Journal of Data Analysis and Information Processing, 2024, 12, 210-231
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

High Dimension Multivariate Data Analysis for Small Group Samples of Chemical Volatile Profiles of African Nightshade Species

Lorna Chepkemoi[1], Daisy Salifu[1]*, Lucy Kananu Murungi[2], Henri E. Z. Tonnang[1]

[1]International Centre of Insect Physiology and Ecology (ICIPE), Nairobi, Kenya
[2]Department of Horticulture and Food Security, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya
Email: *dsalifu@icipe.org

How to cite this paper: Chepkemoi, L., Salifu, D., Murungi, L.K. and Tonnang, H.E.Z. (2024) High Dimension Multivariate Data Analysis for Small Group Samples of Chemical Volatile Profiles of African Nightshade Species. *Journal of Data Analysis and Information Processing*, **12**, 210-231.
https://doi.org/10.4236/jdaip.2024.122012

Received: February 6, 2024
Accepted: May 14, 2024
Published: May 17, 2024

Abstract

Quantitative headspace analysis of volatiles emitted by plants or any other living organisms in chemical ecology studies generates large multidimensional data that require extensive mining and refining to extract useful information. More often the number of variables and the quantified volatile compounds exceed the number of observations or samples and hence many traditional statistical analysis methods become inefficient. Here, we employed machine learning algorithm, random forest (RF) in combination with distance-based procedure, similarity percentage (SIMPER) as preprocessing steps to reduce the data dimensionality in the chemical profiles of volatiles from three African nightshade plant species before subjecting the data to non-metric multidimensional scaling (NMDS). In addition, non-parametric methods namely permutational multivariate analysis of variance (PERMANOVA) and analysis of similarities (ANOSIM) were applied to test hypothesis of differences among the African nightshade species based on the volatiles profiles and ascertain the patterns revealed by NMDS plots. Our results revealed that there were significant differences among the African nightshade species when the data's dimension was reduced using RF variable importance and SIMPER, as also supported by NMDS plots that showed *S. scabrum* being separated from *S. villosum* and *S. sarrachoides* based on the reduced data variables. The novelty of our work is on the merits of using data reduction techniques to successfully reveal differences in groups which could have otherwise not been the case if the analysis were performed on the entire original data matrix characterized by small samples. The R code used in the analysis has been shared herein for interested researchers to customise it for their own data of similar nature.

Keywords

Random Forest, Similarity Percentage, PERMANOVA, ANOSIM, Non-Metric Multi-Dimensional Scaling

1. Introduction

Quantification of plant volatiles or any volatiles emitted by living organisms gives rise to high multidimensional data. Such studies often are set to compare volatile profiles of groups of organisms. The fundamental objective in such studies is to identify compounds that discriminate between groups. In plants, such studies are useful to understand host plant-insect pest interaction, as particular plant species could be host or non-host of insect pest [1]. These volatiles can be emitted from flowers, leaves, fruits, roots or any other part of the plant into the atmosphere or soil, allowing the plant to interact with other organisms. There has been extensive investigation on the significance of volatiles in plant physiology and ecology and their roles in mutualistic interaction with other organisms [2]. For instance, pollinators are attracted by volatiles emitted from floral tissues [1] and conversely play a crucial role in host finding by insect pests in agro-ecosystems [3] [4].

Other studies have demonstrated the beneficial effect of herbivore-induced plant volatile compounds (HIVOCs) as host location signals for parasitoids and herbivore predators [2] [5] [6]. This indirect chemical defense is most likely as significant as direct chemical and physical defenses in inhibiting herbivore damage [5]. Furthermore, volatile signals emitted by injured plants may transmit a signal to surrounding plants, stimulating defensive responses [7]. Plants may release volatiles in response to changes in light, temperature, or other abiotic stressors [2]. Research on plant volatiles has therefore provided insights in understanding variations in plant species with regard to evolutionary origins and ecological consequences in terms of plant-insect interactions and functional responses.

High dimensional multivariate data obtained from chemical volatile analysis has usually been analyzed using usual linear methods such as principal component analysis (PCA) [8] [9], linear discriminant analysis (LDA) [10] [11], multivariate analysis of variance and other methods. In some cases, PCA has been used as a preprocess to reduce dimensionality of data before applying LDA on the principal components [12]. Principal component analysis and linear discriminant analysis are famous feature extraction methods that are subject to small sample sizes. In fact, the effect of small sample sizes for high dimensional data has been discussed by several authors [13] [14] [15]. Generally, sample size n must be greater than the number of variables or features, p. Small sample sizes with PCA tend to provide eigenvectors coefficients (also known as factor loadings) and eigenvalues that are unprecise estimate of population values while

large sample sizes provide precise estimates [13] [16]. The effect of small sample size is even stronger for linear discriminant methods such as LDA which loses performance when faced with small sample sizes in high dimensional variable space [17]. The effect of small sample size on classical statistical methods is further aggravated by the methods' reliance on model assumptions that can hardly be verified when sample sizes are small. This favors use of statistical methods having their model assumptions relaxed for the analysis of such small samples. Consequently, this has led to increased application of non-metric ordination techniques, especially in revealing patterns and producing meaningful results that are easy to interpret in multivariate data [18]. Specifically, non-metric multidimensional scaling (NMDS) has gained immense application in ecological data due to its ability to handle non-linear data [19] [20].

