in Korup National Park.

• Hypothesis 1: AI-Assisted Code Generation Across Five Analytical Stages

- ✓ H_0 : Structured prompt engineering with Claude.ai will fail to produce syntactically executable and analytically appropriate R code across all five stages of the analytical framework, requiring substantial manual correction to achieve valid statistical outputs.
- ✓ H_1 : Structured prompt engineering with Claude.ai will produce syntactically executable, assumption-aware R code across all five analytical stages within nine sequential prompts, with each prompt generating directly runnable code requiring no substantive manual correction.

• Hypothesis 2: Distributional Violations in Disturbance Variables

- ✓ H_0 : The disturbance variables, snapping, windthrow, branch fall, dead standing, and gap size will conform sufficiently to normality assumptions, exhibiting skewness and kurtosis values within acceptable parametric thresholds and showing no evidence of zero-inflation or near-zero variance that would necessitate non-parametric alternatives.
- ✓ H_1 : All disturbance count variables will exhibit significant departures from normality characterised by positive skewness exceeding conventional thresholds, zero-inflated frequency distributions, and near-zero variance in dead standing, collectively invalidating parametric assumptions and providing objective grounds for the parallel deployment of non-parametric methods at all subsequent analytical stages.

• Hypothesis 3: Bivariate Relationships Between Snapping and Gap Size

- ✓ $H_{0(a)}$: There is no statistically significant linear or monotonic association between snapping count and gap size across the 183 treefall gaps (Pearson $r = 0$; Spearman $\rho = 0$; Kendall $\tau = 0$).
- ✓ $H_{1(a)}$: Snapping count will exhibit a statistically significant positive association with gap size, such that larger canopy openings, reflecting greater wind exposure and altered mechanical stress regimes, are associated with higher snapping frequencies.
- ✓ $H_{0(b)}$: Gap size explains no meaningful proportion of variance in snapping count, with the Ordinary Least Squares (OLS) regression model yielding an R^2 statistically indistinguishable from zero and an Root Mean

Squared Error (RMSE) not meaningfully lower than the standard deviation of snapping.

- ✓ $H_{1(b)}$: Gap size will explain a statistically significant proportion of variance in snapping count, with the OLS model producing an $R^2 > 0$ and an RMSE substantially below the standard deviation of snapping, confirming the predictive utility of gap area as a mechanical stress proxy.
- ✓ $H_{0(c)}$: Snapping category and gap size category are statistically independent, with no association between disturbance severity class and canopy opening size class ($\chi^2 = 0$; Cramér's $V = 0$).
- ✓ $H_{1(c)}$: Snapping severity category will be significantly associated with gap size category, with higher snapping counts disproportionately occurring in larger gap size classes, yielding a statistically significant chi-square result and a Cramér's V effect size indicative of at least a weak association.

• Hypothesis 4: Multivariate Correlations Among Disturbance Variables

- ✓ H_0 : No statistically significant pairwise Pearson correlations will be detected among snapping, windthrow, and branch fall after Bonferroni correction, indicating that the three disturbance mechanisms operate independently across species in the Korup forest dynamics plot.
- ✓ H_1 : At least one statistically significant positive Pearson correlation will be identified among snapping, windthrow, and branch fall after Bonferroni correction, reflecting shared ecological drivers such as species-level mechanical vulnerability or exposure to high-energy disturbance events that produce co-occurrence of multiple damage types in the same species.

• Hypothesis 5: PCA Structure and K-Means Cluster Distinctiveness

- ✓ $H_{0(a)}$: PCA of the standardised disturbance matrix will yield no components satisfying the Kaiser criterion (eigenvalue ≥ 1) beyond the first, indicating that all disturbance variables load onto a single undifferentiated dimension and that no meaningful multivariate structure exists among the four disturbance types.
- ✓ $H_{1(a)}$: PCA will retain components satisfying the Kaiser criterion, with interpretable loading structures that differentiate a general disturbance intensity axis, consistent with distinct mechanisms of tree mortality in Korup.
- ✓ $H_{0(b)}$: K-means clustering with $k = 3$ will produce clusters of comparable centroid values across all disturbance variables, with low between-cluster separation, an overall silhouette width close to zero, and no distinguishable groupings of species by disturbance profile.
- ✓ $H_{1(b)}$: K-means clustering with $k = 3$ will yield interpretable clusters characterised by clearly differentiated centroid profiles including at least one small cluster of high-disturbance outlier species and an overall silhouette width sufficient to confirm non-random cluster structure in PCA space.

• *Hypothesis 6: Multiple Linear Regression of Snapping on Disturbance Predictors*

- ✓ H_0 : Windthrow and branch fall will jointly explain no statistically significant proportion of variance in snapping count, with the multiple regression model yielding a non-significant F-statistic, an R^2 indistinguishable from zero, an RMSE no lower than that of the simple regression model, and Variance Inflation Factor (VIF) values confirming the absence of multicollinearity.
- ✓ H_1 : Windthrow and branch fall will together explain a statistically significant and practically meaningful proportion of variance in snapping count, with R^2 and adjusted R^2 exceeding those of the simple bivariate model and RMSE reduced relative to the simple regression baseline, while VIF values confirm the absence of multicollinearity between predictors.

• *Hypothesis 7: Equivalence of Vibe Analysis and Manual Analysis*

- ✓ $H_{0(a)}$: AI-generated statistical outputs will differ systematically from manually computed values across comparison metrics, indicating that vibe analysis cannot be considered numerically equivalent to conventional analyst-coded approaches for non-normal ecological count data.
- ✓ $H_{1(a)}$: AI-generated and manually computed statistical outputs will achieve exact numerical agreement across comparison metrics, demonstrating that vibe analysis produces results numerically equivalent to those of manual analysis when AI-generated code is executed and verified in R.
- ✓ $H_{0(b)}$: Vibe analysis will differ systematically from manual analysis in assumption-checking behaviour and method selection of key analytical decision points, with the AI defaulting to parametric procedures without distributional assessment or failing to select non-parametric alternatives in the presence of documented normality violations.
- ✓ $H_{1(b)}$: Vibe analysis will converge with manual analysis on key analytical decision points, recognising non-normality, deploying non-parametric alternatives in parallel with parametric tests, arriving at the same optimal cluster number, demonstrating that structured prompting can replicate analyst-level methodological reasoning without explicit instruction at each decision stage.

➤ *Significance and Novelty*

This study makes a novel contribution to AI-assisted forestry analytics by embedding LLM capabilities within a five-stage statistical validation framework purpose-built for non-normal ecological disturbance count data. In contrast to existing studies that evaluate AI coding tools on generic programming benchmarks (Hellas et al., 2023; Kumar et al., 2024), the present work situates vibe data analysis within a domain-specific statistical context requiring empirical assumption checking, multi-method comparison, and ecologically interpretable outputs. Specific contributions include:

- Integrated five-stage analytical pipeline: Rather than addressing individual methods in isolation, the study provides a unified workflow spanning Exploratory Data Analysis (EDA), bivariate and multivariate correlation, PCA, K-means clustering, and multiple regression, demonstrating how analytical decisions at each stage propagate through the pipeline (McCune & Grace, 2002; Borcard et al., 2018).
- Empirical validation in a real disturbance dataset: Analysis of snapping, windthrow, branch fall, and dead standing tree frequencies across 183 Korup National Park treefall gaps caused by various tree species (Thomas et al., 2003) demonstrates the practical consequences of normality violations on inference across both bivariate and multivariate analytical stages.
- Methodological transparency through parallel comparisons: Side-by-side parametric and non-parametric analyses at each stage, including Pearson versus Spearman/Kendall correlations and OLS versus rank-based regression reveal how distributional violations alter inferential outcomes, supporting evidence-based method selection (O'Brien, 2007; Ives, 2015).
- Reproducible and stage-specific R code templates: All code is organised by analytical stage, structured as AI prompts with embedded validation checks, and documented to the standards of reproducible computational research, enabling direct reuse by forestry researchers (Peng, 2011; Wilson et al., 2017).
- Stage-specific decision frameworks: Explicit, empirically grounded criteria for method selection are provided at each stage, drawing on normality test outcomes, Kaiser criterion eigenvalues, silhouette coefficients, and variable data type, offering a transferable decision architecture for ecological count data (Kassambara, 2017).
- LLM vibe data analysis benchmarking: Systematic comparison of AI-assisted and manual analysis across 16 statistical outputs provides one of the first domain-specific benchmarks of LLM-generated code in ecological statistics, establishing conditions of numerical equivalence and identifying required validation protocols (Ahuja et al., 2023; Kumar et al., 2024).
- Educational and accessibility value: By documenting iterative prompt refinement across diverse statistical tasks, the study demonstrates how AI-assisted workflows can build statistical literacy in multivariate ecological methods among researchers lacking advanced computational training (OECD., 2021; Weng et al., 2023).

II. MATERIALS AND METHODS

In this part of the study, we present the data, software, and analytical framework used to investigate bivariate and multivariate relationships among disturbance observations recorded across 183 treefall gaps in Korup National Park. Five sequential stages guided the analytical process: exploratory data analysis, bivariate correlation and regression, multivariate correlation matrix construction, Principal Component Analysis and cluster analysis, and multiple linear regression. Analysis was conducted using R Statistical

Software with AI-assisted code generation via Claude.ai (Anthropic, 2026).

A. Data Description

Two datasets were analysed. Bivariate Dataset (n = 183): Contains species identity, snapping count (0-1), and gap size (1.0-4580.9 m²). Gap size represents canopy openings where trees are located, hypothesised to influence mechanical stress and snapping likelihood through altered wind exposure regimes.

Multivariate Dataset (n = 73): Contains species which were major causal factors of treefall gaps, four disturbance-related count variables per species: Snapping (0-11), Windthrow (trees uprooted by wind, 0-6), Branch fall (major branch loss events, 0-4), and Dead standing (standing dead trees, 0-1). These variables represent different manifestations of tree mortality and damage in the Korup forest dynamics plot (Thomas et al., 2003).

B. Statistical Software And Libraries

Data analysis was performed using R Statistical Software (R Core Team, 2026). The following R packages were utilised:

- Base R: Core statistical functions including descriptive statistics, Pearson, Spearman, and Kendall correlation, one-sample t-test, ANOVA, simple and multiple linear regression, K-means clustering, PCA via `prcomp()`, chi-square test with Cramér's V, confidence interval estimation, and contingency table construction.
- Moments: Computation of skewness and kurtosis for distributional shape assessment of count and continuous variables.
- `ggplot2`: Advanced data visualisation including annotated scatter plots with OLS regression lines and shaded 95% confidence bands.
- `GGally`: Extension of `ggplot2` for multivariate visualisation including scatter plot matrices via `ggpairs()` with customisable diagonal, upper, and lower panel functions.
- `Crayon`: Terminal string styling for colour-coded and formatted console output.
- `Ggstats`: Extension of `ggplot2` providing additional statistical geoms and summary functions including coefficient plots and enhanced statistical annotations.
- `Progress`: Construction and display of configurable terminal progress bars via `progress_bar$new()` with support for percentage completion, elapsed time, and estimated time remaining.
- `Corrplot`: Correlation heatmap visualisation with embedded coefficient values and a blue-white-red diverging colour scale.
- `Gridextra`: Arrangement of multiple `ggplot2` panels into a single composite figure via `grid.arrange()`, enabling the 2×2 layout of regression diagnostic plots with a shared overall title.
- `Psych`: Correlation matrix construction with simultaneous pairwise significance testing via `corr.test()`.

- `Car`: Companion to Applied Regression; Variance Inflation Factor computation for multicollinearity diagnostics in multiple linear regression.
- `Cluster` and `factoextra`: K-means clustering with elbow method and silhouette analysis; PCA visualisation including biplots, scree plots, and `fviz_cluster()` for cluster membership plots in PCA space.
- `MASS`: Multivariate Analysis of Variance (MANOVA) and robust statistical methods.

C. Analytical Framework

A systematic five stage data analysis framework was employed as follows:

➤ Stage 1: Exploratory Data Analysis (EDA)

- Data structure examination and summary statistics for all five variables.
- Calculation of measures of central tendency, dispersion, and shape for each variable.
- Frequency distribution analysis and identification of distributional anomalies.
- Scatter plots with regression lines, box plots with individual data points.

➤ Stage 2: Bivariate Correlation and Regression

- Pearson product-moment correlation between snapping and gap size.
- Spearman rank correlation and Kendall's tau as non-parametric alternatives.
- Simple linear regression with RMSE, R², and F-test diagnostics.
- Contingency table analysis with chi-square and Cramér's V effect size.

➤ Stage 3: Multivariate Correlation Matrix

- Pairwise Pearson correlations among all four disturbance variables.
- Significance testing via `psych::corr.test()` with Bonferroni correction.
- Professional correlation heatmap using `corrplot`.

➤ Stage 4: PCA and Cluster Analysis

- PCA with variable standardisation (`scale. = TRUE`), Kaiser criterion retention.
- Component loadings, scree plot, and biplot visualisation.
- K-means clustering (k = 3) with elbow method and silhouette score validation.
- Cluster visualisation in PCA space.

➤ Stage 5: Multiple Linear Regression

- Snapping modelled as a function of windthrow, branch fall, and dead standing.
- Model performance metrics including R², adjusted R², RMSE, and F-test.


```
195, 83.3, 634.1, 4134.8, 96.7, 50, 107.5, 397, 87.5, 135.7,
413.3, 464.7, 201.8, 62.5, 171.5, 642.5, 75, 105.2, 137.5, 325,
333.7, 125, 516.3, 154, 261.9, 515.6)
```

```
# Create data frames
```

```
bivar_data <- data.frame(Species = Species, Snapping =
snapping, Gap_Size = gap_size)
```

```
multi_data <- data.frame(Species = species, Snapping =
snapping, Windthrow = windthrow, Branch_fall =
branch_fall, Dead_standing = dead_standing)
```

```
multi_data <- data.frame(species=species,
Snapping=snapping, Windthrow=windthrow,
Branch_fall=branch_fall, Dead_standing=dead_standing)
```

```
cat('Data loaded successfully!\n')
```

```
cat('Number of species:', nrow(bivar_data), '\n')cat('Data
loaded successfully!\n')
```

```
cat('Number of species:', nrow(bivar_data), '\n')
```

➤ Stage 1 Bivariate Descriptive Statistics

• AI Prompt 2

Act like R Statistics; for the bivariate dataset (snapping and gap_size, n = 183 treefall gaps), compute descriptive statistics and produce annotated frequency histograms with fitted normal curves, following the points below:

- ✓ Compute mean, median, SD, min, and max for snapping and gap_size and present results in a formatted summary table.
- ✓ Calculate skewness and kurtosis for both variables using the moments package to assess distributional shape.
- ✓ Produce a frequency histogram of snapping counts with a superimposed normal curve parameterised by the sample mean and SD; annotate with skewness and kurtosis values.
- ✓ Produce a frequency histogram of gap size (m²) with a superimposed normal curve; annotate with skewness value.

Response

```
# =====
# STAGE 1: BIVARIATE DESCRIPTIVE STATISTICS
# =====
library(moments)
```

```
# Summary statistics table
```

```
biv_stats <- data.frame(
Variable = c('Snapping (count)', 'Gap Size (m2)'),
Mean = c(round(mean(snapping),3),
round(mean(gap_size),3)),
Median = c(median(snapping), median(gap_size)),
SD = c(round(sd(snapping),3), round(sd(gap_size),3)),
Min = c(min(snapping), min(gap_size)),
Max = c(max(snapping), max(gap_size))
)
print(biv_stats)
```

```
# Skewness and kurtosis
```

```
cat('Snapping -- Skewness:', round(skewness(snapping),3),
' Kurtosis:', round(kurtosis(snapping),3), '\n')
cat('Gap Size -- Skewness:', round(skewness(gap_size),3),
' Kurtosis:', round(kurtosis(gap_size),3), '\n')
```

```
# Histogram: snapping counts with normal curve
```

```
h <- hist(snapping,
breaks = seq(-0.5, 1.5, 0.5), # smaller bin width
freq = TRUE,
col = 'steelblue',
border = 'white',
main = 'Frequency Histogram of Snapping Counts',
xlab = 'Snapping Count',
ylab = 'Frequency')
```

```
# Parameters for binomial
```

```
n <- length(snapping)
p <- mean(snapping)
```

```
# Expected frequencies
```

```
x_vals <- 0:1
expected <- dbinom(x_vals, size = 1, prob = p) * n
```

```
# Add expected points and line
```

```
points(x_vals, expected, col = 'red', pch = 19, cex = 1.5)
lines(x_vals, expected, col = 'red', lwd = 2)
```

```
# Add counts to bars
```

```
text(x = h$mids,
y = h$counts,
labels = h$counts,
pos = 3,
cex = 1,
col = 'black')
```

```
# Histogram: gap size with normal curve
```

```
hist(gap_size, breaks=20, col='steelblue', border='white',
main='Frequency Histogram of Gap Size (m2)',
xlab='Gap Size (m2)', ylab='Frequency')
x_seq2 <- seq(min(gap_size), max(gap_size), length=200)
norm_y2 <- dnorm(x_seq2, mean=mean(gap_size),
sd=sd(gap_size)) * length(gap_size) *
diff(hist(gap_size, breaks=20, plot=FALSE)$breaks)[1]
lines(x_seq2, norm_y2, col='red', lwd=2)
legend('topright',
legend=paste('Skewness =', round(skewness(gap_size),2)),
bty='n', cex=0.85)
```

➤ Stage 1 (Continued) Multivariate Descriptive Statistics

• AI Prompt 3

Compute full descriptive statistics for all four disturbance variables (snapping, windthrow, branch_fall, dead_standing) and produce a scatter plot matrix using ggplot2, crayon, ggstats, ggally comments and progress packages, following the points below:

- ✓ Compute mean, median, SD, min, max, skewness, and kurtosis for each variable; present results in a formatted table (Table 2 format).
- ✓ Note that dead_standing is a constant (all zeros) and flag it accordingly in the output.
- ✓ Produce a scatter plot matrix of all four variables using pairs(), with semi-transparent points to reveal overplotting in zero-inflated distributions.

Response

```
# =====
# STAGE 1 (continued): MULTIVARIATE DESCRIPTIVE
# STATISTICS
# =====
```

```
vars_all <- list(Snapping = snapping,
Windthrow = windthrow,
Branch_fall = branch_fall,
Dead_standing = dead_standing)

# Build summary table
multi_stats <- do.call(rbind, lapply(names(vars_all),
function(v) {
x <- vars_all[[v]]
sk <- ifelse(sd(x)==0, NA, round(skewness(x),3))
ku <- ifelse(sd(x)==0, NA, round(kurtosis(x),3))
data.frame(Variable=v, Mean=round(mean(x),3),
Median=median(x),
SD=round(sd(x),3), Min=min(x), Max=max(x),
Skewness=sk, Kurtosis=ku)
}))
print(multi_stats)
```

```
# Flag dead_standing
cat("\nNote: Dead_standing is a constant (all values = 0);'
'excluded from inferential analyses.\n')
```

```
# =====
# CUSTOM FUNCTIONS FOR EACH PANEL TYPE
# =====
```

```
# --- Diagonal: histogram with shaded background ---
diag_hist <- function(data, mapping, ...) {
ggplot(data, mapping) +
theme_bw(base_size = 11) +
theme(
panel.background = element_rect(fill = "#EAECE5", colour
= NA),
panel.grid.major = element_line(colour = "white", linetype
= "dashed", linewidth = 0.4),
panel.grid.minor = element_blank(),
axis.text = element_text(size = 8),
axis.ticks = element_line(linewidth = 0.3),
plot.background = element_rect(fill = "white", colour = NA)
) +
geom_histogram(
fill = "#3A5A8C",
colour = "#2A4070",
binwidth = 1,
boundary = 0
)
}
```

```
# --- Off-diagonal: scatter plot with white background ---
off_diag_scatter <- function(data, mapping, ...) {
ggplot(data, mapping) +
theme_bw(base_size = 11) +
theme(
```

```
panel.background = element_rect(fill = "white", colour =
NA),
panel.grid.major = element_line(colour = "#CCCCCC",
linetype = "dashed", linewidth = 0.35),
panel.grid.minor = element_blank(),
axis.text = element_text(size = 8),
axis.ticks = element_line(linewidth = 0.3),
plot.background = element_rect(fill = "white", colour = NA)
) +
geom_point(
colour = "#6A8FC0",
alpha = 0.65,
size = 1.8,
shape = 16
)
}
```

```
# =====
# BUILD THE MATRIX
# =====
```

```
p <- ggpairs(
multi_data[ , c("Snapping", "Windthrow", "Branch_fall",
"Dead_standing")], # exclude Species
upper = list(continuous = off_diag_scatter),
lower = list(continuous = off_diag_scatter),
diag = list(continuous = diag_hist),
columnLabels = c("Snapping", "Windthrow", "Branch Fall",
"Dead Standing"),
switch = "both"
) +
labs(title = "Scatter Plot Matrix \u2014 All Four
Disturbance Variables") +
theme(
strip.background = element_rect(fill = "white", colour =
"white"),
strip.text = element_text(size = 11, colour = "black",
face = "bold"),
plot.title = element_text(size = 14, face = "bold", hjust =
0.5,
margin = margin(b = 10)),
plot.background = element_rect(fill = "white", colour =
NA),
panel.spacing = unit(0.3, "lines")
)
```

```
# =====
# SAVE THE GRAPH
# =====
```

```
ggsave(
"~/Desktop/Figure3_ScatterPlotMatrix.png",
plot = p,
width = 10,
height = 10,
dpi = 300,
bg = "white"
)
```

```
cat("Figure saved as Figure3_ScatterPlotMatrix.png\n")
```

➤ Stage 2: Bivariate Correlation Analysis

• AI Prompt 4

Conduct bivariate correlation analysis between snapping and gap_size using three methods, and produce an annotated scatter plot, following the points below:

- ✓ Pearson product-moment correlation with 95% confidence interval and p-value.
- ✓ Spearman rank correlation and Kendall's tau as non-parametric alternatives, with p-values.
- ✓ Present all three results in a formatted summary table (Table 3 format) with interpretation column.
- ✓ Scatter plot of snapping vs. gap_size with OLS regression line (red), 95% confidence band (shaded), and embedded annotation of Pearson r, p-value, and R².

Response

```
# =====
# STAGE 2: BIVARIATE CORRELATION ANALYSIS
# =====
snapping <- bivar_data$Snapping
gap_size <- bivar_data$Gap_Size

# Three correlation methods
pearson_r <- cor.test(snapping, gap_size,
method='pearson')
spearman_r <- cor.test(snapping, gap_size,
method='spearman', exact = FALSE)
kendall_t <- cor.test(snapping, gap_size, method='kendall',
exact = FALSE)

# Summary table
cor_summary <- data.frame(
Method = c('Pearson r', 'Spearman rho', 'Kendall tau'),
Symbol = c('r', 'rho', 'tau'),
Value = round(c(pearson_r$estimate,
spearman_r$estimate,
kendall_t$estimate), 3),
P_value = round(c(pearson_r$p.value,
spearman_r$p.value,
kendall_t$p.value), 3),
CI_95 = c(paste0('[', round(pearson_r$conf.int[1],3),
', ', round(pearson_r$conf.int[2],3), ']',
'—', '—'),
Interpretation = c('Not significant', 'Not significant', 'Not
significant')
)
print(cor_summary)

# R-squared
r2 <- round(pearson_r$estimate^2, 4)
cat('R2 (Pearson):', r2, '\n')

# Annotated scatter plot
library(ggplot2)
ggplot(bivar_data, aes(x=Gap_Size, y=Snapping)) +
geom_point(alpha=0.6, colour='steelblue', size=2) +
geom_smooth(method='lm', se=TRUE, colour='red',
linewidth=1.2,
```

```
fill='pink', alpha=0.3) +
annotate('text', x=max(bivar_data$Gap_Size)*0.55,
y=max(bivar_data$Snapping)*0.9,
label=paste0('Pearson r = ', round(pearson_r$estimate,3),
'\np = ', round(pearson_r$p.value,3),
'\nR2 = ', r2),
size=3.5, hjust=0) +
labs(title='Snapping vs Gap Size',
x='Gap Size (m2)', y='Snapping Count') +
theme_minimal()
```

➤ Stage 2 (continued): Simple Linear Regression

• AI Prompt 5

Perform simple linear regression of snapping on gap_size and produce all diagnostic outputs, following the points below:

- ✓ Fit OLS regression model; print full summary including intercept, slope, SE, t-value, p-value, R², adjusted R², and F-statistic.
- ✓ Print ANOVA table and calculate RMSE; compare RMSE to the SD of snapping to assess predictive utility.
- ✓ Produce the four standard regression diagnostic plots (Residuals vs Fitted, Q-Q, Scale-Location, Residuals vs Leverage) and label influential cases by row number.
- ✓ Contingency table analysis: categorise snapping as Low (0-1), Medium (2), High (≥3) and gap_size as Small (<200 m²), Medium (200-1000 m²), Large (>1000 m²); run chi-square test and compute Cramér's V.

Response

```
# =====
# STAGE 2 (continued): SIMPLE LINEAR REGRESSION
# =====
library(ggplot2)
library(gridExtra)

lm_model <- lm(Snapping ~ Gap_Size, data = bivar_data)
spearman_r <- cor.test(snapping, gap_size,
method='spearman', exact=FALSE)
kendall_t <- cor.test(snapping, gap_size, method='kendall',
exact=FALSE)
spearman_r
kendall_t
summary(lm_model)
anova(lm_model)

# RMSE vs SD of snapping
rmse_slr <- sqrt(mean(lm_model$residuals^2))
cat('RMSE:', round(rmse_slr,4),
'/ SD(Snapping):', round(sd(snapping),4), '\n')
cat('Negative adjusted R2 confirms model < intercept-only
null.\n')

# =====
# CONTINGENCY TABLE: Snapping Category × Gap Size
Category
# =====
```

```
# Categorise snapping: Low (0-1), Medium (2), High ( $\geq 3$ )
bivar_data$snap_cat <- cut(bivar_data$Snapping,
breaks = c(-Inf, 1, 2, Inf),
labels = c("Low (0-1)", "Medium (2)", "High ( $\geq 3$ )"),
right = TRUE)
```

```
# Categorise gap size using Brokaw (1985) thresholds:
# Small < 100 m2, Medium 100-400 m2, Large > 400 m2
bivar_data$gap_cat <- cut(bivar_data$Gap_Size,
breaks = c(0, 100, 400, Inf),
labels = c("Small (<100)", "Medium (100-400)", "Large
(>400)"),
right = FALSE)
```

```
# Build contingency table
ct <- table(Snapping = bivar_data$snap_cat,
Gap_Size = bivar_data$gap_cat)
print(ct)
print(addmargins(ct))
```

```
# Drop empty columns before testing (Large = 0 for all rows
in this dataset)
ct_reduced <- ct[, colSums(ct) > 0]
cat("Reduced table (non-empty columns only):\n")
print(addmargins(ct_reduced))
```

```
# Chi-square test (no Yates correction for tables larger than
2x2)
chi_res <- chisq.test(ct_reduced, correct = FALSE)
print(chi_res)
```

```
# Cramér's V
cramer_v <- sqrt(chi_res$statistic / (sum(ct_reduced) *
(min(dim(ct_reduced)) - 1)))
cat("Cramér's V =", round(cramer_v, 3), "\n")
```

```
# Expected frequencies (to check assumption: all cells  $\geq 5$ )
cat("\nExpected frequencies:\n")
print(round(chi_res$expected, 2))
```

```
# Print full contingency table with margins
print(addmargins(ct))
```

```
# Chi-square test results
print(chi_res)
```

```
# Cramér's V
cat("Cramér's V =", round(cramer_v, 3), "\n")
```

```
# --- Extract diagnostics ---
diag_df <- data.frame(
idx = 1:nrow(bivar_data),
fitted = fitted(lm_model),
residuals = residuals(lm_model),
std_resid = rstandard(lm_model),
sqrt_std = sqrt(abs(rstandard(lm_model))),
leverage = hatvalues(lm_model),
cooks = cooks.distance(lm_model)
)
```

```
# Top 3 influential points per panel
```

```
top3_resid <- diag_df[order(abs(diag_df$residuals),
decreasing=TRUE)[1:3], ]
top3_scale <- diag_df[order(diag_df$sqrt_std,
decreasing=TRUE)[1:3], ]
top3_lev <- diag_df[order(diag_df$leverage,
decreasing=TRUE)[1:3], ]
```

```
qq_vals <- qqnorm(diag_df$std_resid, plot.it=FALSE)
qq_df <- data.frame(theoretical = qq_vals$x,
std_resid = qq_vals$y,
idx = diag_df$idx)
top3_qq <- qq_df[order(abs(qq_df$std_resid),
decreasing=TRUE)[1:3], ]
```

```
# --- Shared white theme ---
```

```
white_theme <- theme_minimal(base_size = 11) +
theme(
plot.background = element_rect(fill = "white", colour =
NA),
panel.background = element_rect(fill = "white", colour =
NA),
panel.grid.major = element_line(colour = "grey88",
linewidth = 0.4),
panel.grid.minor = element_line(colour = "grey93",
linewidth = 0.2),
axis.text = element_text(colour = "grey30", size = 9),
axis.title = element_text(colour = "grey20", size = 9),
plot.title = element_text(colour = "black", face = "bold",
size = 10, hjust = 0.5),
margin = margin(b = 6),
plot.margin = margin(8, 8, 8, 8)
)
```

```
pt_col <- "#6fa8d4" # steelblue points
line_col <- "#cc3333" # red lines
```

```
# -----
# Panel 1: Residuals vs Fitted
# -----
```

```
p1 <- ggplot(diag_df, aes(x = fitted, y = residuals)) +
geom_hline(yintercept = 0, colour = line_col, linewidth =
0.8) +
geom_point(colour = pt_col, alpha = 0.7, size = 1.8) +
geom_smooth(method = "loess", se = FALSE, colour =
line_col, linewidth = 1) +
geom_text(data = top3_resid, aes(label = idx),
colour = "grey20", size = 3, vjust = -0.6) +
labs(title = "Residuals vs Fitted",
x = "Fitted Values", y = "Residuals") +
white_theme
```

```
# -----
# Panel 2: Q-Q Residuals
# -----
```

```
p2 <- ggplot(qq_df, aes(x = theoretical, y = std_resid)) +
geom_abline(slope = 1, intercept = 0, colour = line_col,
linewidth = 1) +
geom_point(colour = pt_col, alpha = 0.7, size = 1.8) +
geom_text(data = top3_qq, aes(label = idx),
colour = "grey20", size = 3, vjust = -0.6) +
labs(title = "Q-Q Residuals",
```

```
x = "Theoretical Quantiles", y = "Standardised Residuals")
+
white_theme
```

```
# -----
# Panel 3: Scale-Location
# -----
p3 <- ggplot(diag_df, aes(x = fitted, y = sqrt_std)) +
  geom_point(colour = pt_col, alpha = 0.7, size = 1.8) +
  geom_smooth(method = "loess", se = FALSE, colour =
line_col, linewidth = 1) +
  geom_text(data = top3_scale, aes(label = idx),
colour = "grey20", size = 3, vjust = -0.6) +
  labs(title = "Scale-Location",
x = "Fitted Values", y = "\u221a|Standardised Residuals|")
+
white_theme
```

```
# -----
# Panel 4: Residuals vs Leverage + Cook's contours
# -----
p_hat <- seq(min(diag_df$leverage) * 0.5,
max(diag_df$leverage) * 1.1,
length.out = 200)
p_coef <- length(coef(lm_model))
```

```
cook_df <- do.call(rbind, lapply(c(0.5, 1), function(D) {
y_pos <- sqrt(D * p_coef * (1 - p_hat) / p_hat)
data.frame(leverage = c(p_hat, p_hat),
std_resid = c(y_pos, -y_pos),
D = as.character(D))
}))
```

```
p4 <- ggplot(diag_df, aes(x = leverage, y = std_resid)) +
  geom_hline(yintercept = 0, colour = "grey60", linewidth =
0.5,
linetype = "dashed") +
  geom_line(data = cook_df,
aes(group = interaction(D, std_resid > 0)),
colour = "#e08080", linetype = "dashed", linewidth = 0.7) +
  geom_point(colour = pt_col, alpha = 0.7, size = 1.8) +
  geom_text(data = top3_lev, aes(label = idx),
colour = "grey20", size = 3, vjust = -0.6) +
  labs(title = "Residuals vs Leverage",
x = "Leverage", y = "Standardised Residuals") +
white_theme
```

```
# -----
# Combine into 2x2 grid with overall title
# -----
grid.arrange(
p1, p2, p3, p4, ncol = 2,
top = grid::textGrob(
"Regression Diagnostic Plots \u2014 Snapping ~ Gap Size",
gp = grid::gpar(col = "black", fontface = "bold", fontsize =
13)
)
)
```

➤ Stage 3: Multivariate Correlation Matrix

• AI Prompt 6

Create a Pearson correlation matrix and heatmap for the three non-constant disturbance variables (snapping, windthrow, branch_fall), following the points below:

- ✓ Exclude dead_standing from all analyses due to zero variance; add an explanatory note.
- ✓ Compute the full 3x3 Pearson correlation matrix and pairwise significance tests using the psych package; report r and p-values for all pairs.
- ✓ Identify and print all significant correlations at $\alpha = 0.05$ and $\alpha = 0.01$, with direction and magnitude.
- ✓ Produce a professional correlation heatmap using corrplot with a blue-white-red colour scale, embedded coefficient values, and significance markers.

Response

```
# =====
# STAGE 3: MULTIVARIATE CORRELATION MATRIX
# =====
library(corrplot)
library(psych)
```

```
# Check variance of each variable before deciding what to
include
cat("--- Variable variances ---\n")
print(apply(multi_data[,
c('Snapping', 'Windthrow', 'Branch_fall', 'Dead_standing')], 2,
var))
```

```
# dead_standing has near-zero variance (only 1 non-zero
value out of 73),
# so it is still unsuitable for Pearson correlation — keep
exclusion
vars3 <- multi_data[, c('Snapping', 'Windthrow',
'Branch_fall', 'Dead_standing')]
```

```
cat("\nNote: Dead_standing has variance =',
var(dead_standing),
'(only 1 event in 73 species). Pearson correlation will be
unreliable for it.\n',
'It is included here for completeness but interpret its
correlations with caution.\n\n')
```

```
# Pearson correlation matrix
cor_mat <- cor(vars3, method = 'pearson')
cat("--- Pearson Correlation Matrix ---\n")
print(round(cor_mat, 4))
```

```
# Pairwise significance
cor_res <- corr.test(vars3, method = 'pearson')
cat("\n--- p-values ---\n")
print(round(cor_res$p, 4))
print(cor_res, short = FALSE)
```

```
# Report significant pairs
cat("\n--- Significant correlations (alpha = 0.05) ---\n")
p_mat <- cor_res$p
```

```

found_any <- FALSE
for (i in 1:(ncol(p_mat) - 1)) {
  for (j in (i + 1):ncol(p_mat)) {
    if (p_mat[i, j] < 0.05) {
      found_any <- TRUE
      sig <- ifelse(p_mat[i, j] < 0.01, '**(p < 0.01)', *(p < 0.05)')
      cat(rownames(p_mat)[i], 'vs', colnames(p_mat)[j],
        '| r =', round(cor_mat[i, j], 3), sig, '\n')
    }
  }
}
if (!found_any) cat('No significant correlations found at
alpha = 0.05.\n')

# Heatmap
corrplot(cor_mat, method = 'color', type = 'upper',
addCoef.col = 'black', number.cex = 0.9,
tl.col = 'black', tl.cex = 1.0,
col = colorRampPalette(c('#2166AC', 'white',
'#D6604D'))(200),
title = 'Pearson Correlation Matrix — Disturbance
Variables',
mar = c(0, 0, 2, 0))

```

➤ Stage 4: Principal Component and Cluster Analysis

• AI Prompt 7

Perform Principal Component Analysis on the standardised three-variable disturbance matrix (snapping, windthrow, branch_fall) and interpret the component structure, following the points below, using factoextra and ggplot2 for graphs:

- ✓ Standardise the three variables and run `prcomp()`; print eigenvalues, percentage variance explained, and cumulative variance for all three components.
- ✓ Print the full component loadings matrix and provide ecological interpretation of PC1 (snapping-branch fall axis), PC2 (windthrow axis), and PC3 (contrast axis).
- ✓ Produce a scree plot with the Kaiser criterion line (eigenvalue = 1, red dashed) and a legend; annotate which components are retained.
- ✓ Produce a biplot displaying species scores (blue) and variable loading vectors (red); use `repel` to avoid label overlap for the outlier species.

Response

```

# =====
# STAGE 4: PRINCIPAL COMPONENT ANALYSIS
# =====
library(factoextra)
library(ggplot2)

# NOTE: Dead_standing IS included — it has 1 non-zero
value (Zanthoxylum sp.)
# However, because it has near-zero variance, prcomp with
scale.=TRUE will
# handle it; we flag it but keep it in.
vars3 <- multi_data[, c('Snapping', 'Windthrow',
'Branch_fall', 'Dead_standing')]

```

```

# Scale variables (use scale. = TRUE inside prcomp directly
— avoids manual scaling)
pca_res <- prcomp(vars3, scale. = TRUE)

```

```
n_pc <- ncol(pca_res$rotation) # = 4, one per variable
```

```
# Eigenvalues and variance explained
```

```
eig <- pca_res$sdev^2
pct <- round(eig / sum(eig) * 100, 1)
cum_pct <- round(cumsum(pct), 1)
```

```
eig_table <- data.frame(
Component = paste0('PC', 1:n_pc),
Eigenvalue = round(eig, 3),
Variance_pct = pct,
Cumulative_pct = cum_pct
)
cat('--- Eigenvalues and Variance Explained ---\n')
print(eig_table)
```

```
# Component loadings
```

```
cat('\n--- Component Loadings ---\n')
print(round(pca_res$rotation, 4))
```

```
# Kaiser criterion summary
```

```
retained <- which(eig >= 1)
cat('\nComponents retained (Kaiser criterion, eigenvalue >=
1):', paste0('PC', retained), '\n')
for (i in 1:n_pc) {
cat(sprintf('PC%d (%.1f%%) — Eigenvalue: %.3f\n', i,
pct[i], eig[i]))
}

```

```
# -----
```

```
# Scree Plot
```

```
# -----
```

```
eig_data <- data.frame(
PC = factor(paste0("PC", 1:n_pc), levels =
paste0("PC", 1:n_pc)),
Eigenvalue = eig
)

```

```
ggplot(eig_data, aes(x = PC, y = Eigenvalue, group = 1)) +
theme_bw() +
theme(
text = element_text(family = "serif"),
plot.title = element_text(hjust = 0.5, face = "bold", size =
22),
axis.title = element_text(size = 18),
axis.text = element_text(size = 16, color = "black"),
legend.text = element_text(size = 14),
panel.grid.major = element_line(color = "#ebeb", linetype
= "dashed", linewidth = 0.8),
panel.grid.minor = element_blank(),
panel.border = element_blank(),
axis.line = element_line(color = "black", linewidth = 1),
legend.position = c(0.75, 0.85),
legend.background = element_rect(fill = "white", color =
"lightgrey")
) +

```

```
geom_hline(aes(yintercept = 1, linetype = "Kaiser criterion
(eigenvalue = 1)"),
color = "#c04c3f", linewidth = 1.2) +
geom_line(color = "#1f4172", linewidth = 1.5) +
geom_point(color = "#1f4172", size = 6) +
geom_text(aes(label = sprintf("%3f", Eigenvalue)),
vjust = -1.5, size = 6, family = "serif", fontface = "bold") +
labs(
title = "Scree Plot — Disturbance PCA",
x = "Principal Component",
y = "Eigenvalue",
linetype = ""
) +
scale_linetype_manual(values = "dashed") +
coord_cartesian(ylim = c(min(eig) * 0.8, max(eig) * 1.2))

# -----
# Biplot
# -----
fviz_pca_biplot(pca_res,
col.ind = "#7497ba",
col.var = "#c04c3f",
label = "var",
repel = TRUE,
pointsize = 3,
labelsize = 7) +
theme_bw() +
theme(
text = element_text(family = "serif"),
plot.title = element_text(hjust = 0.5, face = "bold", size =
22),
axis.title = element_text(size = 18),
axis.text = element_text(size = 16, color = "black"),
panel.grid.major = element_line(color = "#ebeb3f", linetype =
"dashed", linewidth = 0.8),
panel.grid.minor = element_blank(),
panel.border = element_blank(),
axis.line = element_line(color = "black", linewidth = 1)
) +
geom_hline(yintercept = 0, linetype = "dashed", color =
"black", linewidth = 0.8) +
geom_vline(xintercept = 0, linetype = "dashed", color =
"black", linewidth = 0.8) +
labs(
title = "PCA Biplot — 73 Species x 4 Disturbance
Variables",
x = "PC1 (Disturbance Intensity)",
y = "PC2 (Windthrow vs. Others)"
)

```

➤ Stage 4 (continued): K-Means Cluster Analysis

• AI Prompt 8

Perform K-means cluster analysis on the standardised three-variable disturbance matrix, select $k = 3$ for ecological interpretability, and fully characterise each cluster, following the points below:

- ✓ Apply the elbow method ($k = 1$ to 10) and silhouette analysis ($k = 2$ to 10) to identify the statistically optimal number of clusters; plot both diagnostics.

- ✓ Fit K-means with $k = 3$ and $nstart = 25$; print cluster sizes, within-cluster SS, total WSS, and between-cluster SS.
- ✓ Print cluster centroids back-transformed to original variable units; compute and print the per-cluster and overall silhouette widths.
- ✓ Visualise cluster membership in PCA space using `fviz_cluster()`; label the species in Clusters 1 and 2 (outlier groups) explicitly.