Data on chemical profile of volatiles is characterized by small samples n and large number of variables p, thus often $p > n$. The crucial problem then is the presence of variables not significantly contributing to discrimination of samples but capable of contributing to random noise which potentially could obscure differences in groups. The use of variable selection methods to reduce dimensionality can lead to improvements in discrimination of samples and enhanced visualization. Therefore, in this study we use random forest (RF) technique and similarity percentage (SIMPER) as variable reduction techniques on chemical profiles of volatile compounds of three African nightshade plants prior to application of NMDS and hypothesis testing using permutational multivariate analysis of variance (PERMANOVA) and analysis of similarities (ANOSIM).

The random forest (RF) technique is an ensemble classifier that generates several decision trees from a sample obtained from the original dataset [21]. Each decision tree uses a different bootstrap sample in building the tree by randomly selecting with replacement a sample from the dataset. The bootstrap sample is then fed as input to base learners which are combined using a majority vote [22]. Since the decision trees are unrelated, the decision made as a majority vote is better than the decision made by each individual tree [23]. Machine learning (ML) models have been shown to be robust in handling small sample size compared to other well-established models such as linear discriminant analysis [24]. In particular, RF has shown superior performance over other ML models especially with high dimensional small-sample datasets ($p > n$) [25] [26] [27]. Further, RF offers variable importance measures which are used to rank variables based on their predictive ability and we leverage on this aspect of RF to reduce the dimension of the data in this study prior to application of NMDS and hypothesis tests. Gini importance and permutation importance (mean decrease in accuracy) are the two commonly used variable importance measures [28] and are described in the methodology of this study.

Similarity percentage (SIMPER), proposed by Clarke [29] compares groups of sampling units pairwise and calculates each samples' contribution to the average between-group Bray-Curtis dissimilarity [30] [31]. The existence of good dis-

criminator variables in samples result in high quantitative presence yielding high average dissimilarity [30]. This allows identifying variables that significantly contribute to the dissimilarity between samples [32].

The study aims to assess the similarity or dissimilarity of three African nightshade species (*Solanum sarrachoides* Sendtner, *S. scabrum* Miller and *S. villosum* Mill.) using NMDS with data preprocessed for variable reduction using RF and SIMPER. We demonstrate a step-by-step analysis using NMDS and show the merits of reducing the variables using RF and SIMPER as also confirmed by hypothesis tests using permutational multivariate analysis of variance (PERMANOVA) and analysis of similarities (ANOSIM). We demonstrate that using RF and SIMPER to reduce data dimension, enhances the visualization of the projection of the nightshade species on the volatile compounds and increases the power of hypothesis tests in PERMANOVA and ANOSIM. The R code used in the analysis is shared here for interested researchers to customise it for their own datasets of similar nature.

2. Materials and Methods

2.1. Data

Our study used secondary data on amount of volatile organic compounds (VOCs) obtained from intact plants of three African nightshade species namely, *S. sarrachoides*, *S. scabrum* and *S. villosum*. The volatile chemical analyses were performed using gas chromatography-mass spectrometry (GC/MS) on three samples from each African nightshade species. A total of 58 volatile organic compounds were identified. A full description of the data and methodology is found in Murungi *et al.* [33].

2.2. Data Reduction Techniques

In this study, we take advantage of the fundamental outcome of RF to reduce the dimension of the 58 volatile organic compounds of the three African nightshade species prior to application of NMDS, PERMANOVA and ANOSIM.

Random forest technique generates several decision trees from a sample obtained from the original dataset. The parameters under consideration in the implementation of the RF algorithm are therefore, number of features for growing each tree (*mtry*) and number of trees to be generated (*ntree*). Here, *ntree* was fixed at default value 500 while *mtry* was evaluated by searching for the optimal *mtry* value using the tune function implemented in random Forest package in R. Approximately two-thirds of the samples (in-bag samples) are used to train the decision trees, with the remaining one-third (out-of-bag samples) used during an internal cross-validation procedure to estimate the performance of RF algorithm [21] [34]. There is no pruning of trees in RF as ensemble and bootstrapping schemes help it to overcome overfitting issues [25]. RF produces variable importance measures namely Gini importance and permutation importance (mean decrease in accuracy), which are used to rank variables based on their

predictive ability. Gini importance has been disputed as having undesirable properties such as being biased in favor of variables with many categories while permutation importance has been proposed as a corrective measure to this biasness [35]. Therefore, variable importance was derived using permutation importance.

Additionally, SIMPER was used to identify volatile compounds that showed significant difference between the African nightshade species ($\alpha = 0.05$) to augment the volatile compounds selected under the RF variable importance measure. SIMPER analysis works at the univariate level by computing the relative contribution of each variable to the overall average Bray-Curtis dissimilarities by pairwise comparison of groups.