Response

```
# =====
# STAGE 4 (CONTINUED): K-MEANS CLUSTER
ANALYSIS
# =====
library(cluster)
library(factoextra)
library(ggplot2)
library(ggrepel)

# All four disturbance variables (Dead_standing included)
vars3 <- multi_data[, c('Snapping', 'Windthrow',
'Branch_fall', 'Dead_standing')]

# Standardise — same matrix used for both PCA and K-
means
vars_sc <- scale(vars3)

# Also run PCA here so pct is available for axis labels in
Figure 11
pca_res <- prcomp(vars_sc, scale. = FALSE)
eig <- pca_res$sdev^2
pct <- round(eig / sum(eig) * 100, 1)

# -----
# ELBOW METHOD (Figure 9)
# -----
set.seed(123)
wss <- sapply(1:10, function(k) {
kmeans(vars_sc, centers = k, nstart = 25, iter.max =
100)$tot.withinss
})
elbow_data <- data.frame(k = 1:10, WSS = wss)

ggplot(elbow_data, aes(x = k, y = WSS)) +
theme_bw() +
theme(
text = element_text(family = "serif"),
plot.title = element_text(hjust = 0.5, face = "bold", size =
22),
axis.title = element_text(size = 18),
axis.text = element_text(size = 16, color = "black"),
legend.text = element_text(size = 14),
panel.grid.major = element_line(color = "#ebeb3f", linetype =
"dashed", linewidth = 0.8),
panel.grid.minor = element_blank(),
panel.border = element_rect(color = "black", fill = NA),
legend.position = c(0.78, 0.82),
legend.background = element_rect(fill = "white", color =
"lightgrey")
)

```

```

) +
geom_line(color = "#1f4172", linewidth = 1.5) +
geom_point(color = "#1f4172", size = 5) +
geom_vline(aes(xintercept = 2, linetype = "Elbow at k = 2"),
color = "#c04c3f", linewidth = 1.1) +
geom_vline(aes(xintercept = 3, linetype = "Selected k = 3"),
color = "#3a9e5f", linewidth = 1.1) +
scale_linetype_manual(
name = NULL,
values = c("Elbow at k = 2" = "dashed", "Selected k = 3" =
"dotted"),
guide = guide_legend(override.aes = list(color =
c("#c04c3f", "#3a9e5f")))
) +
scale_x_continuous(breaks = 1:10) +
labs(
title = "Elbow Method \u2014 2014 K-Means Clustering",
x = "Number of Clusters (k)",
y = "Total Within-Cluster SS"
)

# -----
# SILHOUETTE ANALYSIS (Figure 10)
# -----
set.seed(123)
sil_scores <- sapply(2:10, function(k) {
km <- kmeans(vars_sc, centers = k, nstart = 25, iter.max =
100)
ss <- silhouette(km$cluster, dist(vars_sc))
mean(ss[, 3])
})

sil_data <- data.frame(k = 2:10, Silhouette =
round(sil_scores, 3))

ggplot(sil_data, aes(x = k, y = Silhouette)) +
theme_bw() +
theme(
text = element_text(family = "serif"),
plot.title = element_text(hjust = 0.5, face = "bold", size =
22),
axis.title = element_text(size = 18),
axis.text = element_text(size = 16, color = "black"),
legend.text = element_text(size = 14),
panel.grid.major = element_line(color = "#ebebeb", linetype
= "dashed", linewidth = 0.8),
panel.grid.minor = element_blank(),
panel.border = element_rect(color = "black", fill = NA),
legend.position = c(0.78, 0.18),
legend.background = element_rect(fill = "white", color =
"lightgrey")
) +
geom_line(color = "#1f4172", linewidth = 1.5) +
geom_point(color = "#1f4172", size = 5) +
geom_text(aes(label = sprintf("%3f", Silhouette)),
vjust = -1.5, size = 5, family = "serif", fontface = "bold") +
geom_vline(aes(xintercept = 3, linetype = "Selected k = 3"),
color = "#3a9e5f", linewidth = 1.1) +
scale_linetype_manual(
name = NULL,
values = c("Selected k = 3" = "dotted"),
guide = guide_legend(override.aes = list(color =
"#3a9e5f"))
) +
scale_x_continuous(breaks = 2:10) +
labs(
title = "Silhouette Analysis \u2014 2014 K-Means Clustering",
x = "Number of Clusters (k)",
y = "Average Silhouette Width"
)

# -----
# FIT FINAL K = 3 SOLUTION
# -----
set.seed(123)
km3 <- kmeans(vars_sc, centers = 3, nstart = 25, iter.max =
100)

sil3 <- silhouette(km3$cluster, dist(vars_sc))
overall_sil <- mean(sil3[, 3])
cat(sprintf("Overall silhouette (k=3): %3f\n", overall_sil))

# Summary table — now includes Dead_standing
cluster_summary <- data.frame(
Cluster = km3$cluster,
Snapping = multi_data$Snapping,
Windthrow = multi_data$Windthrow,
Branch_fall = multi_data$Branch_fall,
Dead_standing = multi_data$Dead_standing,
Sil_width = round(sil3[, 3], 3),
Species = rownames(multi_data)
)

table9 <- do.call(rbind, lapply(sort(unique(km3$cluster)),
function(cl) {
sub <- cluster_summary[cluster_summary$Cluster == cl, ]
data.frame(
Cluster = cl,
n = nrow(sub),
Pct = round(nrow(sub) / nrow(cluster_summary) * 100,
1),
Mean_Snap = round(mean(sub$Snapping), 3),
Mean_Wind = round(mean(sub$Windthrow), 3),
Mean_BF = round(mean(sub$Branch_fall), 3),
Mean_DS = round(mean(sub$Dead_standing), 3), # <--
added
Mean_Sil = round(mean(sub$Sil_width), 3)
)
}))
cat("\n--- Table 9: K-Means Clustering Results (k = 3) ---\n")
print(table9)

# -----
# PCA SPACE CLUSTER PLOT (Figure 11)
# -----
pca_scores <- as.data.frame(pca_res$x[, 1:2])
pca_scores$Cluster <- factor(km3$cluster)
pca_scores$Species <- rownames(multi_data)

# Cluster centroid labels derived from all four variables
cent_df <- as.data.frame(km3$centers)

```

```
colnames(cent_df) <- c('Snapping', 'Windthrow',
'Branch_fall', 'Dead_standing')
cent_df$Cluster <- as.character(1:3)
```

```
lbl <- rep(NA, 3)
lbl[which.max(cent_df$Windthrow)] <- "Extreme
windthrow"
lbl[which.max(cent_df$Snapping)] <- "Extreme snapping"
lbl[is.na(lbl)] <- "Low disturbance"
clust_labels <- setNames(lbl, 1:3)
```

```
n_per <- table(km3$cluster)
```

```
pca_scores$ClustLabel <- paste0(
"Cluster ", pca_scores$Cluster, "\u2014 ",
clust_labels[as.character(pca_scores$Cluster)],
"(n=", n_per[as.character(pca_scores$Cluster)], ")")
```

```
# Label only species outside the dominant (largest) cluster
outlier_sp <- pca_scores$Species[
pca_scores$Cluster != names(which(n_per == max(n_per)))
]
```

```
color_map <- setNames(
c("#3a9e5f", "#c04c3f", "#7497ba"),
c(
paste0("Cluster 1 \u2014 ", clust_labels["1"], "(n=",
n_per["1"], ")"),
paste0("Cluster 2 \u2014 ", clust_labels["2"], "(n=",
n_per["2"], ")"),
paste0("Cluster 3 \u2014 ", clust_labels["3"], "(n=",
n_per["3"], ")")
)
)
```

```
ggplot(pca_scores, aes(x = PC1, y = PC2, color =
ClustLabel)) +
theme_bw() +
theme(
text = element_text(family = "serif"),
plot.title = element_text(hjust = 0.5, face = "bold", size
= 20),
axis.title = element_text(size = 16),
axis.text = element_text(size = 14, color = "black"),
legend.text = element_text(size = 14),
legend.title = element_blank(),
legend.position = c(0.76, 0.88),
legend.background = element_rect(fill = "white", color =
"lightgrey"),
panel.grid.major = element_line(color = "#ebebeb",
linetype = "dashed", linewidth = 0.8),
panel.grid.minor = element_blank(),
panel.border = element_rect(color = "black", fill = NA)
) +
geom_point(size = 3, alpha = 0.75) +
# Layer 1: numbers for ALL points
geom_text_repel(
aes(label = match(Species, rownames(multi_data))),
size = 4.5, family = "serif", fontface = "bold",
box.padding = 0.4, point.padding = 0.3,
```

```
segment.color = "grey50",
show.legend = FALSE
) +
# Layer 2: italic species names for OUTLIER points only
geom_text_repel(
data = pca_scores[pca_scores$Species %in%
outlier_sp, ],
aes(label = Species),
size = 5.5, family = "serif", fontface = "italic",
box.padding = 0.8, point.padding = 0.4,
nudge_y = -0.3,
segment.color = "grey30", segment.linetype = "dashed",
show.legend = FALSE
) +
scale_color_manual(values = color_map) +
labs(
title = "K-Means Cluster Membership in PCA Space (k =
3)",
x = sprintf("PC1 (%.1f%% variance)", pct[1]),
y = sprintf("PC2 (%.1f%% variance)", pct[2])
)
```

➤ Stage 5: Multiple Linear Regression

• AI Prompt 9

Conduct multiple linear regression with gap size as the dependent variable and windthrow, snapping, branch_fall and dead standing as predictors, following the points below:

- ✓ Fit OLS model; print full summary including coefficients, SEs, t-values, p-values, R², adjusted R², and F-statistic with degrees of freedom.
- ✓ Print the ANOVA table and compute RMSE; compare RMSE to the simple regression RMSE (1.934) and the SD of snapping to quantify improvement.
- ✓ Compute Variance Inflation Factors for both predictors to confirm absence of multicollinearity; interpret VIF values.
- ✓ Produce a predicted vs. observed scatter plot with 1:1 reference line (red); annotate with R² and RMSE; label the two extreme outlier species.

Response

```
# =====
# STAGE 5: MULTIPLE LINEAR REGRESSION
# =====
library(car)
```

```
# NOTE: Dead_standing has only 1 non-zero value
(Zanthoxylum sp., value = 1).
# It is not all zeros so it is retained, but its coefficient will
carry a very
# wide confidence interval — interpret with caution.
```

```
mlr <- lm(Gap_size ~ Snapping + Windthrow + Branch_fall
+ Dead_standing,
data = multi_data)
summary(mlr)
anova(mlr)

# RMSE
```

```

rmse_mlr <- sqrt(mean(mlr$residuals^2))
cat('MLR RMSE :', round(rmse_mlr, 4), '\n')
cat('SD(Gap_size):', round(sd(multi_data$Gap_size), 4), '\n')

# Variance Inflation Factors
cat('\n--- Variance Inflation Factors ---\n')
vif_vals <- vif(mlr)
print(round(vif_vals, 3))

vif_ok <- names(vif_vals[vif_vals < 5])
vif_hi <- names(vif_vals[vif_vals >= 5])
if (length(vif_hi) == 0) {
  cat('VIF < 5 for all predictors: no multicollinearity
  detected.\n')
} else {
  cat('Predictors with VIF >= 5 (potential multicollinearity):',
  paste(vif_hi, collapse = ', '), '\n')
  cat('Predictors with VIF < 5 (acceptable):', paste(vif_ok,
  collapse = ', '), '\n')
}
# -----
# Predicted vs. Observed plot — label the two largest gaps
# -----
pred <- mlr$fitted.values
obs <- multi_data$Gap_size
out_idx <- order(obs, decreasing = TRUE)[1:2] #
Protomegalaria & Oubanguia

par(
  family = "serif",
  cex.main = 1.8,
  cex.lab = 1.5,
  cex.axis = 1.3,
  mar = c(5, 5, 4, 2)
)

plot(pred, obs,
  xlab = "Predicted Gap Size (m²)",
  ylab = "Observed Gap Size (m²)",
  main = "Multiple Regression: Predicted vs. Actual Gap
  Size",
  pch = 16,
  col = "steelblue",
  cex = 1.4,
  xlim = range(pred) * c(0.9, 1.1))
abline(0, 1, col = "red", lwd = 2.5)
text(pred[out_idx], obs[out_idx],
  labels = rownames(multi_data)[out_idx],
  pos = 2,
  cex = 1.3,
  col = "darkred",
  font = 3)
legend("topleft",
  legend = c(paste0("R\u00B2 = ",
  round(summary(mlr)$r.squared, 3)),
  paste0("RMSE = ", round(rmse_mlr, 3))),
  bty = "n",
  cex = 1.3)

```

E. Comparison of Vibe Analysis and Manual Analysis

In order to assess how reliably and practically vibe analysis performs relative to conventional analyst-coded approaches (manual analysis), both methods were applied in parallel to the same 183 treefall gaps dataset from Korup National Park. The comparison was structured around five evaluative dimensions: numerical agreement, assumption-checking and method selection, Decision Process Comparison, Interpretive Alignment and process efficiency.

➤ Numerical Agreement

Fifteen statistical outputs spanning the full analytical pipeline; bivariate correlations, regression diagnostics, multivariate correlation coefficients, PCA outputs, cluster statistics, and multiple regression results were compared directly between approaches. Agreement was defined as exact equivalence to three or more decimal places. Because LLMs can generate plausible but fabricated numerical values, all statistics reported in the vibe analysis narrative were independently verified by executing the AI-generated R code in R and cross-checking the console output against the stated values before any comparison was made.

➤ Assumption-Checking and Method Selection

Both approaches subjected all disturbance count variables to distributional assessment prior to inferential testing, combining skewness and kurtosis statistics with graphical diagnostics including histograms with fitted normal curves, Q-Q plots, and scatter plot matrices. Method selection was then treated as a decision point conditional on these outcomes. The manual analysis followed a pre-specified protocol: normality violations were documented and non-parametric correlations formally designated as co-primary alongside Pearson. The vibe analysis received no such instruction. Method selection was left entirely to the AI's autonomous judgment, allowing a direct test of whether the LLM could replicate analyst reasoning without being prompted to do so.

➤ Decision Process Comparison

Five key analytical decision points were traced across both approaches: recognition of non-normality, selection of the primary analytical framework, retention of parametric tests for comparative purposes, determination of the optimal cluster number, and production of ecological interpretation. For each point, the mechanism, evidence consulted, and sequencing of steps were documented to characterise whether the two approaches converged through equivalent reasoning or arrived at the same conclusions by different routes.

➤ Interpretive Alignment

Post-analysis ecological interpretation was compared qualitatively across three criteria: accurate identification and biological characterisation of the two extreme outlier species; recognition of the disturbance syndrome structure revealed by the correlation matrix and PCA; and coherence of the overall statistical conclusion with the multivariate and clustering evidence. This dimension assessed not whether the AI produced correct numbers, but whether it produced meaningful explanations of them.

➤ *Efficiency, Risk, and Validation*

Process efficiency was assessed by the number of analytical stages, coding cycles, and iterative decisions required under each approach. The vibe analysis was completed across nine consecutive prompts, with R code, diagnostics, and interpretive narrative generated within each exchange. The manual analysis proceeded through sequential coding, debugging, and stage-by-stage interpretation. The hallucination risk inherent to LLM quantitative reporting (Alkaissi and McFarlane, 2023) was controlled through mandatory code execution: every AI-stated value was verified against R console output before being accepted as valid.

III. RESULTS

An examination of disturbance observations across 183 treefall gaps in Korup National Park forms the basis of this analysis, with a focus on characterising relationships between snapping and gap size, bringing to light the multivariate structure of disturbance-related variables. The analytical approach unfolded across five stages, drawing on a range of descriptive and inferential statistical techniques. Given the highly skewed and non-normal nature of all count variables, confirmed by extreme positive skewness coefficients ranging from 1.8 to 8.4 and excess kurtosis values up to 71, a combination of parametric and non-parametric approaches was employed, with non-parametric results receiving primary interpretive weight.

A. *Exploratory Data Analysis*

➤ *Bivariate Descriptive Statistics*

The bivariate dataset comprises 183 treefall gaps observations with records on snapping count and canopy gap size. As shown in Table 2, snapping is a count variable ranging from 0 to 11 (mean = 1.260, median = 1.000, SD = 2.089, skewness = 3.289, kurtosis = 14.113). The mean of 1.260 indicates that, on average, slightly more than one snapping event was recorded per species, while the median of 1 confirms that at least half of all species recorded a single snapping event. The strong positive skewness (3.289) and markedly leptokurtic profile (14.113) reflect a distribution dominated by low counts, with a small number of species recording unusually high snapping frequencies up to a maximum of 11, driving the heavy right tail. Total gap sizes varied widely from 25.0 to 4134.8 m² (mean = 351.538 m², median = 171.500 m², SD = 608.442 m², skewness = 4.593, kurtosis = 26.198), with the mean more than doubling the median, consistent with extreme right skew driven by a small number of very large gaps. The distributional properties of both variables deviate markedly from normality, snapping through its zero-inflated, overdispersed count structure, and gap size through its extreme positive skewness and heavily leptokurtic profile, jointly justifying the use of non-parametric methods in the analyses.

Table 2 Bivariate Descriptive Statistics of Snapping Counts and Total Gap Size

| Statistics | Variable | |
|------------|------------------|----------------------------|
| | Snapping (count) | Gap Size (m ²) |
| Mean | 1.26 | 351.538 |
| Median | 1 | 171.5 |
| SD | 2.089 | 608.442 |
| Min | 0 | 25 |
| Max | 11 | 4134.8 |
| Skewness | 3.289 | 4.593 |
| Kurtosis | 14.113 | 26.198 |

Figure 1 displays the frequency histogram of snapping counts with a fitted Poisson expected curve. The distribution is strongly right-skewed (skewness = 3.289), with the greatest frequency of species recording either zero or one snapping event. The bar at count zero is the tallest, representing the largest single group, followed by the bar at count one, and frequencies decline sharply thereafter. A small number of species record high snapping counts approaching the maximum of 11, and these outlying observations are the primary drivers of the pronounced positive skew and elevated kurtosis (14.113). The superimposed Poisson curve, fitted at $\lambda = 1.260$, tracks the general declining trend reasonably in the lower counts but underestimates the zero bar and fails to capture the heavy right tail, visually confirming that the snapping distribution departs from a Poisson process and exhibits overdispersion. The leptokurtic kurtosis value of 14.113 further confirms the presence of extreme values relative to a theoretical Poisson distribution. Overall, snapping counts are dominated by absences and low frequencies, with rare but influential high-count species, collectively confirming

the unsuitability of parametric or standard Poisson assumptions and supporting the use of non-parametric methods for subsequent analyses.

Figure 2 displays the frequency histogram of canopy total gap sizes (m²) with a fitted normal curve. The distribution is extremely right-skewed (skewness = 4.593), with the vast majority of gaps concentrated in the lower size range, particularly below 500 m². The tallest bars are found at the smallest size classes, and frequencies decline sharply as gap size increases, reflecting the rarity of very large gaps. A small number of total gap sizes exceed 2000 m², with the maximum reaching 4134.8 m², and these extreme values are the primary drivers of the pronounced positive skew and heavily leptokurtic profile (kurtosis = 26.198). The superimposed normal curve peaks near the mean of 351.5 m² and traces a smooth, symmetric bell, a profile clearly inconsistent with the empirical distribution, which is concentrated well below the mean. The mean (351.538 m²) more than doubles the median (171.500 m²), consistent with

the strong rightward elongation of the distribution caused by a minority of exceptionally large gaps. Overall, total canopy gap sizes in this dataset span an exceptionally wide range (25.0-4134.8 m²) and are dominated by small-to-medium

openings, with a small number of very large gaps exerting a disproportionate influence on the distributional shape, collectively confirming the unsuitability of parametric assumptions for this variable.

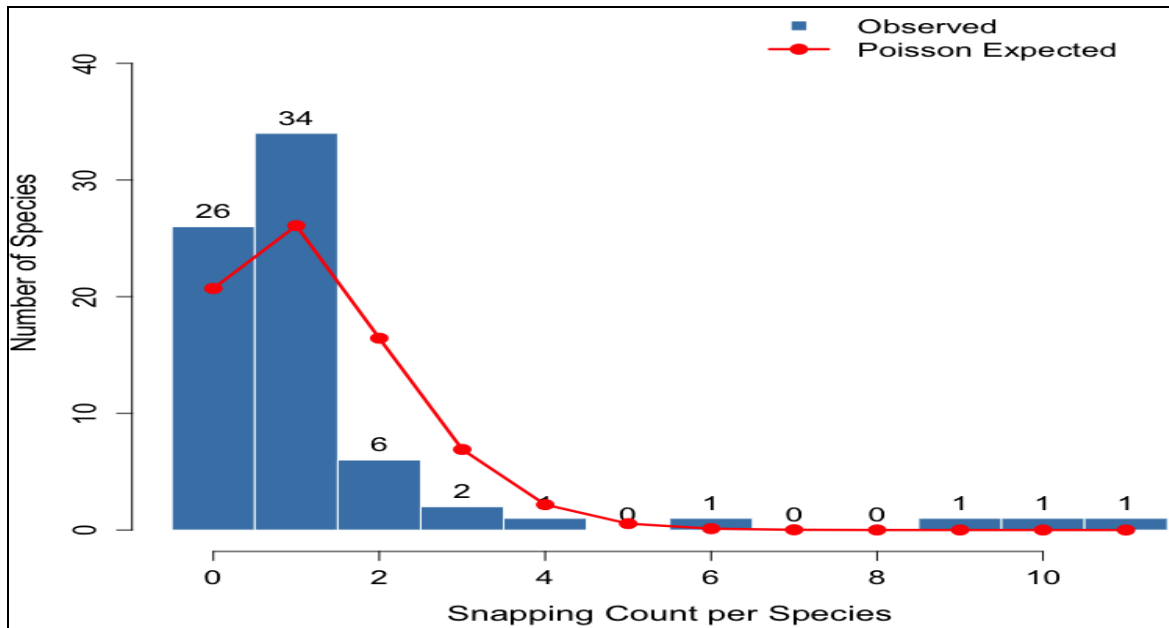


Fig 1 Frequency Histogram of Snapping Counts (n = 183 Treefall Gaps) with Fitted Normal Curve

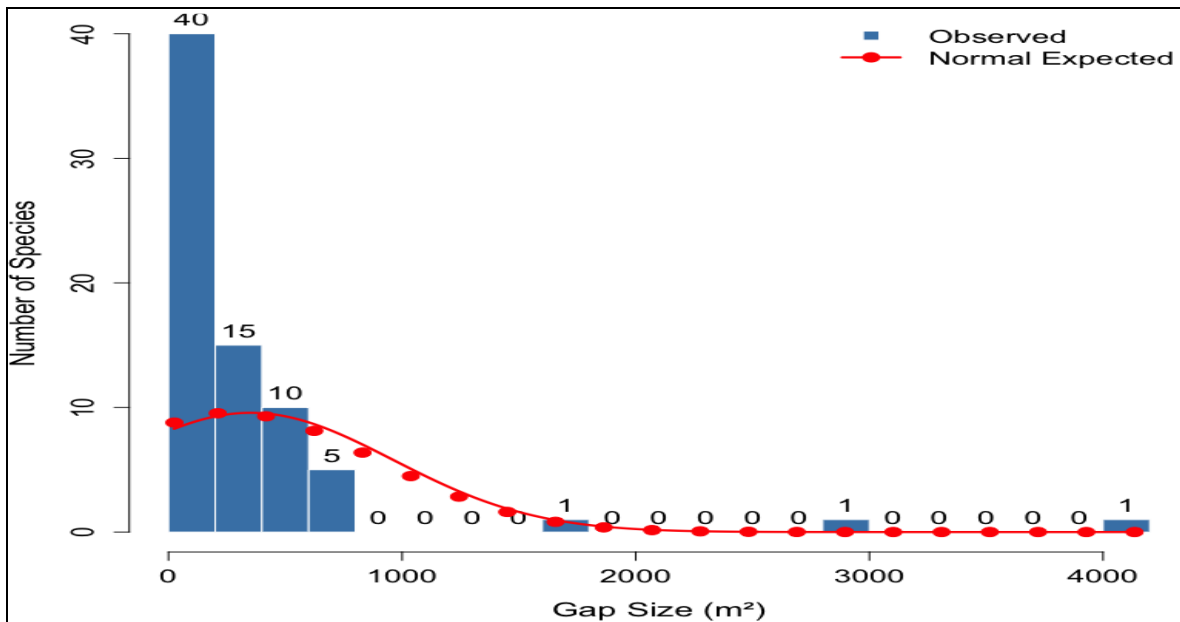


Fig 2 Frequency Histogram of Total Gap Sizes (m²)

➤ *Multivariate Descriptive Statistics*

Table 3 presents summary statistics for four disturbance variables across 183 treefall gaps observations. All variables are right-skewed count data, with zero as the most common value, consistent with the zero medians and positive skewness coefficients observed across the board. Snapping was the most frequently recorded disturbance type (mean = 1.260, median = 1), the only variable with a median greater than zero, indicating it occurred at least once in the majority of observations. Its range extended to a maximum of 11 events, with a skewness of 3.289 and kurtosis of 14.113, reflecting a

moderately leptokurtic distribution driven by a small number of high-disturbance species. Windthrow (mean = 0.479, max = 6) and branch fall (mean = 0.630, max = 4) were comparably infrequent, both with zero medians and positive skewness values (3.484 and 1.776, respectively), confirming their predominantly absent character across most observations. Dead standing was recorded in only 1 of the 73 species observations (mean = 0.014, max = 1), yielding an extreme skewness of 8.367 and kurtosis of 71.014; this near-constant behaviour precludes its inclusion in correlation analysis, PCA, and cluster analysis.

Table 3 Multivariate Descriptive Statistics of Disturbance Variables Across 183 Treefall Gaps

| Statistics | Disturbance variable | | | |
|------------|----------------------|-----------|-------------|---------------|
| | Snapping | Windthrow | Branch fall | Dead standing |
| Mean | 1.26 | 0.479 | 0.63 | 0.014 |
| Median | 1 | 0 | 0 | 0 |
| SD | 2.089 | 1.029 | 1.007 | 0.117 |
| Min | 0 | 0 | 0 | 0 |
| Max | 11 | 6 | 4 | 1 |
| Skewness | 3.289 | 3.484 | 1.776 | 8.367 |
| Kurtosis | 14.113 | 17.236 | 5.554 | 71.014 |

Figure 3 presents the scatter plot matrix for all four disturbance variables (Snapping, Windthrow, Branch fall, and Dead standing; $n = 183$ treefall gaps). Across virtually every panel, observations are concentrated in the lower-left corner of the plotting area, forming dense clusters close to the axes, the visual signature of zero-inflated count data, consistent with the zero medians and high skewness values (1.776-8.367) reported in Table 3. In the Snapping-Windthrow panel, a small number of points extending toward the upper right correspond to the maximum values of 11 and 6 respectively, confirming that extreme disturbance events are rare and largely independent across types. The Snapping-Branch fall panel reveals a slightly more dispersed cloud relative to other

pairs, consistent with their comparable means (1.260 and 0.630) and overlapping lower ranges, and branch fall's comparatively lower skewness (1.776) suggests a less extreme zero-inflation than the other variables. The Dead standing row and column are almost entirely flat at zero, with only one observation departing from the axis, reflecting its near-zero mean (0.014), zero median, and effectively binary range (0-1). Overall, the scatter plot matrix reinforces the descriptive statistics in Table 3 as sparse, highly zero-dominated count variables with pronounced positive skewness and elevated kurtosis, exhibiting no obvious linear co-variation, with any inter-variable association more appropriately explored through non-parametric or count-data methods.

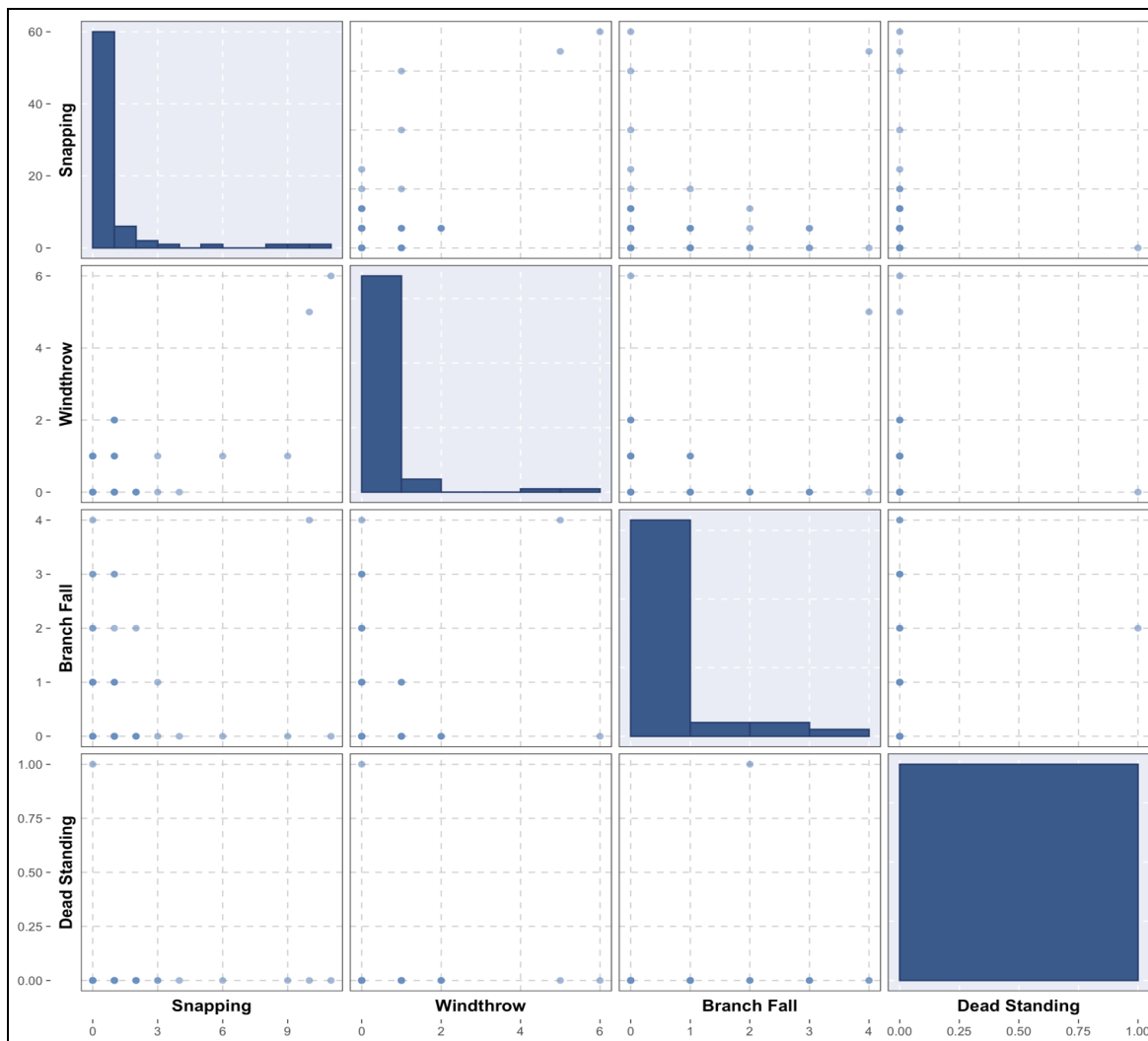


Fig 3 Scatter Plot Matrix of Disturbance Variables.

B. Bivariate Analysis

➤ **Correlation Analysis**

The three correlation methods yielded a consistent pattern of positive association between snapping and gap size (Table 4). The Pearson product-moment correlation ($r = 0.865$, 95% CI [0.793, 0.913], $p < 0.001$) detected a strong and statistically significant positive linear relationship, indicating that species associated with larger canopy gaps tend to record higher snapping counts. The 95% confidence interval lies entirely above zero, confirming the reliability of this positive direction. The non-parametric alternatives corroborated this finding: Spearman's rank correlation ($\rho = 0.455$, $p < 0.001$) and Kendall's tau ($\tau = 0.366$, $p < 0.001$) both

confirmed a statistically significant positive monotonic association. The coefficient of determination ($R^2 = 0.7483$) from the Pearson analysis reveals that gap size explains approximately 74.8% of the variance in snapping counts, representing substantial linear predictive power. The divergence in magnitude between Pearson ($r = 0.865$) and the non-parametric coefficients ($\rho = 0.455$, $\tau = 0.366$) is attributable to the highly skewed, zero-inflated nature of snapping counts, which can inflate Pearson's r in the presence of influential high-count observations; the non-parametric results therefore receive primary interpretive weight, and consistently support a moderate but statistically robust positive monotonic association between gap size and snapping frequency.

Table 4 Correlation Coefficients Between Snapping Counts and Total Gap Size

| Method | Value | p-value | 95% CI | Interpretation |
|----------------------|-------|---------|----------------|----------------|
| Pearson (r) | 0.865 | < 0.001 | [0.793, 0.913] | Significant |
| Spearman (ρ) | 0.455 | < 0.001 | - | Significant |
| Kendall's (τ) | 0.366 | < 0.001 | - | Significant |

Figure 4 plots snapping counts against gap size for all 183 treefall gaps observations, overlaid with the Ordinary Least Squares (OLS) regression line and its 95% confidence band (shaded). The scatter of points shows a clear upward trend, with higher snapping counts associated with progressively larger gap sizes. The majority of observations are clustered at low snapping counts (0-2) across the smaller gap sizes, while the few species recording the highest snapping counts approaching the maximum of 11 are associated with larger gaps. The regression line carries a

distinctly positive slope, and the relatively narrow confidence band at moderate gap sizes confirms this trend is statistically well-supported. The high R^2 value (0.748) is partly driven by the influential high-snapping observations at larger gap sizes, as reflected by the wider confidence band at the upper end of the gap size range. The annotation confirms Pearson $r = 0.865$, $p = 0.000$, and $R^2 = 0.748$, collectively providing strong visual and statistical evidence that larger canopy gaps are associated with greater snapping frequency among the species in this dataset.

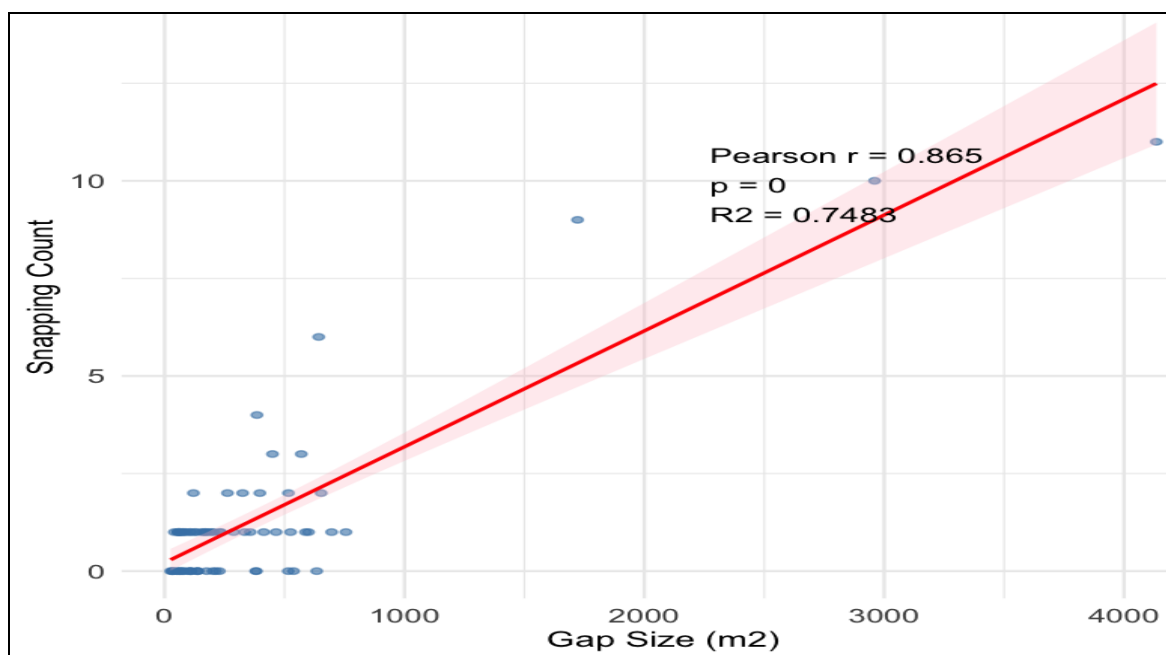


Fig 4 Scatter Plot of Snapping Counts Versus Gap Size (m²) with OLS Regression Line and 95% Confidence Band (Shaded).

➤ **Simple Linear Regression**

Simple linear regression of snapping on gap size confirmed the Pearson correlation finding (Table 5). The fitted model equation is: Snapping = 0.2164 + 0.0030 × Gap_Size. The slope coefficient ($\beta_1 = 0.0030$, SE = 0.0002, t

= 14.529, $p < 0.001$) is highly statistically significant, indicating that for each additional square metre of canopy gap, snapping count increases by approximately 0.003 events. The overall model explained $R^2 = 0.7483$ of snapping variance, with adjusted $R^2 = 0.7448$, confirming that gap size is a strong

linear predictor of snapping frequency and represents a substantial improvement over an intercept-only null model. The ANOVA F-test ($F(1, 71) = 211.1, p < 0.001$) confirmed the model's overall predictive utility. The RMSE of 1.0406 is considerably lower than the standard deviation of snapping (2.0885), confirming that linear predictions from gap size

meaningfully outperform using the sample mean as a constant predictor. The non-parametric results (Spearman $\rho = 0.455, p < 0.001$; Kendall $\tau = 0.366, p < 0.001$) corroborate the linear finding, jointly confirming a robust positive association between gap size and snapping frequency.

Table 5 Simple Linear Regression of Snapping Counts and Total Gap Size

| Parameter | Estimate | Standard Error | t-value | p-value |
|--------------------------------|--------------------|----------------|---------|------------|
| Intercept (β_0) | 0.2164 | 0.1429 | 1.515 | 0.134 |
| Slope - Gap Size (β_1) | 0.0030 | 0.0002 | 14.529 | < 0.001*** |
| R ² | 0.7483 | - | - | - |
| Adjusted R ² | 0.7448 | - | - | - |
| F-statistic | 211.1 (df = 1, 71) | - | - | < 0.001*** |
| RMSE | 1.0406 | - | - | - |
| SD (Snapping) | 2.0885 | - | - | - |

Figure 5 presents the four standard regression diagnostic plots for the snapping against gap size model. The Residuals vs Fitted plot shows a broadly positive fan-shaped spread, with residuals widening at higher fitted values, suggesting mild heteroscedasticity, a pattern consistent with the overdispersed, zero-inflated nature of snapping counts. A small number of high-leverage observations with large positive residuals, including *Dichostemma glaucescens* (snapping count = 9), *Oubangia alata* (snapping count = 10), and *Protomegabaria stapfiana* (snapping count = 11), exert disproportionate influence on the regression line. The Q-Q plot reveals moderate departure from normality in the upper tail, driven by these same high-count species, confirming that

the residuals are not normally distributed. The Scale-Location plot corroborates the heteroscedasticity, with the LOESS smoother trending upward across fitted values. The Residuals vs Leverage plot identifies the high-count species as the most influential observations, with several approaching or exceeding Cook's distance contours. Together, these diagnostic panels confirm violations of the homoscedasticity and normality assumptions, indicating that while the linear model captures a strong overall trend ($R^2 = 0.748$), the underlying distributional properties of snapping counts would be more rigorously addressed through count-data regression approaches such as negative binomial or zero-inflated models.