2.3. Non-Metric Multidimensional Scaling and How It Works

NMDS is a rank-based approach whose algorithm works by first randomly placing samples in an ordination space, with the desired number of dimensions defined a *priori*. The placement of samples is by an iterative process that attempts to find an ordination based on a stress function, in which ordinated sample distance closely match the order of sample dissimilarities in the original distance matrix [36]. This means that the original distance data is substituted with ranks. Samples are represented as points in a two or three-dimensional space such that the relative distances of all points are in the same rank order as the relative similarities of the samples [37] [38]. The mapping of samples using ranks preserves their ranked differences which enhances rescaling or rotation of axes for better visualization and interpretation [36]. Several iterations are implemented in the algorithm to obtain the lowest stress value possible, thus the stress function measures the goodness of fit of the distance adjustment in the reduced variable space configuration. Therefore, the lower the stress value, the better the data are represented in an ordination. The commonly utilized stress measure is Kruskal's stress [39] defined as;

$$\text{stress}_1^2 = \frac{\sum_{ij}\left(d_{uv} - \hat{d}_{uv}\right)}{\sum_{uv} d_{uv}^2} \tag{1}$$

where d_{uv} represents the actual distance between samples u and v in an ordination space;

$\hat{d}_{uv}$ represents the fitted distance between samples u and v.

Stress values are measured on a scale of 0 to 1 [39], with a stress value of 0 indicating similarity between all rank order distances in the input data and final ordination. Stress values reduce with increasing NMDS dimensionality. Stress value less than 0.05 gives an excellent representation with no prospect of misinterpretation while stress values greater than 0.2 are likely to yield NMDS plots that are hard to interpret [29]. In this study, we take advantage of the fundamental outcome of RF to reduce the dimension of the 58 volatile organic compounds of the three African nightshade species prior to application of NMDS, PERMANOVA and ANOSIM.

2.4. Distance Measures Used in Non-Metric Multidimensional Scaling

NMDS uses distance measures for ordination and some of these distances include Euclidean, Manhattan, Bray-Curtis, Kulczynski as described below.

2.4.1. Euclidean

Euclidean distance measures distance between two samples in multidimensional space and is calculated as the square root of the sum, over all the variables, of the square of the difference between values of a pair of samples [40]. It has no upper limit and is strongly affected by large number of zeros in the data, which often lead to high similarities between samples not sharing same variables. Moreover, it is a symmetrical index which treats double zeros in the same way as double presences resulting to shrinking of distance between two samples. To make the resulting Euclidean distances asymmetrical, the data is first transformed using either Chord, Hellinger or chi-square transformation [41]. The Euclidean distance measure is given as:

$$D = \sqrt{\sum_{v=1}^{k} \left(y_{1v} - y_{2v} \right)^2} \tag{2}$$

where D is the distance measure, k is the number of variables, and y_{1v} and y_{2v} are values of variable v in sample 1 and 2, respectively.

2.4.2. Manhattan

Manhattan distance is obtained by computing the sum of the absolute differences between distances of a pair of samples [42]. It has the same properties as Euclidean distance and is majorly dominated by variables with large values. The distance measure is given as:

$$D = \sum_{v=i}^{k} \left| \left(y_{1v} - y_{2v} \right) \right| \tag{3}$$

2.4.3. Bray-Curtis

Bray-Curtis is a modification of Manhattan distance measure where the sum of differences between samples across variables is standardized by the sum of variable values across samples, also summed across variables. Standardization was introduced to ensure that each variable is maximum-adjusted to equalize their contributions, and to relativize samples to reduce the effect of differing summed quantities. Bray-Curtis distance ranges between zero (completely similar variables) and one (completely dissimilar variables) [43]. The distance measure is given as:

$$D = \frac{\sum_{v=1}^{k} \left| y_{1v} - y_{2v} \right|}{\sum_{v=1}^{k} \left(y_{1v} + y_{2v} \right)} \tag{4}$$

2.4.4. Kulczynski

The distance measure calculates dissimilarities between pairs of samples. It is calculated by summing variable minima and dividing this value by each sam-

pling unit's total. The distance between the two sampling units is one minus the average of these two values [44]. The distance measure is given as:

$$D = 1 - \frac{\left(\frac{\sum_{v=1}^{k} \min\left(y_{1v}, y_{2v}\right)}{\sum_{v=1}^{k}\left(y_{1v}\right)} + \frac{\sum_{v=1}^{k} \min\left(y_{1v}, y_{2v}\right)}{\sum_{v=1}^{k}\left(y_{2v}\right)} \right)}{2} \tag{5}$$

2.5. Test of Difference in Groups

To compare overall variation in volatile compounds composition between the African nightshade species, analysis of similarities (ANOSIM) and permutational multivariate analysis of variance (PERMANOVA) were used.

Permutational multivariate analysis of variance (PERMANOVA) is a semiparametric method which tests and estimates sizes of main effects or interaction terms while retaining important statistical properties of rank based non-parametric multivariate methods such as flexibility to base the analysis on a dissimilarity measure of choice and distribution-free inferences achieved by permutations, with no assumption of multivariate normality [45]. Pseudo F-ratio is used as a test statistic in PERMANOVA [44] and is given as:

$$F = \frac{\mathrm{SSB}/(\beta - 1)}{\mathrm{SSW}/(N - \beta)} \tag{6}$$

where SSB is the sum of squared dissimilarities between groups; SSW is the sum of squared dissimilarities within groups; $(\beta - 1)$ is the degrees of freedom associated with grouping variable; and $(N - \beta)$ is the degrees of freedom associated with residuals.