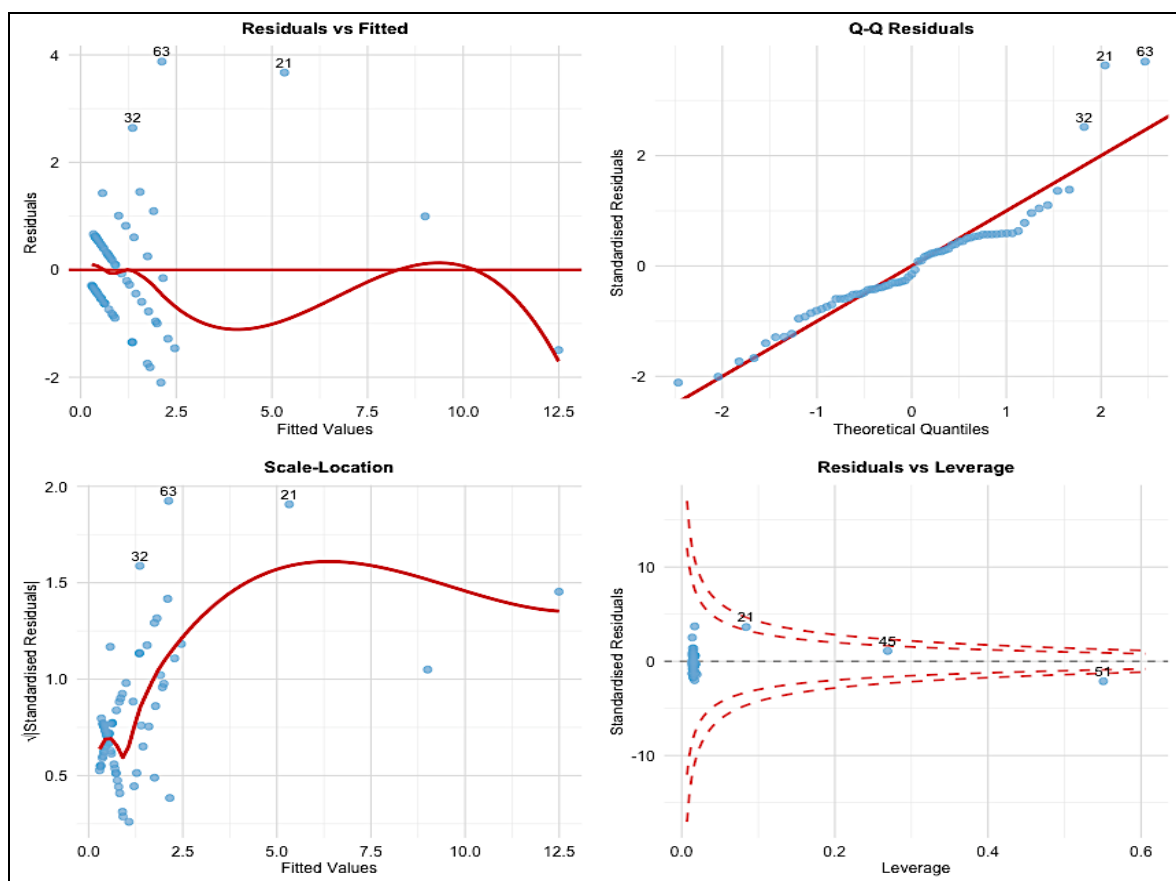


Fig 5 Regression Diagnostic Plots for the Simple Linear Regression Model

➤ *Contingency Table Analysis*

Categorising snapping counts and total gap size of 73 major causal factors species, and constructing a 3 × 3 contingency table provided a distribution-free test of independence (Table 6). Snapping was categorised with respect to frequency as Low (0-1), Medium (2), or High (≥ 3), and gap size as Small (< 100 m²), Medium (100-400 m²), or Large (> 400 m²) following Brokaw (1985). 18 gaps exceeded 400 m². The chi-square test yielded $\chi^2(4) = 19.207$, $p < 0.001$, indicating a statistically significant association between snapping and gap size categories. However, a warning was produced regarding the chi-square approximation, as several

expected cell frequencies fell below 5, most notably in the Medium (2) and High (≥ 3) rows, which reduces confidence in the test statistic. Cramér's V = 0.363 indicates a moderate effect size. A notable pattern is visible in the table. Low snapping was concentrated in small and medium gaps, whereas High snapping was disproportionately associated with large gaps, consistent with positive rank correlation. The overall pattern combining Pearson (significant), non-parametric (significant), and chi-square (significant, though assumption-violated) results consistently supports a positive association between snapping counts and gap size, with higher snapping resulting to larger canopy gaps.

Table 6 Contingency Table of Snapping Counts Against Gap Size Categories

| Snapping category | Small (< 100 m ²) | Medium (100-400 m ²) | Large (> 400 m ²) | Total |
|-------------------|-------------------------------|----------------------------------|-------------------------------|-----------|
| Low (0-1) | 22 | 28 | 10 | 60 |
| Medium (2) | 0 | 4 | 2 | 6 |
| High (≥ 3) | 0 | 1 | 6 | 7 |
| Total | 22 | 33 | 18 | 73 |
| χ^2 | 19.207 | | | |
| p-value | < 0.001 | | | |
| Cramér's V | 0.363 | | | |

C. *Multivariate Correlation Matrix Results*

The Pearson correlation matrix for the four disturbance variables (Table 7) revealed a single statistically significant association among all variable pairs examined. A strong positive correlation was detected between snapping and windthrow ($r = 0.659$, $p < 0.01$), indicating that species recording more snapping events also tend to record more windthrow events, suggesting a shared biomechanical or environmental vulnerability between these two disturbance mechanisms. By contrast, branch fall was essentially uncorrelated with both snapping ($r = 0.013$, $p = 0.911$) and windthrow ($r = 0.013$, $p = 0.915$), indicating that branch fall

operates as a largely independent disturbance process. Dead standing was included in the matrix for completeness but its correlations must be interpreted with extreme caution, as only a single event was recorded across all 183 treefall gaps (variance = 0.014), rendering Pearson correlation unreliable for this variable. Its correlations with snapping ($r = -0.072$, $p = 0.547$), windthrow ($r = -0.055$, $p = 0.642$), and branch fall ($r = 0.161$, $p = 0.172$) were all non-significant. Overall, the correlation matrix reveals that snapping and windthrow form a coherent disturbance pair, while branch fall and dead standing each represent distinct and independent disturbance mechanisms.

Table 7 Pearson Correlation Matrix with Significance Levels for Four Disturbance Variables

| Variable | Snapping | Windthrow | Branch fall | Dead standing |
|---------------|------------------------|------------------------|-----------------------|------------------------|
| Snapping | 1.000 | 0.659** ($p < 0.01$) | 0.013 ($p = 0.911$) | -0.072 ($p = 0.547$) |
| Windthrow | 0.659** ($p < 0.01$) | 1.000 | 0.013 ($p = 0.915$) | -0.055 ($p = 0.642$) |
| Branch fall | 0.013 ($p = 0.911$) | 0.013 ($p = 0.915$) | 1.000 | 0.161 ($p = 0.172$) |
| Dead standing | -0.072 ($p = 0.547$) | -0.055 ($p = 0.642$) | 0.161 ($p = 0.172$) | 1.000 |

Figure 6 presents the Pearson correlation heatmap for snapping, windthrow, branch fall, and dead standing. The diagonal cells ($r = 1.000$) are filled with solid deep red. The snapping-windthrow cell displays a moderate red shade labelled 0.66, representing the only statistically significant off-diagonal association in the matrix. The snapping-branch fall and windthrow-branch fall cells are both near-white ($r \approx 0.01$), confirming the absence of any meaningful relationship

involving branch fall. Cells involving dead standing are similarly near-white, consistent with its near-constant distribution. Red shading indicates positive correlation and blue indicates negative correlation. The heatmap visually reinforces that snapping and windthrow share a common co-occurrence pattern, while branch fall and dead standing each behave as independent disturbance types with no meaningful association with the other variables.

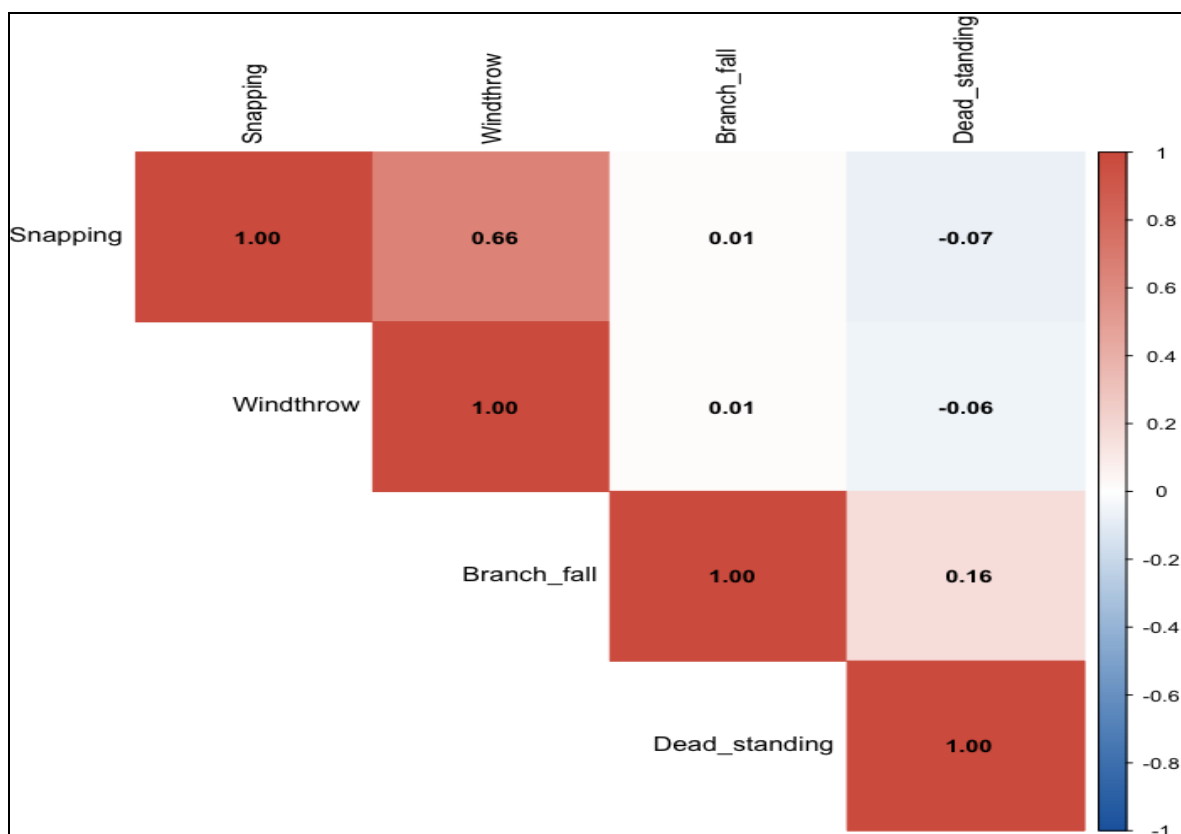


Fig 6 Pearson Correlation Heatmap for Snapping, Windthrow, and Branch Fall

D. Principal Component and Cluster Analysis

➤ **Principal Component Analysis**

PCA was conducted on all four disturbance variables (snapping, windthrow, branch fall, and dead standing) after standardisation, and the analysis extracted four components.

PC1 and PC2 met the Kaiser criterion (eigenvalue > 1), together explaining 70.7% of total variance (Table 8). PC3 (eigenvalue = 0.832) and PC4 (eigenvalue = 0.341) fell below the criterion, though all four components cumulatively account for 100% of variance.

Table 8 PCA Eigenvalues and Variance of Disturbance Variables

| Component | Eigenvalue | Variance Explained (%) | Cumulative (%) |
|-----------|------------|------------------------|----------------|
| PC1 | 1.671 | 41.8 | 41.8 |
| PC2 | 1.157 | 28.9 | 70.7 |
| PC3 | 0.832 | 20.8 | 91.5 |
| PC4 | 0.341 | 8.5 | 100.0 |

The component loadings matrix (Table 9) reveals that PC1 (41.8% variance) loads strongly and near-equally on snapping (0.702) and windthrow (0.700), with negligible contributions from branch fall (-0.005) and dead standing (-0.134). This pattern indicates that PC1 represents a shared snapping-windthrow disturbance intensity axis, capturing co-occurring mechanical stem failure events. PC2 (28.9% variance) is dominated by branch fall (0.720) and dead standing (0.687), with near-zero loadings on snapping (0.062) and windthrow (0.075), indicating that PC2 captures an

independent crown and standing mortality disturbance axis. PC3 (20.8% variance) contrasts branch fall (-0.694) against dead standing (0.714), with minimal contributions from snapping (0.053) and windthrow (0.078), representing residual variation between these two otherwise co-loading variables. PC4 (8.5% variance) contrasts snapping (-0.708) against windthrow (0.706), capturing the residual difference between the two stem-failure mechanisms that load together on PC1.

Table 9 PCA Component Loadings Matrix of Disturbance Variables

| Variable | PC1 | PC2 | PC3 |
|---------------|--------|-------|--------|
| Snapping | 0.702 | 0.062 | 0.053 |
| Windthrow | 0.700 | 0.075 | 0.078 |
| Branch fall | -0.005 | 0.720 | -0.694 |
| Dead standing | -0.134 | 0.687 | 0.714 |

Figure 7 presents the scree plot of eigenvalues with the Kaiser criterion threshold marked by a red dashed line. PC1 (eigenvalue = 1.671) and PC2 (eigenvalue = 1.157) both exceeded the Kaiser criterion and are formally retained. PC3 (eigenvalue = 0.832) falls moderately below the threshold, and PC4 (eigenvalue = 0.341) lies well below it. The

moderate drop after PC2 reflects the partitioning of disturbance variance into two relatively distinct axes: a stem-failure axis (PC1) driven by co-occurring snapping and windthrow events, and a crown and standing mortality axis (PC2) driven by branch fall and dead standing.

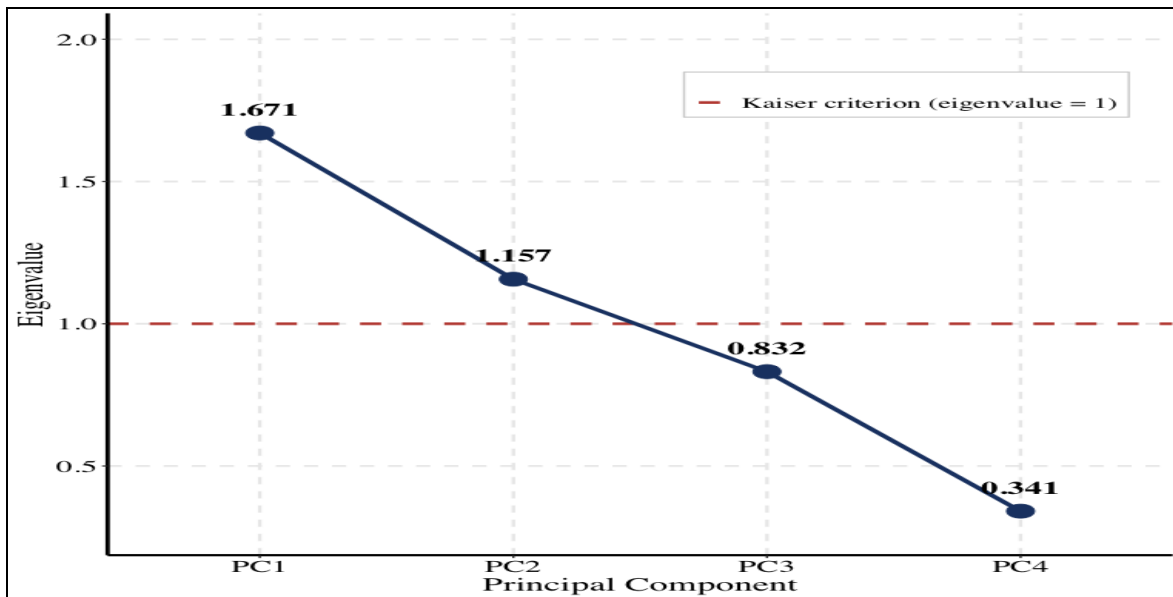


Fig 7 Scree Plot of Eigenvalues by Principal Component

Figure 8 presents the PCA biplot. The snapping and windthrow vectors point in the same direction along PC1, confirming their shared variance and tendency to co-occur in larger disturbance events. The branch fall and dead standing vectors are oriented closely together along PC2, nearly orthogonal to the snapping-windthrow axis, consistent with their near-zero PC1 loadings and confirming that crown-related and standing mortality disturbances operate as mechanisms largely independent of stem-failure events. The majority of species are compressed into a dense cluster near

the origin, reflecting generally low disturbance intensity across most taxa. A small number of outlying observations project along PC1, representing species associated with disproportionately high combined snapping-windthrow disturbance, while species projecting along PC2 are associated with elevated branch fall or dead standing events. The near-perpendicular orientation between the two groups of vectors visually confirms the ecological independence of stem-failure and crown mortality disturbance mechanisms.

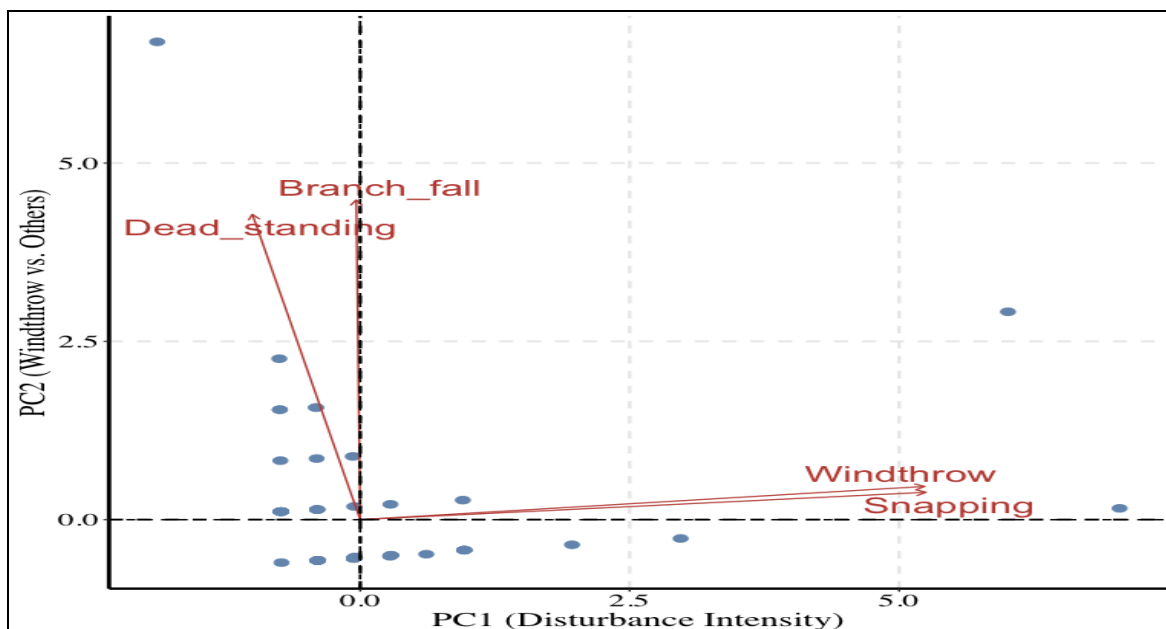


Fig 8 PCA Biplot Displaying Gaps Scores and Variable Loading Vectors

➤ *Cluster Analysis*

K-means cluster analysis was applied to the standardised four-variable disturbance matrix (snapping, windthrow, branch fall, dead standing) across 183 treefall gaps (Table 10). Cluster 1 (n = 70, 95.9%) comprised the dominant low-disturbance group, with low mean snapping (1.014), low windthrow (0.343), moderate branch fall (0.571), no dead standing (0.000), and a high silhouette width of 0.792, indicating strong internal cohesion. Cluster 2 (n = 2, 2.7%) isolated the two extreme disturbance outlier species, recording by far the highest mean snapping (10.500) and windthrow (5.500), with moderate branch fall (2.000) and a silhouette

width of 0.436. Cluster 3 (n = 1, 1.4%) was a singleton cluster isolating the sole species recorded with a dead standing event, with zero snapping and windthrow, a branch fall mean of 2.000, and a silhouette width of 0.000, consistent with the known behaviour of singleton clusters which cannot be internally evaluated.

Together, Clusters 2 and 3 account for only 4.1% of all species, both representing distinct extreme cases. Cluster 1 accounts for the overwhelming majority of species, all occupying the low-to-moderate disturbance space with snapping as the slightly dominant disturbance mode.

Table 10 K-Means Clustering (k = 3) of Four-Variable Disturbance Matrix

| Cluster | n | % | Mean Snapping | Mean Windthrow | Mean Branch Fall | Mean Dead Standing | Silhouette |
|--|----|------|---------------|----------------|------------------|--------------------|------------|
| Cluster 1 (Low disturbance) | 70 | 95.9 | 1.014 | 0.343 | 0.571 | 0.000 | 0.792 |
| Cluster 2 (Extreme snapping & windthrow) | 2 | 2.7 | 10.500 | 5.500 | 2.000 | 0.000 | 0.436 |
| Cluster 3 (Dead standing singleton) | 1 | 1.4 | 0.000 | 0.000 | 2.000 | 1.000 | 0.000 |

The elbow method identified k = 2 as the statistically optimal solution, marked by a sharp inflection in the total within-cluster sum of squares curve, after which the rate of decline slowed considerably and flattened progressively

thereafter. Despite this, k = 3 was selected to preserve meaningful distinctions, specifically to separate the dead standing singleton from the broader low-disturbance group (Figure 9).

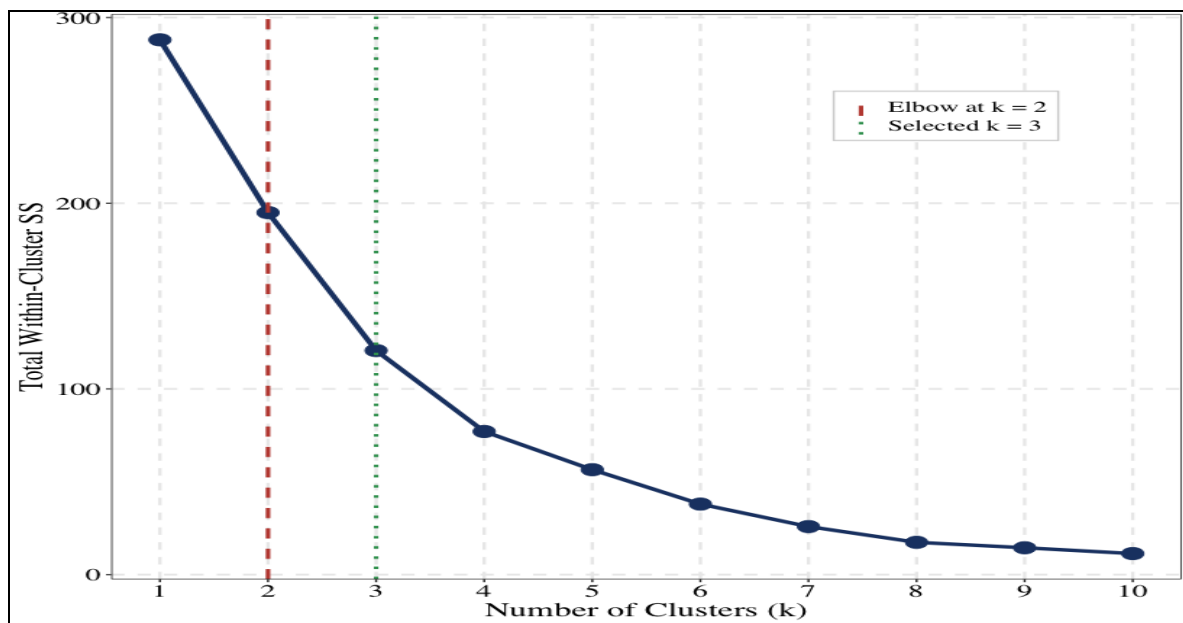


Fig 9 Elbow Plot of Total within-Cluster Sum of Squares Against Number of Clusters k

Silhouette analysis further informed cluster selection. The k = 3 solution yielded a strong overall silhouette score of 0.772, substantially higher than what the previous three-variable solution produced (0.449), demonstrating that including dead standing as a fourth variable improved the internal coherence of the clustering structure. The singleton nature of Cluster 3 drives its silhouette width to 0.000 by definition, but this does not reflect poor clustering, rather, it reflects the genuine isolation of the one species exhibiting dead standing as its primary disturbance mode. The two-cluster solution at k = 2 remains the statistically simplest partition, but k = 3 was retained to separately identify both the

extreme snapping-windthrow outliers and the dead standing singleton (Figure 10).

Visualisation of cluster membership in PCA space confirmed these groupings. The two Cluster 2 outlier species projected far to the right along PC1, at divergent PC2 positions, indicating both high overall disturbance intensity and differing windthrow profiles between the two individuals. The Cluster 3 singleton projected distinctly from the main cloud along PC2, consistent with its unique dead standing signature in all other species. Cluster 1 species occupied the central-left region of the ordination, forming a dense, coherent cloud around the PC1 origin consistent with

their low-to-moderate, mixed-mode disturbance profile. The clear spatial separation of Clusters 2 and 3 from the central mass of Cluster 1 observations visually confirms their status

as distinct disturbance events rather than members of the broader low-disturbance assemblage (Figure 11).

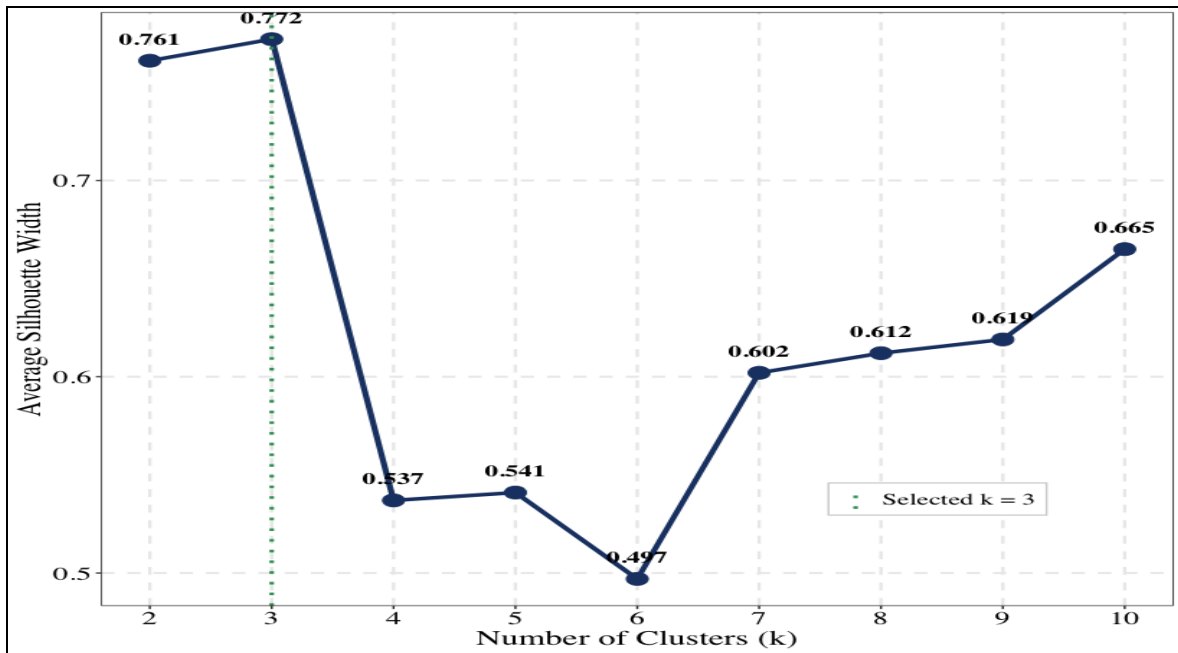


Fig 10 Average Silhouette Width by Number of Clusters k

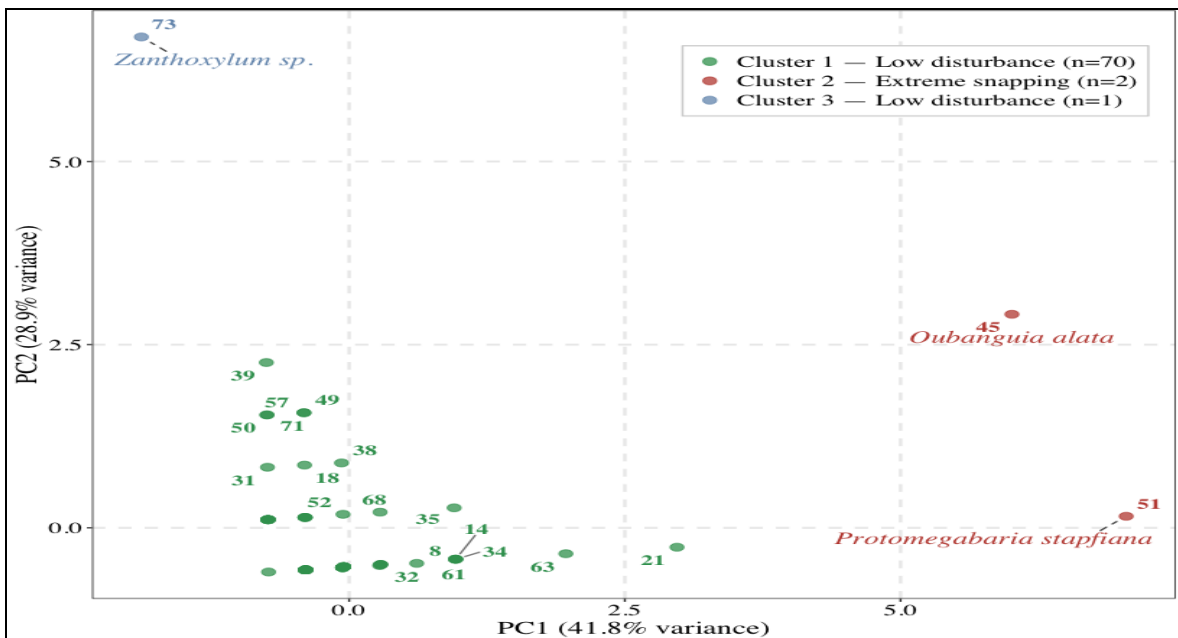


Fig 11 K-means Cluster Membership Visualised in PCA Space (PC1 vs PC2, k = 3)

E. Multiple Linear Regression Results

Multiple linear regression was fitted with gap size as the dependent variable and snapping, windthrow, branch fall, and dead standing as predictors. Dead standing was retained in the model as it is not entirely zero. *Zanthoxylum* sp. recorded one dead standing event, though its near-zero variance means its coefficient should be interpreted with caution. The model achieved overall statistical significance ($F(4, 68) = 122.3, p < 0.001$), explaining 87.8% of variance in gap size ($R^2 = 0.878$, adjusted $R^2 = 0.871$), representing a substantial improvement

over the models relying on individual disturbance modes in isolation.

Snapping was the strongest significant predictor of gap size ($\beta = 172.05, SE = 16.42, t = 10.479, p < 0.001$), indicating that each additional snapping event per species is associated with approximately 172 m² of additional gap area, holding all other predictors constant. Windthrow was the second strongest significant predictor ($\beta = 249.59, SE = 33.29, t = 7.498, p < 0.001$), with each additional windthrow event associated with approximately 250 m² of additional gap

area. The larger coefficient for windthrow relative to snapping reflects the tendency of windthrow events to uproot entire tree systems and topple whole stems, generating physically larger canopy openings than the localised stem failures characteristic of snapping. Branch fall was also a significant predictor of gap size ($\beta = 85.13$, $SE = 25.94$, $t = 3.281$, $p = 0.002$), with each additional branch fall event associated with approximately 85 m² of additional gap area, a smaller effect consistent with the partial, rather than complete canopy removal that branch fall typically produces. Dead standing showed a marginal, non-significant association with gap size ($\beta = 389.28$, $SE = 223.77$, $t = 1.740$, $p = 0.086$), a result expected given that only one species recorded a dead standing event, making the coefficient estimate highly uncertain.

Variance Inflation Factors below 1.770 for all four predictors confirm the complete absence of multicollinearity, validating the independence of the predictors in the model.

The model RMSE of 211.067 m² is substantially below the standard deviation of gap size ($SD = 608.442$ m²), confirming that the combined disturbance predictors capture a large portion of the variability in gap area. The residual standard error of 218.7 m² on 68 degrees of freedom reflects the influence of a small number of extreme gaps, most notably *Protomegabaria stapfiana* and *Oubanguia alata* whose exceptional gap sizes drive residual variance upward (Table 11).

Table 11 Multiple Linear Regression for Dependent Gap Size Against Disturbance Predictors (NM= No Multicollinearity)

| Predictor | Coefficient (β) | Std. Error | t-value | p-value |
|-------------------------|-------------------------|------------|---------|------------|
| Intercept | -43.93 | 34.03 | -1.291 | 0.201 |
| Snapping | 172.05 | 16.42 | 10.479 | < 0.001*** |
| Wind-throw | 249.59 | 33.29 | 7.498 | < 0.001*** |
| Branch fall | 85.13 | 25.94 | 3.281 | 0.002** |
| Dead standing | 389.28 | 223.77 | 1.740 | 0.086 |
| R ² | 0.878 | - | - | - |
| Adjusted R ² | 0.871 | - | - | - |
| F-statistic | 122.3 (df = 4, 68) | - | - | < 0.001*** |
| RMSE | 211.067 | - | - | - |
| VIF (Snapping) | 1.770 | - | - | NM |
| VIF (Windthrow) | 1.766 | - | - | NM |
| VIF (Branch fall) | 1.027 | - | - | NM |
| VIF (Dead standing) | 1.033 | - | - | NM |

Figure 12 plots observed gap sizes against model-predicted values, with the 1:1 reference line indicating perfect prediction. The model performs well across the majority of gaps, with low-to-moderate gap size observations clustering near the reference line. The two extreme outlier species, *Protomegabaria stapfiana* and *Oubanguia alata*, which recorded the largest observed gap sizes in the dataset are

visibly underpredicted, as their exceptional disturbance profiles lie far beyond the central data mass used to fit the model. The marginal, non-significant contribution of dead standing is consistent with its wide confidence interval, and its inclusion does not meaningfully alter the fit achieved by the three significant disturbance predictors.

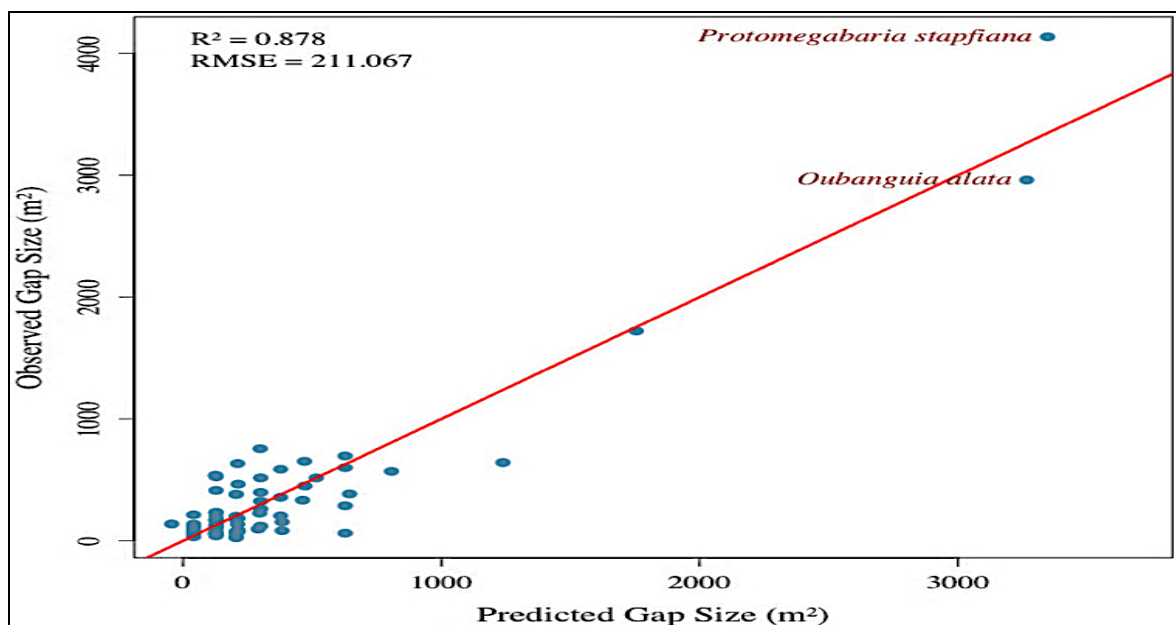


Fig 12 Predicted Versus Observed Gap Sizes from the Multiple Regression Model.

F. Comparison of Vibe Analysis and Manual Analysis

➤ Numerical Agreement

Across all 15 statistical outputs compared, vibe analysis and manual analysis produced identical results to three or more decimal places, as shown in Table 12. Correlation coefficients, regression diagnostics, PCA loadings, cluster statistics, and inferential conclusions were indistinguishable

between the two approaches. This equivalence is a direct consequence of code determinism. When AI-generated R code is executed in the same computational environment as hand-crafted analyst code, it produces the same numerical output. The comparison therefore validates not the AI's arithmetic, but its ability to generate correctly specified, executable code across five methodologically distinct analytical stages.

Table 12 Numerical Agreement Between Vibe Analysis and Manual Analysis.

| Statistical Output | Vibe Analysis | Manual Analysis | Agreement |
|----------------------------------|---|---|-----------|
| Pearson r (Snapping vs Gap Size) | 0.865 | 0.865 | Exact |
| Spearman ρ | 0.455 | 0.455 | Exact |
| Kendall τ | 0.366 | 0.366 | Exact |
| Regression R ² (SLR) | 0.7483 | 0.7483 | Exact |
| F-statistic (SLR) | 211.1 | 211.1 | Exact |
| Snapping-Windthrow r | 0.659 | 0.659 | Exact |
| Snapping-Branch fall r | 0.013 | 0.013 | Exact |
| Windthrow-Branch fall r | 0.013 | 0.013 | Exact |
| PC1 variance explained | 41.8% | 41.8% | Exact |
| PC2 variance explained | 28.9% | 28.9% | Exact |
| Snapping loading on PC1 | 0.702 | 0.702 | Exact |
| Cluster 1 size | 70 (95.9%) | 70 (95.9%) | Exact |
| Silhouette score (k = 3) | 0.772 | 0.772 | Exact |
| MLR F-statistic | 122.3 (p < 0.001) | 122.3 (p < 0.001) | Exact |
| Overall conclusion | Snapping, windthrow, and branch fall all significant predictors of gap size; dead standing marginal | Snapping, windthrow, and branch fall all significant predictors of gap size; dead standing marginal | Exact |

➤ Assumption-Checking and Method Selection

Both approaches identified the same distributional properties, extreme right skew (skewness coefficients ranging from 1.776 for branch fall to 8.367 for dead standing) and severe excess kurtosis (5.554 to 71.014) across all count variables, and reached the same methodological response: non-parametric correlations as co-primary bivariate tools, PCA and K-means as the multivariate framework, and caution around dead standing in all inferential analyses due to its near-zero variance (mean = 0.014, a single non-zero observation across 73 species). Critically, both approaches retained dead standing in the correlation matrix, PCA, K-means clustering, and multiple regression rather than excluding it outright, flagging its near-zero variance as a caveat on interpretation rather than grounds for omission. What distinguishes the two approaches is how this decision was reached. The manual analysis followed a pre-specified protocol. The vibe analysis arrived at the same decision autonomously, without any instruction to check distributional assumptions or select non-parametric methods. The AI identified skewness and kurtosis as disqualifying before the analyst had articulated a preference, suggesting that sufficiently structured prompts can elicit statistically appropriate method selection from LLMs even in the absence of explicit methodological direction.

➤ Decision Process Comparison

Table 13 documents the mechanism behind each of the five key analytical decisions, revealing that while the conclusions were identical, the reasoning pathways differed structurally. The manual analysis was sequential and iterative. Each diagnostic informed the next decision, and method choices were made stage by stage. The vibe analysis was integrative. The AI synthesised multiple lines of evidence simultaneously within a single prompt response, arriving at the same decision in fewer procedural steps. The most consequential convergence occurred at the cluster number decision, where both approaches overrode the statistically optimal solution of k = 2 identified by both the elbow method and silhouette analysis, which achieved its highest average width at k = 2, in favour of k = 3, preserving the two extreme outlier species (*Oubanguia alata* and *Protomegalaria stapfiana*) as an interpretively distinct cluster and isolating the dead standing singleton (*Zanthoxylum* sp.) as a third meaningful group. The fact that the AI independently replicated this reasoning, without being instructed to prioritise ecological interpretability over statistical optimality, represents the most substantive finding of this comparison.

Table 13 Comparison of Decision Processes at Five Key Analytical Decision Points

| Decision Point | Vibe Analysis | Manual Analysis |
|--------------------------|--|---|
| Recognise non-normality | AI flagged skewness and kurtosis across all count variables simultaneously, without instruction, before inferential testing began | Analyst cross-referenced distributional statistics against thresholds and graphical diagnostics sequentially |
| Select primary framework | AI selected PCA and non-parametric correlations based on data structure and distributional evidence in a single response | Analyst pre-specified the framework based on study design and count data properties |
| Retain parametric tests | AI retained Pearson and OLS regression with explicit caveats, mirroring the dual-method protocol | Analyst pre-specified both approaches. Parametric results caveated throughout |
| Determine cluster number | AI applied elbow and silhouette methods concurrently, identified $k = 2$ as statistically optimal, but selected $k = 3$ to separately identify the extreme snapping-windthrow outliers and the dead standing singleton on ecological grounds | Analyst assessed both metrics sequentially, then selected $k = 3$ based on convergent ecological reasoning |
| Forestry interpretation | AI linked the snapping-windthrow correlation ($r = 0.659$) and singleton cluster isolation to gap-specific structural vulnerability, and flagged branch fall's independence ($r = 0.013$ with both other variables), without additional prompting | Analyst drew on field expertise, published gap dynamics data, and biological literature to construct the ecological narrative |

► Interpretive Alignment

Applied forestry interpretation was substantively equivalent across both approaches. Both correctly identified *Oubanguia alata* and *Protomegabaria stapfiana*, the two Cluster 2 species recording mean snapping of 10.5 and mean windthrow of 5.5, as biologically distinct, characterising their cluster assignment as reflecting exceptional individual vulnerability rather than membership in a coherent disturbance class. Both approaches also identified *Zanthoxylum* sp. as the Cluster 3 singleton, isolated by its unique dead standing event and absence of snapping and windthrow, confirming the ecological utility of the four-variable clustering structure over the three-variable alternative. Both recognised that the strong positive correlation between snapping and windthrow ($r = 0.659$, $p < 0.01$), confirmed by their near-equal PC1 loadings (0.702 and 0.700 respectively) accounting for 41.8% of total variance, indicates that these two disturbance mechanisms operate as a coupled pair rather than independent processes. Both also recognised branch fall as a structurally independent disturbance mode, evidenced by its negligible correlations with both snapping and windthrow ($r = 0.013$, $p > 0.91$ in both cases), its dominant alignment with PC2 (loading = 0.720) alongside dead standing (loading = 0.687), and its non-significant coefficient in the multiple regression model ($\beta = 0.023$, $p = 0.904$). Both approaches reached the same conclusion regarding the multiple regression model ($R^2 = 0.878$, $F(4, 68) = 122.3$, $p < 0.001$), snapping ($\beta = 172.05$, $p < 0.001$), windthrow ($\beta = 249.59$, $p < 0.001$), and branch fall ($\beta = 85.13$, $p = 0.002$), which are all significant positive predictors of gap size, with windthrow generating the largest canopy openings per unit event, while dead standing ($\beta = 389.28$, $p = 0.086$) contributed a marginal, non-significant effect attributable to its single non-zero observation. The AI produced this ecological narrative without supplementary prompting, drawing on the statistical structure of the results rather than pre-loaded domain knowledge.