The test statistic compares the total sum of squared ranked dissimilarities among samples in different groups to those belonging to the same group. The p-value is used to validate the significance of Pseudo F-ratio. On the other hand, ANOSIM is a hypothesis testing procedure that uses a dissimilarity measure to test for differences among groups. The null hypothesis being tested is that the average rank dissimilarities among samples within groups are the same as the average rank dissimilarities among samples from different groups. ANOSIM test statistic (R) is based on the rank differences between the average between-group $\left(\overline{r}_B\right)$ and within-group $\left(\overline{r}_W\right)$ given as:

$$R = \frac{\overline{r}_B - \overline{r}_W}{n(n-1)/4} \tag{7}$$

R is scaled within the range −1 to 1 with values greater than zero suggesting differences between groups, with more dissimilarity between groups than within groups. R values less than zero indicate more dissimilarities within groups than between groups, while R values of zero indicate that the dissimilarity within groups is the same as dissimilarity from different groups.

The workflow of our study in terms of methodology is summarized in **Figure 1**. The data under study is subjected to variable reduction techniques; thus, random forest and similarity percentage (SIMPER) prior to analysis by NMDS,

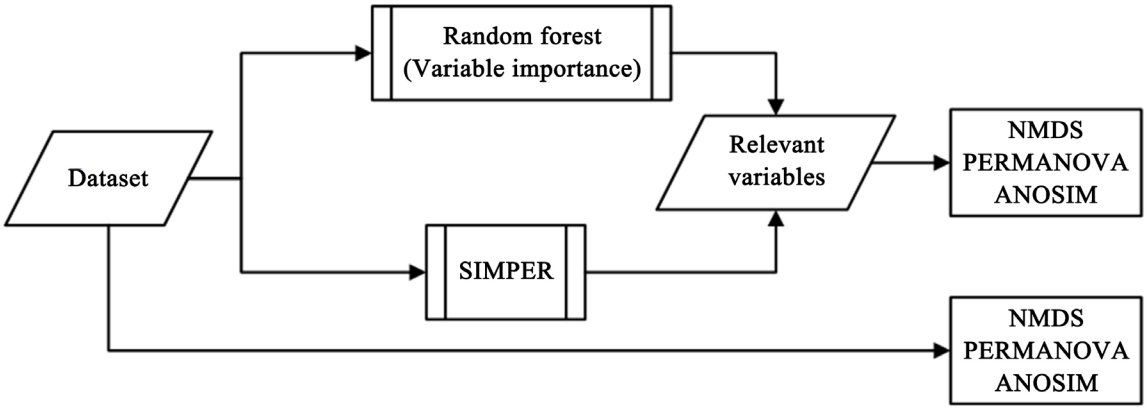

Figure 1. The workflow of the analysis procedures. The dataset is subjected to variable reduction techniques; random forest and similarity percentage (SIMPER) prior to analysis by NMDS, PERMANOVA and ANOSIM. NMDS, PERMANOVA and ANOSIM are also performed on the entire dataset for comparison purposes.

PERMANOVA and ANOSIM. The NMDS, PERMANOVA and ANOSIM are likewise performed on the entire dataset to compare out with that of reduced dimension.

2.6. The Analysis

Random forest technique was used to reduce the 58 volatile organic compounds (VOCs) under study based on variable importance. The top 12 compounds were selected to be used in the discrimination of the three African nightshade species (**Figure 2**). On the other hand, SIMPER was also used to identify variables that showed significant difference between the African nightshade species at $\alpha = 0.05$ level of significance and a total of 13 compounds were obtained. The outcomes of RF and SIMPER were combined to give a total of 16 "relevant" variables displayed in the Venn diagram (**Figure 3**). The 16 volatile compounds were then used in NMDS analysis. Bray Curtis distance which was determined as the suitable distance for these data was used to obtain pairwise similarity matrix, which determines the ecological distance between all pairs of nightshade species. Suitable k dimension for the NMDS plot was determined using scree plot, which is a plot of stress values versus number of dimensions. PERMANOVA and ANOSIM were performed to test for significant difference in volatile compound profiles of the African nightshade species. The NMDS, PERMANOVA and ANOSIM output on the reduced dataset (16 VOCs) was compared to NMDS, PERMANOVA and ANOSIM implemented on the entire dataset (58 volatile compounds).

All analyses were implemented in R version 4.1.3 [46] using the following packages; randomForest [47] and vip [48] for Random Forest variable importance analysis, vegan [49] for NMDS, PERMANOVA and ANOSIM; goeveg [50], ggplot2 [51] and ggforce [52] for scree plot and NMDS plots. The R script for commands used in the study is available at https://github.com/icipe-official/non-metric-multidimensional-scaling/blob/main/r-code