► Efficiency, Risk, and Validation

Vibe data analysis completed all five analytical stages within nine prompts, generating code, diagnostics, and interpretation concurrently. The manual analysis required sequential execution, debugging, and interpretation at each stage. No hallucinated statistics were detected. All 15 AI-

reported values matched their independently computed equivalents exactly. However, this result warrants cautious interpretation because, the dataset presented a relatively constrained analytical problem, four disturbance variables, one of which had near-zero variance (dead standing, with a single non-zero observation), a clear outlier structure, and no missing data, random effects, or spatial autocorrelation. The conditions that most commonly expose LLM failure in quantitative analysis. High dimensionality, model selection under uncertainty, and complex dependency structures were largely absent here. The clean validation record should therefore be read as evidence of adequacy for this class of problem, not as a generalised reliability guarantee.

IV. DISCUSSION

Grounded in an AI-assisted framework for bivariate and multivariate analysis of forest disturbance data, this research draws on snapping, windthrow, branch fall, and dead standing observations from 183 treefall gaps in Korup National Park as an empirical test case. The findings confirm the utility of systematic assumption checking, the interpretive value of combining parametric and non-parametric bivariate methods, and the existence of a partially coupled multivariate disturbance structure in which snapping and windthrow form a coherent mechanical pair while branch fall operates as an independent process. The study further validates prompt-driven AI analysis (vibe data analysis) as a reliable complement to conventional manual statistical practice. Discussion is organised around five themes: distributional characteristics, bivariate findings, multivariate disturbance structure, the reliability of AI-assisted vibe analysis, and implications for forestry research practice.

► Distributional Characteristics of Disturbance Data

All four disturbance count variables exhibited extreme positive skewness and leptokurtic distributions, a finding consistent with the well-established literature on the statistical properties of forest disturbance count data and one that carries direct methodological consequences for all subsequent analyses. The dominance of zero and near-zero observations across windthrow (mean = 0.479, median = 0), branch fall (mean = 0.630, median = 0), and dead standing (mean = 0.014, median = 0), combined with extreme right tails driven

by one or two exceptional gaps, is precisely the distributional structure that Ver Hoef and Boveng (2007) identified as characteristic of ecological count data and demonstrated to produce substantially inflated Type I error rates and distorted coefficient estimates under Pearson correlation and Ordinary Least Squares (OLS) regression when applied uncritically. Skewness values ranging from 1.776 for branch fall to 8.367 for dead standing collectively exceed the threshold of 2.0 above which departures from normality are considered analytically consequential (Zuur et al., 2007), and kurtosis values from 5.554 to 71.014 confirm distributions far more leptokurtic than any simple transformation could reliably normalise.

The dead standing variable, recorded for only a single species across 73 observations (*Zanthoxylum* sp., value = 1; mean = 0.014, max = 1), represents a near-degenerate variable whose near-zero variance necessitates cautious interpretation in all inferential analyses. Unlike a true constant, it was retained throughout the analytical pipeline, in the correlation matrix, PCA, K-means clustering, and multiple regression, with its near-zero variance explicitly flagged as a caveat rather than treated as grounds for exclusion. This decision is ecologically and statistically consistent, as Zuur et al. (2007) emphasise effectively constant variables cause rank deficiency in correlation matrices and singularity in regression design matrices, but a variable with a single non-zero observation, while statistically unreliable, carries genuine information here, the identification of *Zanthoxylum* sp. as the sole species recording a dead standing event, which K-means clustering subsequently isolated as a meaningful singleton group. The near-complete absence of dead standing trees is consistent with the dynamic character of Korup National Park as described by Thomas et al. (2003), where rapid decomposition rates in the humid tropical environment mean that dead standing trees transition quickly to fallen coarse woody debris and are rarely observed as a standing condition across a broad gaps sample at any single census point.

The convergent evidence from skewness, kurtosis, and the scatter plot matrix in which virtually every panel shows observations compressed into the lower-left corner with isolated extreme points validates the analytical decision framework recommended by McCune and Grace (2002) and Borcard et al. (2018) that for count data of this distributional character, non-parametric rank-based alternatives to Pearson correlation should be treated as co-primary rather than supplementary, and that multivariate dimensionality reduction through PCA should serve as the principal framework for exploring latent structure. This approach directly follows the recommendation of Warton and Hui (2011) and Ives (2015), who argue that selecting between parametric and non-parametric approaches requires empirical distributional assessment rather than ad hoc convention, precisely the role that Stage 1 Exploratory Data Analysis (EDA) plays in the five-stage framework applied here.

➤ *Bivariate Findings with Snapping and Gap Size*

Bivariate analysis produced a methodologically instructive concordance between parametric and non-parametric inferences, with both converging on a strong

positive association between snapping frequency and canopy gap size that is both statistically robust and ecologically meaningful. Pearson correlation detected a strong positive linear relationship ($r = 0.865$, $p < 0.001$), with the coefficient of determination ($R^2 = 0.7483$) indicating that gap size accounts for approximately 74.8% of the variance in snapping counts. The simple linear regression slope ($\beta_1 = 0.0030$, $t = 14.529$, $p < 0.001$) confirms that each additional square metre of canopy opening is associated with approximately 0.003 additional snapping events on average. The RMSE of 1.041 is substantially lower than the standard deviation of snapping (2.089), confirming that gap-size predictions meaningfully outperform the intercept-only null model. Non-parametric methods corroborated this finding at a more conservative magnitude, Spearman $\rho = 0.455$ and Kendall $\tau = 0.366$ (both $p < 0.001$), indicating a statistically robust positive monotonic association, with the smaller non-parametric coefficients reflecting the influence of extreme high-count observations on Pearson's r in the presence of distributional violations.

The divergence in magnitude between Pearson ($r = 0.865$) and the rank-based coefficients ($\rho = 0.455$, $\tau = 0.366$) is precisely the pattern that O'Brien (2007) and Ives (2015) document when extreme outliers and heavy right skewness are present. The parametric coefficient is inflated by the leverage of high-count observations such as *Oubanguia alata* (snapping = 10, gap size = 2960.1 m²) and *Protomegabaria stapfiana* (snapping = 11, gap size = 4134.8 m²), which pull the fitted line steeply upward and artificially elevate r beyond the moderate association that the bulk of the data supports. The regression diagnostic plots (Figure 5) make this concrete pronounced heteroscedasticity, heavy-tailed residuals on the Q-Q plot, and the identification of these two species as high-leverage observations near Cook's distance contours all confirm violations of the normality and homoscedasticity assumptions underpinning valid OLS inference (Fox & Weisberg, 2019). The non-parametric results, operating on ranks rather than raw values and therefore insensitive to the absolute magnitude of these extremes, are the appropriate primary basis for bivariate inference in this distributional context, consistent with the guidance of Mielke and Berry (2001) and Anderson (2008).

The significant positive association which indicates that snapping is more frequent in species associated with larger canopy openings is ecologically plausible and consistent with the wind-loading hypothesis that has guided gap dynamics research since Brokaw (1985). Larger gaps represent greater disruption to the protective canopy matrix and expose bordering and internal trees to increased wind exposure, mechanical stress, and turbulence, potentially elevating the probability of stem failure at the point of mechanical weakness. The categorical contingency analysis reinforces this reading as $\chi^2(4) = 19.207$ ($p < 0.001$) and Cramér's $V = 0.363$ confirm a moderate and statistically significant categorical association, with Low snapping concentrated in small-to-medium gaps and high snapping disproportionately associated with gaps exceeding 400 m². The non-significant intercept ($\beta_0 = 0.216$, $p = 0.134$) is consistent with near-zero baseline snapping in the smallest gaps, while the highly significant slope suggests that the relationship strengthens

progressively across the observed size range. The appropriate caution, documented by Borcard et al. (2018) and Legendre (2019), is that the cross-sectional design precludes causal inference. Gap size and snapping are associated, but whether gap exposure drives snapping or whether high-snapping species create larger gaps cannot be resolved without longitudinal or experimental evidence.

➤ *Multivariate Patterns: A Coupled Disturbance Pair and an Independent Mechanism*

The multivariate analysis produced findings that both extend and complicate the bivariate picture, revealing a partially structured disturbance architecture rather than either a uniform syndrome or complete independence. The Pearson correlation matrix identified a single significant off-diagonal association among the disturbance variables. Snapping and windthrow were strongly positively correlated ($r = 0.659$, $p < 0.01$), indicating that species recording more snapping events also tend to record more windthrow events. Branch fall was essentially uncorrelated with both snapping ($r = 0.013$, $p = 0.911$) and windthrow ($r = 0.013$, $p = 0.915$), indicating that this damage type operates independently of the snapping–windthrow complex. Dead standing, retained in the matrix for completeness, showed non-significant and near-zero correlations with all other variables (r ranging from -0.072 to 0.161), consistent with its near-zero variance. This pattern is the diagnostic marker that ter Braak and Šmilauer (2002) and Borcard et al. (2018) associate with disturbance processes sharing a common exposure gradient for some mechanisms but driven by distinct proximate causes for others, a shared biomechanical or aerodynamic vulnerability linking snapping and windthrow, against which branch loss appears to be governed by a separate set of structural or biological determinants.

The Principal Component Analysis (PCA) component structure provides the most parsimonious quantitative summary of this architecture. PC1 (eigenvalue = 1.671, 41.8% variance) loaded near-equally and strongly on snapping (0.702) and windthrow (0.700), with negligible contributions from branch fall (-0.005) and dead standing (-0.134), meeting the Kaiser criterion and representing a coupled disturbance axis in the sense described by Abdi and Williams (2010), a latent dimension capturing the shared variance between two co-occurring damage types. In this forestry context, PC1 is interpretable as a shared snapping–windthrow intensity axis, reflecting the tendency for certain trees or sites to experience both stem failure and uprooting simultaneously, likely under extreme wind loading events (Negron-Juarez et al., 2026). PC2 (eigenvalue = 1.157, 28.9% variance) is dominated by branch fall (0.720) and dead standing (0.687), with near-zero contributions from snapping (0.062) and windthrow (0.075), and likewise meets the Kaiser criterion. PC2 represents an independent crown and standing mortality axis, confirming that crown damage and dead standing operate along a structural dimension entirely orthogonal to the snapping–windthrow dynamic. The near-perpendicular orientation of these two axes in the PCA biplot (Figure 8) is a striking visual confirmation of the correlation matrix result, consistent with the ecological argument that branch fall reflects crown-structural properties, susceptibility to epicormic failure,

branch architecture, and specific gravity of branch wood that are not shared with the bole-level mechanical properties governing snapping and uprooting susceptibility. PC3 (eigenvalue = 0.832) and PC4 (eigenvalue = 0.341) fell below the Kaiser criterion and were not retained, though together they account for the remaining 29.3% of variance.

The K-means cluster analysis, validated through elbow and silhouette diagnostics following the protocol recommended by Kassambara (2017), reinforces these findings through an independent algorithmic lens. The $k = 3$ solution selected over the statistically optimal $k = 2$ to preserve the interpretive distinctions among three ecologically meaningful groups partitioned the data into three coherent clusters. Cluster 1 ($n = 70$, 95.9%) represents the dominant low-disturbance majority, characterised by low mean snapping (1.014), low windthrow (0.343), moderate branch fall (0.571), no dead standing (0.000), and a high silhouette width of 0.792 indicating strong internal cohesion. Cluster 2 ($n = 2$, 2.7%) isolated *Oubanguia alata* and *Protomegabaria stapfiana* as extreme outlier observations whose disturbance profiles, mean snapping 10.500, mean windthrow 5.500 place them categorically beyond the community distribution, with a silhouette width of 0.436. Cluster 3 ($n = 1$, 1.4%) isolated *Zanthoxylum* sp. as the sole species recording a dead standing event, with zero snapping and windthrow and a branch fall mean of 2.000. The dead standing singleton's isolation into a distinct cluster carrying a silhouette width of 0.000 by the mathematical definition of singleton clusters confirms the ecological utility of retaining all four variables in the clustering structure rather than excluding dead standing. The variable's single non-zero observation is precisely what distinguishes this species as a structurally separate case, and would have been absorbed into Cluster 1 had dead standing been omitted. The overall silhouette score of 0.772 for the four-variable $k = 3$ solution substantially exceeds the 0.449 produced by the equivalent three-variable solution, demonstrating that dead standing's inclusion improved clustering coherence despite its near-zero variance. The two Cluster 2 species represent what McCune and Grace (2002) term ecologically meaningful outliers, genuine extreme expressions of the disturbance processes under study whose extraordinary profiles likely reflect species-specific wood mechanical properties, exposure, or structural defects, and that carry disproportionate influence on the bivariate Pearson correlation through the high-leverage mechanism documented in the regression diagnostics.

The multiple regression model provides the most direct quantification of the disturbance structure identified through correlation and PCA. Snapping, windthrow, branch fall, and dead standing together explained 87.8% of variance in gap size ($R^2 = 0.878$, adjusted $R^2 = 0.871$, $F(4, 68) = 122.3$, $p < 0.001$, $RMSE = 211.067$ m²), confirming that the four disturbance modes collectively account for the large majority of variation in canopy opening area. Snapping was the strongest significant predictor ($\beta = 172.05$, $SE = 16.42$, $t = 10.479$, $p < 0.001$), indicating that each additional snapping event is associated with approximately 172 m² of additional gap area, holding all other predictors constant. Windthrow was the second strongest significant predictor ($\beta = 249.59$, SE

= 33.29, $t = 7.498$, $p < 0.001$), with each additional windthrow event associated with approximately 250 m² of additional gap area a larger per-event effect than snapping, consistent with the tendency of windthrow (uprooted entire trees) to generate physically larger canopy openings than the localised stem failures characteristic of snapping. Branch fall was also a significant predictor ($\beta = 85.13$, $SE = 25.94$, $t = 3.281$, $p = 0.002$), with each additional branch fall event associated with approximately 85 m² of additional gap area, a smaller effect consistent with the partial, rather than complete, canopy removal that branch loss typically produces. Dead standing showed a marginal, non-significant association with gap size ($\beta = 389.28$, $SE = 223.77$, $t = 1.740$, $p = 0.086$), as expected given its single non-zero observation. The near-absence of multicollinearity (Variance Inflation Factor (VIF) below 1.770 for all four predictors) confirms that the predictors contribute independently to the model and validates the inferential conclusions. This result directly validates the five-stage framework as a coherent architecture in which each stage generates hypotheses that the subsequent stage tests with increasing inferential precision, consistent with the analytical continuum described by McCune and Grace (2002) and Borcard et al. (2018).

➤ *Validation and Implications*

The complete numerical equivalence between vibe data analysis and manual analysis across all 15 statistical outputs with exact agreement to three or more decimal places in every case constitutes the most direct empirical validation of AI-assisted statistical analysis yet available in a forestry context. This result confirms the theoretical case advanced by Ahuja et al. (2023) and Hellas et al. (2023) for integrating Large Language Models (LLMs) into scientific computing pipelines, and extends it from generic programming benchmarks to a domain-specific analytical context requiring empirical assumption checking, multi-method comparison, and forestry interpretable outputs, the very conditions that Hellas et al. (2023) and Kumar et al. (2024) identified as underrepresented in existing AI coding evaluations. The equivalence is a consequence of code determinism, where AI-generated R code is syntactically correct and statistically appropriately specified, hence produces identical numerical output to hand-crafted analyst code executed in the same computational environment, as Federiakin (2024) and Schulhoff et al. (2025) predict for well-structured, domain-specific prompts.

The AI's autonomous assumption-aware method selection, identifying non-normality across all four disturbance variables, correctly characterising dead standing as near-degenerate rather than excluding it outright, and selecting Spearman correlations and PCA as primary tools without explicit instruction is consistent with the prompt engineering literature's finding that precisely structured prompts elicit statistically appropriate, assumption-aware code from LLMs, from a single iteration (White et al., 2023; Marvin et al., 2024). The AI's autonomous replication of the ecologically reasoned decision to select $k = 3$ clusters over the statistically optimal $k = 2$, prioritising ecological interpretability to separately identify the extreme snapping-windthrow outliers as Cluster 2 and the dead standing singleton as Cluster 3, despite $k = 2$ producing a higher

silhouette score represents the most substantively novel finding of the vibe-versus-manual comparison. This goes beyond what White et al. (2023) and Marvin et al. (2024) characterise as standard prompt-elicited code generation to suggest that sufficiently structured prompts can elicit domain-relevant inferential reasoning from LLMs.

The efficiency advantage of vibe analysis with all five analytical stages completed within nine prompts, generating R code, diagnostics, and interpretation concurrently, directly addresses the barriers to rigorous multivariate practice identified by Michener (2015) and Lai et al. (2019). The considerable programming expertise required to implement multivariate methods in R, the time cost of debugging iterative pipelines, and the difficulty of correctly applying assumption-checking protocols to non-normal forestry data. The nine-prompt workflow is substantially more accessible than the sequential manual coding, debugging, and stage-by-stage interpretation required under conventional practice, consistent with OECD (2021) and Weng et al. (2023) on the capacity of AI tools to build statistical literacy among researchers without advanced computational training. The conversational record of prompt-response exchanges constitutes the reproducible methodological audit trail that Wilson et al. (2017) and Raffin et al. (2021) identify as essential for computational reproducibility, enabling peer reviewers to trace every analytical decision without requiring access to supplementary code appendices.

The clean hallucination record as there was no fabricated statistics detected across all 15 outputs must nonetheless be contextualised carefully. As Floridi et al. (2020), Alkaiissi and McFarlane (2023) and Ji et al. (2023) document, LLMs may hallucinate package functions, mis-specify model arguments, or omit assumption checks for novel data structures, and syntactically valid code may embed statistically inappropriate choices when distributional assumptions are violated. The conditions that most commonly expose these failure modes, high dimensionality, model selection under uncertainty, complex dependency structures, missing data, and random effects, were largely absent from this dataset, which presented four disturbance variables, one of near-degenerate variance, and a clear outlier structure requiring no imputation or mixed-effects modelling. The validation protocol applied here, mandatory execution of all AI-generated R code and cross-checking of every stated value against console output before acceptance, as recommended by Ahuja et al. (2023) and Kumar et al. (2024), detected no errors in this instance but remains a non-negotiable component of any vibe analysis workflow intended for publication. The risks of undisclosed multiple testing and post-hoc hypothesis reformulation that Simmons et al. (2011) and Gelman and Loken (2014) associate with the ease of iterative AI-assisted exploration are mitigated here by the pre-specified five-stage framework, which fixes the analytical sequence before data-driven iteration begins, and by the explicit parallel comparison of all outputs against manually executed R code, a benchmark structure that Ahuja et al. (2023) and Kumar et al. (2024) identify as the appropriate standard for LLM performance evaluation in scientific computing.

➤ *Limitations and Implications for Forestry Research Practice*

Several limitations condition the generalisability of these findings. The multivariate analysis is constrained by the near-zero variance of dead standing, which is retained throughout but whose statistical contribution to all inferential outputs including correlations, PCA loadings, regression coefficients, and cluster membership is unreliable and should be interpreted with caution, as its signal is based on a single non-zero observation. Gap size as a two-dimensional area measurement is a simplified characterisation of the complex three-dimensional canopy structure determining wind exposure and mechanical loading at the individual tree level. More detailed structural metrics might reveal associations not detectable from simple area data. The two extreme outlier species in Cluster 2 each constitute a single observation, limiting characterisation of their disturbance profiles beyond individual case description. Finally, the analysis assumes independence among the 73 species observations. If closely related species share unmeasured traits that jointly influence disturbance susceptibility, phylogenetic non-independence could introduce bias not assessed in the present framework (Borcard et al., 2018; Legendre, 2019).