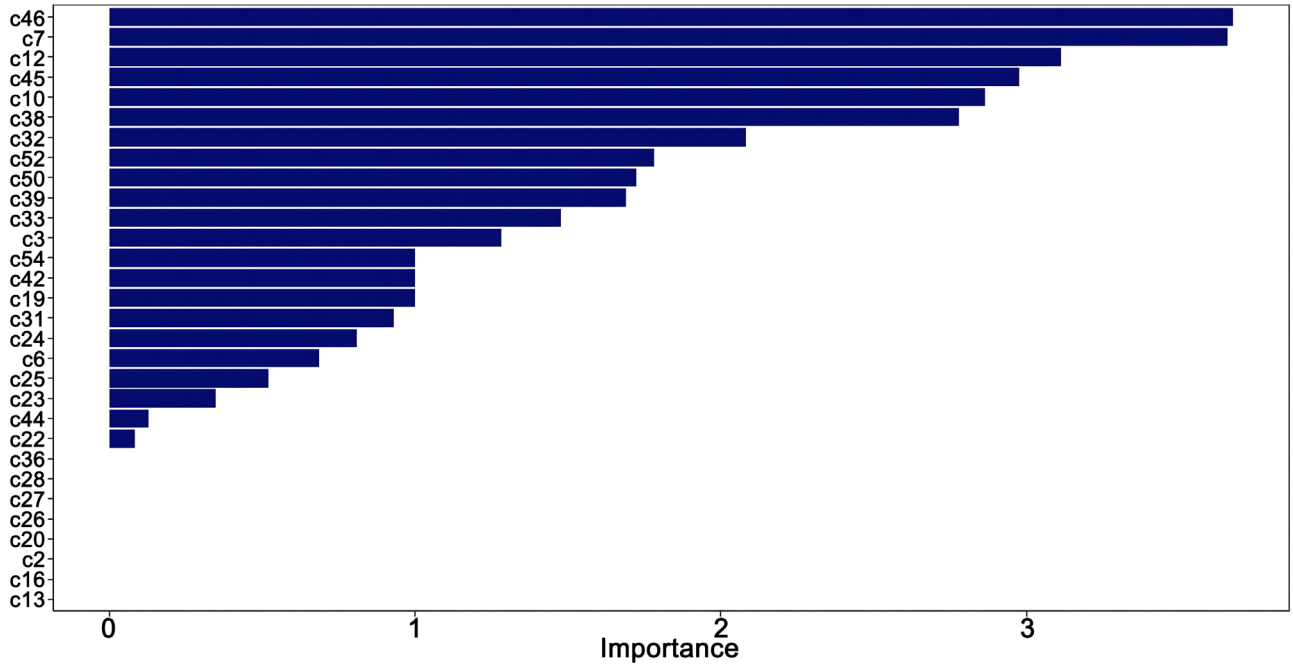

Figure 2. Random Forest variable importance plot based on permutation importance (mean decrease in accuracy). Only the top 12 VOCs were selected.

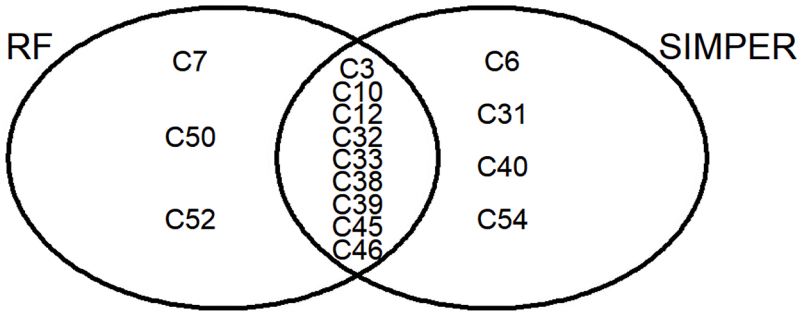

Figure 3. Venn diagram of the 16 selected "relevant" volatile compounds based on Random Forest and similarity percentage (SIMPER). c3, c6, c7, c10, c12, c31, c32, c33, c38, c39, c40, c45, c46, c50, c52, c54 are the labels of the volatile organic compounds.

3. Results

3.1. Similarities of VOCs in the Three Nightshade Species

The pairwise similarity matrix of VOCs in *S. sarrachoides*, *S. scabrum* and *S. villosum* indicated that all distances ranged between 0.210 to 0.902. This explains the variation in the VOCs emitted by the three African nightshade species (Table 1).

3.2. Variables Selected

From the RF and SIMPER procedures, a total of 16 volatile organic compounds out of 58 were selected, nine of which were the same, while three were unique to RF and four unique to SIMPER according to our selection criterion. The 16 variables are displayed in Figure 3. RF variable importance results had c46 as the

Table 1. Pairwise similarity matrix of volatile organic compounds concentration in three Af-rican nightshade species (*S. sarrachoides*, *S. scabrum* and *S. villosum*) based on Bray-Curtis distance.

		African nightshade species								
		S. sarrachoides			*S. scabrum*			*S. villosum*		
	rep	1	2	3	1	2	3	1	2	3
S. sarrachoides	1									
	2	0.8962								
	3	0.7798	0.5532							
S. scabrum	1	0.4873	0.8102	0.5505						
	2	0.8738	0.3453	0.3856	0.7370					
	3	0.8080	0.5722	0.2101	0.6152	0.4046				
S. villosum	1	0.2813	0.9020	0.7664	0.4852	0.8743	0.8134			
	2	0.7141	0.7760	0.6048	0.6932	0.6989	0.5879	0.7664		
	3	0.7003	0.7375	0.4968	0.5552	0.6518	0.5732	0.6781	0.4922	

top-most important volatile compound (Figure 2; Table S2). This was consistent with SIMPER results as c46 was significantly different between *S. sarrachoides* and *S. scabrum*, and *S. villosum* and *S. scabrum*, respectively (Table S1; Table S2). These "relevant" volatile compounds were used to perform NMDS, PERMANOVA and ANOSIM analysis.

3.3. NMDS on Reduced Dataset and Full Dataset

Non-metric multi-dimensional scaling based on Bray Curtis distance was performed with dimension $k = 3$ as suggested by the scree plot (Figure 4) on the full dataset (58 VOCs). Given a scree plot, the value of the dimension of NMDS is at the elbow of the line plot which is the value beyond which additional dimensions do not substantially lower the stress value. Such value provides a suitable dimension for visualizing the NMDS plot.

We evaluated the NMDS ordination at different dimensions to obtain the stress value that optimizes the ordination fit based on Bray Curtis distance for both reduced dataset and full dataset. The stress values obtained at different dimensions with different number of input variables are presented in Table 2.

Table 2 indicates that as the number of dimensions increase, the stress value reduces. Stress values are also higher for high-dimensional data in variable space as compared to low-dimensional data in variable space. The NMDS ordination algorithm could not converge when the "relevant" variables with dimension $k = 3$ were used, as the stress value was nearly zero. Consequently, the reduced dataset (16 VOCs) with lower stress value for dimension $k = 2$ was the ordination of choice as was also supported by the shephard plot that indicated goodness of fit with linear fit, $r^2 = 0.993$ and non-metric fit, $R^2 = 0.998$ (Figure 5(b)) which were both higher compared to the NMDS ordination using full dataset (Figure 5(a)).

Table 2. NMDS ordination dimension and the corresponding stress values based on Bray Curtis distance for the full dataset of 58 VOCs and the reduced dataset of 16 VOCs.

Dimension	Stress value (full dataset 58 VOCs)	Stress value (reduced dataset 16 VOCs)
1	0.2453	0.1998
2	0.0806	0.0403
3	0.0320	No convergence

Stress value < 0.05-Excellent; Stress value > 0.2-Poor.

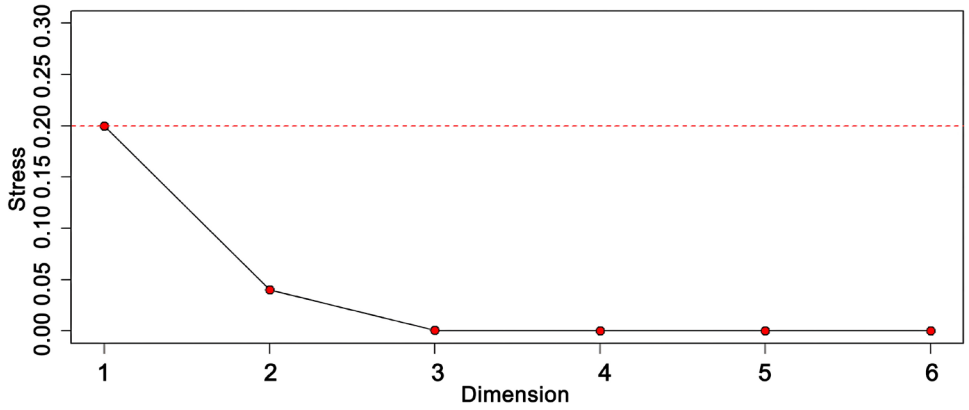

Figure 4. Scree plot based on Bray-Curtis distance to determine the dimensionality of NMDS ordination using all the 58 VOCs. Here, the suitable dimension *k* is 3, thus the value at the elbow of the line plot.