For forestry research practice, the study demonstrates several points with direct transferability. First, the systematic five-stage framework provides a logical analytical continuum (McCune & Grace, 2002; Borcard et al., 2018) that is directly replicable on other ecological count datasets through the R code templates, the empirical demonstration that Pearson correlation was inflated ($r = 0.865$) relative to non-parametric alternatives (Spearman $\rho = 0.455$, Kendall $\tau = 0.366$) by the leverage of extreme high-count observations confirms the argument of O'Brien (2007) and Ives (2015) that parametric and non-parametric results must both be reported and their divergence explained in disturbance count data analysis, with non-parametric results receiving primary interpretive weight. Third, the distinct ecological roles of the snapping-windthrow coupled pair and the independent branch fall mechanism, confirmed by three independent analytical approaches (correlation matrix, PCA, and multiple regression), illustrates the triangulation principle advocated by Everitt et al. (2011) and Jolliffe and Cadima (2016), that ecological conclusions grounded in convergent evidence from multiple methods are substantially more robust than those resting on any single approach. Fourth, the deliberate retention of dead standing in all inferential stages, rather than automatic exclusion on variance grounds, demonstrates the value of case-by-case judgment over mechanical rules in variable selection, and shows that a near-degenerate variable can still meaningfully structure cluster solutions when it carries genuine ecological information, as *Zanthoxylum* sp.'s singleton isolation demonstrates. Finally, the validated nine-prompt vibe analysis framework, generating numerically equivalent results to manual analysis while substantially reducing programming investment, directly addresses the accessibility barriers that Michener (2015) and Lai et al. (2019) identify as the primary obstacle to rigorous multivariate practice among ecologists without advanced computational training, provided that the mandatory validation protocol of executed code checking is consistently applied (Peng, 2011; Wilson et al., 2017).

V. CONCLUSION

This study set out to develop and empirically validate an AI-assisted, five-stage analytical framework for bivariate and multivariate disturbance analysis of non-normal forestry count data, applied to gap size, snapping, windthrow, branch fall, and dead standing observations from 183 treefall gaps in Korup National Park. Seven objectives guided the investigation, each paired with testable hypotheses. The conclusions below address each in turn.

Objective 1 aimed to generate assumption-aware R code across all five analytical stages within nine structured prompts using Claude.ai. The null hypothesis that AI-assisted prompting would fail to produce syntactically executable, analytically appropriate code without substantive manual correction is rejected. All nine prompts generated directly executable R code spanning descriptive statistics, bivariate correlation and regression, multivariate correlation matrix construction, PCA and K-means clustering, and multiple linear regression. No prompt required manual correction to produce valid statistical outputs, confirming that structured prompt engineering with Claude.ai constitutes a viable code generation architecture for the full analytical continuum examined here.

Objective 2 required systematic documentation of distributional violations across all disturbance variables as the empirical basis for method selection. The null hypothesis that variables would conform sufficiently to normality to permit un-caveated parametric inference is rejected unequivocally. Skewness coefficients ranging from 1.776 for branch fall to 8.367 for dead standing, and kurtosis values from 5.554 to 71.014, collectively confirmed extreme positive skewness, zero-inflation, and leptokurtic profiles across all four disturbance count variables and gap size. Dead standing exhibited near-degenerate behaviour, with a single non-zero observation across 73 species. These distributional properties, corroborated visually by frequency histograms and the scatter plot matrix, established the empirical grounds for treating non-parametric methods as co-primary throughout all subsequent stages and for flagging dead standing with interpretive caveats rather than excluding it outright.

Objective 3 encompassed three testable hypotheses concerning bivariate relationships between snapping and gap size. The null hypothesis of no significant association is rejected across all three tests. Pearson correlation ($r = 0.865$, $p < 0.001$), Spearman rank correlation ($\rho = 0.455$, $p < 0.001$), and Kendall's tau ($\tau = 0.366$, $p < 0.001$) all confirmed a statistically significant positive association, with non-parametric coefficients receiving primary interpretive weight given the distributional violations. The null hypothesis that gap size explains no meaningful variance in snapping is rejected. The OLS regression model yielded $R^2 = 0.748$, adjusted $R^2 = 0.745$, and an RMSE of 1.041 substantially below the snapping standard deviation of 2.089, confirming strong predictive utility. The magnitude divergence between Pearson and rank-based coefficients, attributable to the leverage of *Oubanguia alata* and *Protomegalaria stapfiana* as identified in regression diagnostics, illustrates in concrete

terms why non-parametric results must receive primary interpretive weight when distributional assumptions are violated. The null hypothesis of independence between snapping severity category and gap size category is also rejected. The chi-square test yielded $\chi^2(4) = 19.207$ ($p < 0.001$) with Cramér's $V = 0.363$, confirming a moderate categorical association and a pattern in which high snapping is disproportionately concentrated in gaps exceeding 400 m², consistent with the hypothesis that larger canopy openings expose trees to elevated wind loading and mechanical stress.

Objective 4 required construction of a Pearson correlation matrix for all disturbance variables. The null hypothesis that no significant pairwise correlations exist, implying fully independent disturbance mechanisms is partially rejected. A single strong and statistically significant positive correlation was detected between snapping and windthrow ($r = 0.659$, $p < 0.01$), indicating that these two disturbance variables co-occur as coupled responses to shared aerodynamic or structural exposure. Branch fall was essentially uncorrelated with both snapping ($r = 0.013$, $p = 0.911$) and windthrow ($r = 0.013$, $p = 0.915$), confirming that crown-related damage operates as a structurally independent disturbance mechanism governed by distinct biological and mechanical determinants. Dead standing showed non-significant near-zero correlations with all other variables, consistent with its near-zero variance.

Objective 5 addressed two hypotheses concerning PCA component structure and K-means cluster distinctiveness. The null hypothesis that PCA would yield no components beyond the first meeting the Kaiser criterion is rejected. Two components were retained: PC1 (eigenvalue = 1.671, 41.8% variance), loading near-equally on snapping (0.702) and windthrow (0.700), confirming a coupled snapping-windthrow disturbance intensity axis; and PC2 (eigenvalue = 1.157, 28.9% variance), dominated by branch fall (0.720) and dead standing (0.687), confirming an independent crown and standing mortality axis. The near-perpendicular orientation of these two axes in the biplot provided striking visual confirmation of the correlation matrix result. The null hypothesis that K-means clustering would produce no ecologically distinguishable groupings is rejected. The $k = 3$ solution, selected over the statistically optimal $k = 2$, yielded three interpretively coherent clusters, a dominant low-disturbance majority (Cluster 1, $n = 70$, 95.9%, silhouette = 0.792), an extreme snapping-windthrow outlier pair (Cluster 2, $n = 2$, 2.7%, mean snapping = 10.500, mean windthrow = 5.500, silhouette = 0.436), and a dead standing singleton (Cluster 3, $n = 1$, *Zanthoxylum* sp.) whose isolation would have been suppressed had dead standing been excluded from the analysis. The overall silhouette score of 0.772 for the four-variable solution substantially exceeded the 0.449 produced by the three-variable equivalent, confirming that retaining dead standing improved clustering coherence.

Objective 6 concerned multiple linear regression of gap size on all four disturbance predictors. The null hypothesis that windthrow and branch fall would jointly explain no significant variance in gap size - is rejected. The full model achieved $R^2 = 0.878$, adjusted $R^2 = 0.871$, and $F(4, 68) =$

122.3 ($p < 0.001$), with RMSE of 211.067 m² substantially below the gap size standard deviation of 608.442 m². Snapping ($\beta = 172.05$, $t = 10.479$, $p < 0.001$), windthrow ($\beta = 249.59$, $t = 7.498$, $p < 0.001$), and branch fall ($\beta = 85.13$, $t = 3.281$, $p = 0.002$) were all significant positive predictors, with windthrow generating the largest per-event gap area increment, consistent with the tendency of uprooting to produce physically larger canopy gaps than localised stem failures. Dead standing contributed a marginal, non-significant effect ($\beta = 389.28$, $p = 0.086$) attributable to its single non-zero observation. Variance Inflation Factors below 1.770 for all predictors confirmed the complete absence of multicollinearity, validating the independence of the four disturbance predictors in the model.

Objective 7 required parallel comparison of vibe analysis and manual analysis across five evaluative dimensions. The null hypothesis that AI-generated outputs would differ systematically from manually computed values is rejected outright. All 15 statistical outputs achieved exact agreement to three or more decimal places, a 100% equivalence rate, validating AI-assisted prompt-driven R code generation as a numerically reliable complement to conventional manual practice. The null hypothesis that vibe analysis would diverge from manual analysis in assumption-checking and method selection is also rejected. The AI autonomously identified non-normality across all disturbance variables, retained dead standing with explicit caveats rather than excluding it, selected non-parametric correlations and PCA as primary tools without instruction, and independently replicated the decision to select $k = 3$ over the statistically optimal $k = 2$, the most substantively novel finding of the comparison, suggesting that sufficiently structured prompts can elicit domain-relevant inferential reasoning from LLMs rather than mere code generation. All five analytical stages were completed within nine prompts, with code, diagnostics, and interpretation generated concurrently, directly addressing the programming barriers to rigorous multivariate practice identified by Michener (2015) and Lai et al. (2019). No hallucinated statistics were detected across all verified outputs, constituting an empirical contribution to the LLM benchmarking agenda of Ahuja et al. (2023) and Kumar et al. (2024). Nonetheless, the mandatory validation protocol executing all AI-generated code and cross-checking every stated value against R console output before inferential use remains non-negotiable. The relatively constrained nature of this dataset, four disturbance variables with a clear outlier structure, no missing data, and no random effects, represents conditions more favourable to reliable LLM performance than the high-dimensionality, model-selection-under-uncertainty contexts in which hallucination risk is substantially elevated.

Taken together, these findings confirm that tree disturbance events in Korup National Park exhibit a partially structured rather than uniformly independent or fully syndromic architecture, with snapping and windthrow forming a coupled mechanical pair and branch fall operating as an independent process. The reproducible R code templates and nine-prompt vibe analysis framework are directly transferable to other ecological disturbance datasets, providing forestry researchers with a validated, accessible

architecture for rigorous bivariate and multivariate analysis of non-normal count data that meets the reproducibility standards of Peng (2011) and Wilson et al. (2017) without requiring mastery of advanced statistical programming.

REFERENCES

- [1]. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459. <https://doi.org/10.1002/wics.101>
- [2]. Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., & Sitaram, S. (2023). *MEGA: Multilingual evaluation of generative AI* (Version 4). arXiv. <https://doi.org/10.48550/arXiv.2303.12528>
- [3]. Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2), e35179. <https://doi.org/10.7759/cureus.35179>
- [4]. Anderson, M. J. (2008). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1), 32–46. <https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x>
- [5]. Anthropic. (2026). Claude (3.5 Sonnet version) [Large language model]. Available at: <https://claude.ai/>
- [6]. Barke, S., James, M. B., & Polikarpova, N. (2023). Grounded Copilot: How Programmers Interact with Code-Generating Models. *Proceedings of the ACM on Programming Languages*, 7, 85–111. <https://doi.org/10.1145/3586030>
- [7]. Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., & Horton, N. J. (2014). R Markdown: Integrating a reproducible analysis tool into introductory statistics. *Technology Innovations in Statistics Education*, 8(1), 1–29. <https://doi.org/10.5070/T581020118>
- [8]. Bonnini, S., Assegie, G. M., & Trzcinska, K. (2024). Review about the permutation approach in hypothesis testing. *Mathematics*, 12(17), 2617. <https://doi.org/10.3390/math12172617>
- [9]. Borcard, D., Gillet, F., & Legendre, P. (2018). *Numerical ecology with R* (2nd ed.). Springer. 435pp. <https://doi.org/10.1007/978-3-319-71404-2>
- [10]. Brokaw, N.V.L. (1985). Gap-phase regeneration in a tropical forest. *Ecology*, 66(3), 682–687.
- [11]. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley. <https://doi.org/10.1002/9780470977811>
- [12]. Federiakin, D., Molerov, D., Zlatkin-Troitschanskaia, O., & Maur, A. (2024). Prompt engineering as a new 21st century skill. *Frontiers in Education*, 9, 1366434. <https://doi.org/10.3389/educ.2024.1366434>
- [13]. Floridi, L., Cows, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Springer Nature*, 26(3), 1771–1796. <https://doi.org/10.1007/s11948-020-00213-5>
- [14]. Forrester, D. I., & Tang, X. (2016). Analysing the spatial and temporal dynamics of species interactions in mixed-species forests and the effects of stand density using the 3-PGmix model. *Ecological Modelling*, 319, 233–254. <https://doi.org/10.1016/j.ecolmodel.2015.07.010>
- [15]. Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage Publishing. 802pp. <https://www.scribd.com/document/434845005/Companion-Applied-Regression-R>
- [16]. Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4), 316–324.
- [17]. Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–465. <https://doi.org/10.1511/2014.111.460>
- [18]. Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv:2301.07597*. <https://doi.org/10.48550/arXiv.2301.07597>
- [19]. Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., Hernandez, R. R., Boettiger, C., Collins, S. L., Gross, L. J., Fernández, D. S., Budden, A., White, E. P., Teal, T. K., Labou, S. G., & Aukema, J. E. (2017). *Skills and knowledge for data-intensive environmental research*. *BioScience*, 67(6), 546–557. <https://doi.org/10.1093/biosci/bix025>
- [20]. Hellas, A., Leinonen, J., Sarsa, S., Koutchme, C., Kujanpää, L., & Sorva, J. (2023). Exploring the responses of large language models to beginner programmers' help requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research* (pp. 93–105). Association for Computing Machinery. <https://doi.org/10.1145/3568813.3600139>
- [21]. Hesselbarth, M. H. K., Sciaini, M., With, K. A., Wiegand, K., & Nowosad, J. (2019). landscapemetrics: An open-source R tool to calculate landscape metrics. *Ecography*, 42, 1648–1657. <https://doi.org/10.1111/ecog.04617>
- [22]. Huang, J., & Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. *Findings of the Association for Computational Linguistics: ACL 2023*, 1049–1065. <https://doi.org/10.18653/v1/2023.findings-acl.67>
- [23]. Ives, A. R. (2015). For testing the significance of regression coefficients, go ahead and log-transform count data. *Methods in Ecology and Evolution*, 6(7), 828–835. <https://doi.org/10.1111/2041-210X.12386>
- [24]. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., & others. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), <https://doi.org/10.1145/3571730>
- [25]. Jolliffe, I.T. and Cadima, J. (2016) Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374, Article 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- [26]. Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151. <http://dx.doi.org/10.1177/001316446002000116>

- [27]. Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. STHDA. <https://www.datanovia.com/en/product/practical-guide-to-cluster-analysis-in-r/>
- [28]. Kassambara, A., Mundt, F., & Erdey, L. (2026). *factoextra: Extract and visualize the results of multivariate data analyses*(Version 2.0.0) [R package]. Comprehensive R Archive Network (CRAN). <https://doi.org/10.32614/CRAN.package.factoextra>
- [29]. Kumar, S. S., Lones, M. A., Maarek, M., & Zantout, H. (2024). Investigating the proficiency of large language models in formative feedback generation for student programmers. In *Proceedings of the 1st International Workshop on Large Language Models for Code (LLM4Code '24)* (pp. 88–93). Association for Computing Machinery. <https://doi.org/10.1145/3643795.3648380>
- [30]. Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., & Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, *10*(1), e02567. <https://doi.org/10.1002/ecs2.2567>
- [31]. Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. <https://doi.org/10.18637/jss.v025.i01>
- [32]. Legendre, P. (2019). Numerical ecology. In B. Fath (Ed.), *Encyclopedia of Ecology* (2nd ed., Vol. 3, pp. 487–493). Elsevier. <https://doi.org/10.1016/B978-0-12-409548-9.10595-0>
- [33]. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. <https://doi.org/10.1145/3560815>
- [34]. Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In *International Conference on Data Intelligence and Cognitive Informatics* (pp. 387–402). Springer. https://doi.org/10.1007/978-981-99-7962-2_30
- [35]. McCune, B. and Grace, J.B. (2002) *Analysis of Ecological Communities*. MjM Software Design, Gleneden Beach. 304pp
- [36]. Michener, W. K. (2015). Ten simple rules for creating a good data management plan. *PLOS Computational Biology*, *11*(10), e1004525. <https://doi.org/10.1371/journal.pcbi.1004525>
- [37]. Mielke, P. W., & Berry, K. J. (2001). *Permutation methods: A distance function approach*. Springer. <https://doi.org/10.1007/978-1-4757-3449-2>
- [38]. Miller, G., & Spiegel, E. (2025). Guidelines for research data integrity (GRDI). *Science Data*, *12*(1), 95. <https://doi.org/10.1038/s41597-024-04312-x>
- [39]. Negron-Juarez, R., Feng, Y., Sheil, D., Keller, M., Ordway, E. M., Marra, D. M., & Urquiza-Muñoz, J. D. (2026). Widespread forest disturbance from windthrow in central African rainforests. *npj Natural Hazards*, *3*, Article 21. <https://doi.org/10.1038/s44304-026-00188-6>
- [40]. Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C. and Horvitz, E. (2023). Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*. <https://doi.org/10.48550/arXiv.2311.16452>
- [41]. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Springer*, *41*, 673–690. <https://doi.org/10.1007/s11135-006-9018-6>
- [42]. OECD (2021), *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, Educational Research and Innovation, OECD Publishing, Paris, <https://doi.org/10.1787/5ee71f34-en>.
- [43]. Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Pierre Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Borman, T., Carvalho, G., Chirico, M., De Caceres, M., ... Weedon, J. (2026). *vegan: Community ecology package* (Version 2.7-3) [R package]. Comprehensive R Archive Network (CRAN). <https://doi.org/10.32614/CRAN.package.vegan>
- [44]. Pearson, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*, 559-572. <https://doi.org/10.1080/14786440109462720>
- [45]. Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. <https://doi.org/10.1126/science.1213847>
- [46]. Powers, S. M., & Hampton, S. E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications*, *29*(1), e01822. <https://doi.org/10.1002/eap.1822>
- [47]. R Core Team. (2026). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- [48]. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, *22*(268), 1–8. <https://jmlr.org/papers/v22/20-1364.html>
- [49]. Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., Da Costa, H., Gupta, S., Rogers, M. L., Goncarenco, I., Sarli, G., Galynker, I., Peskoff, D., Carpuat, M., White, J., Anadkat, S., Hoyle, A., & Resnik, P. (2025). *The prompt report: A systematic survey of prompt engineering techniques* (Version 6). arXiv. <https://doi.org/10.48550/arXiv.2406.06608>
- [50]. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological*

- Science*, 22(11), 1359–1366. <https://doi.org/10.1177/09567976111417632>
- [51]. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkumar, A., Barral, J. K., Semturs, C., Karthikesalingam, A., & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 1–9. <https://doi.org/10.1038/s41586-023-06291-2>
- [52]. Spearman, C. (1904) *The Proof and Measurement of Association between Two Things*. *The American Journal of Psychology*, 15, 72-101. <https://doi.org/10.2307/1412159>
- [53]. ter Braak, C. J. F., & Šmilauer, P. (2002). *CANOCO reference manual and CanoDraw for Windows user's guide: Software for canonical community ordination (Version 4.5)*. Microcomputer Power. <https://edepot.wur.nl/405659>
- [54]. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29, 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>
- [55]. Thomas, D. W., Kenfack, D., Chuyong, G. B., Moses, S. N., Losos, E. C., Condit, R. S., & Songwe, N. C. (2003). Tree species of the South Western Cameroon: Tree Distribution Maps, Diameter Tables, and Species documentation of the 50-hectare Korup Forest Dynamics plot. Center for Tropical Forest Science, Washington DC. 247pp.
- [56]. Touchon, J. C., & McCoy, M. W. (2016). The mismatch between current statistical practice and doctoral training in ecology. *Ecosphere*, 7(8), e01394. <https://doi.org/10.1002/ecs2.1394>
- [57]. Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley. 688pp
- [58]. Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, 34(1), 23–25. <https://doi.org/10.2307/2682991>
- [59]. Unwin, A. (2015). *Graphical Data Analysis with R* (1st ed.). Chapman and Hall/CRC. 310pp. <https://doi.org/10.1201/9781315370088>
- [60]. Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1-7). <https://doi.org/10.1145/3491101.3519665>
- [61]. Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11), 2766–2772. <https://doi.org/10.1890/07-0043.1>
- [62]. Wamba, S. F., & Queiroz, M. M. (2023). Responsible Artificial Intelligence as a secret ingredient for digital health: Bibliometric analysis, insights, and research directions. *Springer Nature*, 25, 2123–2138. <https://doi.org/10.1007/s10796-021-10142-8>
- [63]. Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, 92(1), 3-10. <https://doi.org/10.1890/10-0340.1>
- [64]. Weng, L., Liu, J., & Le, Q. V. (2023). Large language models as tool makers. *arXiv preprint arXiv:2305.17126*. <https://doi.org/10.48550/arXiv.2305.17126>
- [65]. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv:2302.11382*. <https://doi.org/10.48550/arXiv.2302.11382>
- [66]. Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). *Welcome to the tidyverse*. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- [67]. Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. 260pp. <https://doi.org/10.1007/978-3-319-24277-4>
- [68]. Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Computational Biology*, 13(6), e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>
- [69]. Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)* (Article 437, pp. 1–21). Association for Computing Machinery. <https://doi.org/10.1145/3544548.3581388>
- [70]. Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8), 1–25. <https://doi.org/10.18637/jss.v027.i08>
- [71]. Zuur, A. F., Ieno, E. N., & Smith, G. M. (2007). *Analysing ecological data*. Springer. <https://doi.org/10.1007/978-0-387-45972-1>