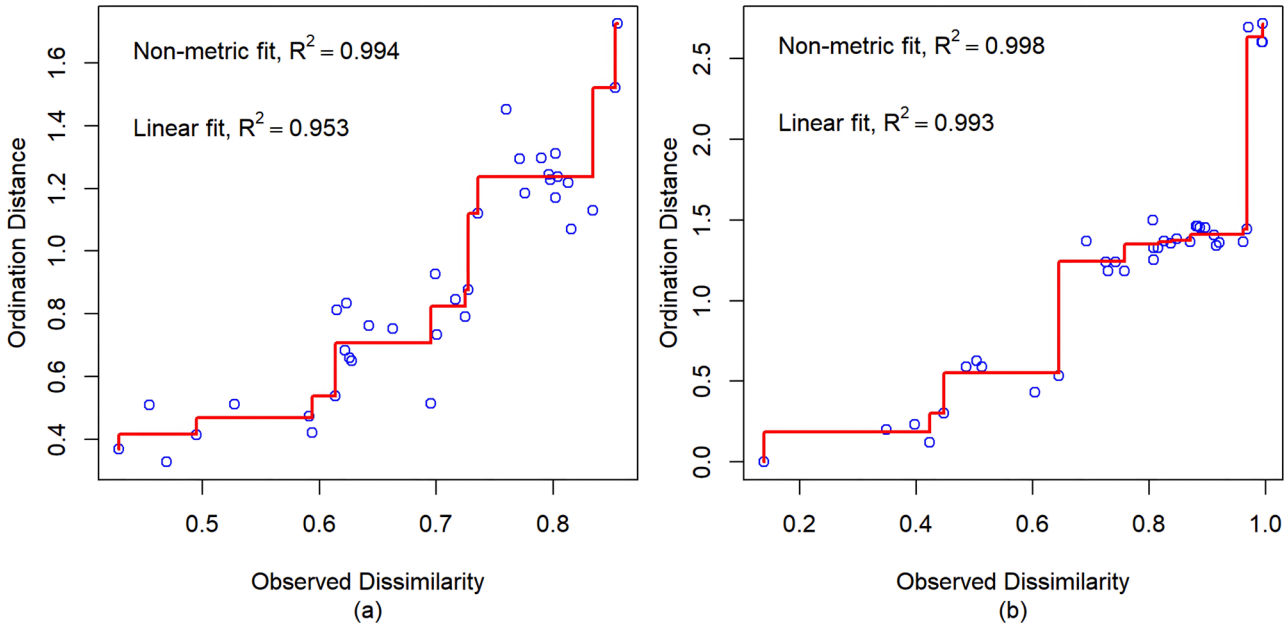

Figure 5. Shepards plot for (a) all 58 variables (stress = 0.0806) and (b) 16 selected variables (stress = 0.0403) showing goodness of fit metrics for the NMDS ordination.

3.4. NMDS Ordination Plots

When all volatile organic compounds were used, the ordination ellipses showed

heavy overlaps among the three African nightshade species (Figure 6). On the other hand, when only the selected "relevant" variables were put into consideration, it was possible to distinguish the three African nightshade species (Figure 7). NMDS plots the distances between points in the same rank order as distances (or similarities) in the original matrix. The closer two samples are on the plot the more similar those samples are in terms of the underlying data. Further, the samples enclosed within an ellipse or close to the ellipse belong to the same group. The NMDS plots showed that *S. scabrum* had dissimilar volatile compounds as compared to *S. villosum* and *S. sarrachoides*. Further, *S. sarrachoides* had dissimilar volatiles profiles as compared to *S. scabrum* and *S. villosum* (Figure 7).

The dissimilarities of VOCs in the three African nightshade species showed by NMDS ordination plots, was supported by hypothesis testing using PERMANOVA and ANOSIM. PERMANOVA results showed that there was no significant difference between the three African nightshade species (p-value = 0.554) when all the VOCs were considered while there was significant difference between the plants (p-value = 0.022) when only "relevant" VOCs were considered. This was further supported by ANOSIM test that showed similar conclusion (Table 3).

4. Discussion

The data in this study are characterized by high dimensionality and small sample size, which tends to reduce the statistical power of tests. The classical statistical methods do not appropriately regulate type 1 error rate when sample sizes are

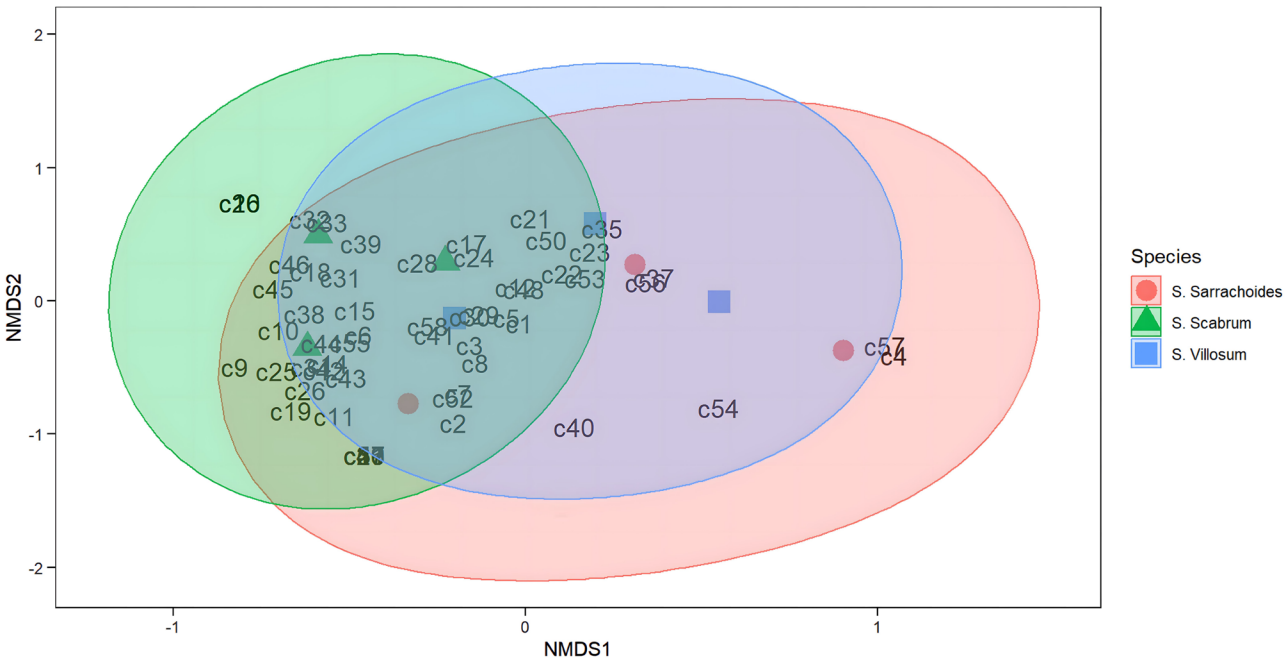

Figure 6. Non-metric multidimensional scaling (NMDS) ordination biplot (stress = 0.08, k = 2) based on Bray Curtis distance showing similarities of the African nightshade species (*S. sarrachoides*, *S. scabrum* and *S. villosum*). c1, c2, …, c58 are the labels of all the volatile organic compounds.

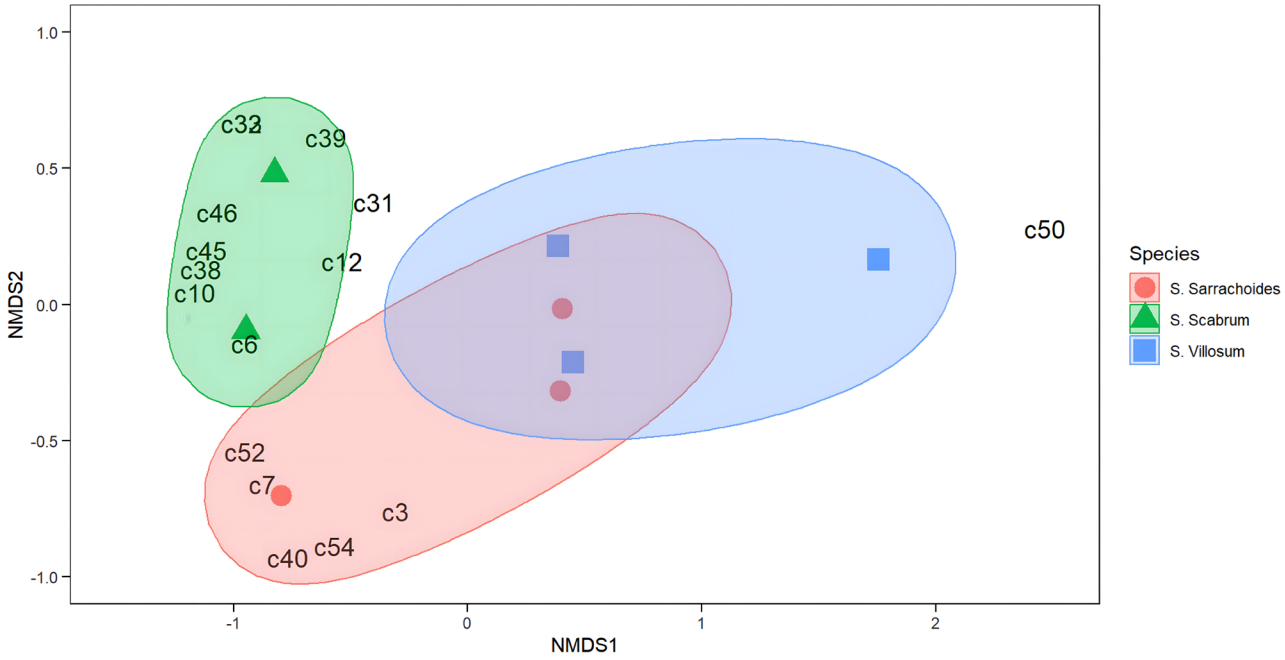

Figure 7. Non-metric multidimensional scaling (NMDS) ordination biplot (stress = 0.04, k = 2) based on Bray Curtis distance showing similarities of the African nightshade species (*S. sarrachoides*, *S. scabrum* and *S. villosum*). c3, c6, c7, c10, c12, c31, c32, c33, c38, c39, c40, c45, c46, c50, c52, c54 are the labels of the selected "relevant" volatile organic compounds for discrimination of the species.

Table 3. PERMANOVA (number of permutations = 999) and ANOSIM hypothesis tests on volatile organic compounds for the three African nightshade species.

Test		58 Volatile organic compounds	16 Volatile organic compounds
PERMANOVA	Pseudo F-ratio	0.8176	2.7595
	p-value	0.554	0.022
ANOSIM	ANOSIM R statistic	−0.07	0.4733
	p-value	0.589	0.018

very small as they require moderate to large sample sizes for analysis. Such methods as multivariate analysis of variance (MANOVA) either behave liberally and over-reject the null hypothesis, or behave conservatively [53]. Chang *et al.* [54] highlighted the need of a large sample size for an accurate type 1 error control.

Non-metric multidimensional scaling has been used in small sample situation to reveal patterns in multivariate datasets visualized in a reduced dimension space. As such its application in chemical ecology characterized by high dimensional small sample size datasets has become of notable interest. For instance, Hufnagel [55] visualized differences in the amount of glycoalkaloid α-solanine among *Solanum tuberosum* L., *S. chacoense* Bitter, *S. pinnatisectum* Dunal and *S. immite* Dunal using NMDS. Suinyuy *et al.* [56] analyzed volatile composition of male and female of African cycad species using headspace technique and gas

chromatography-mass spectrometry (GC-MS) where the species were clustered using NMDS according to shared chemical volatiles. NMDS efficiency in the chemical ecology field has been contributed by the technique's properties such as being less sensitive to variation in species response curve [57] and its requirement of only two dimensions to visualize similarity patterns compared to other ordination techniques which requires a minimum of three dimensions [43].

In our study, NMDS results revealed that when all the chemical volatiles were used in the analysis, the nightshade species were highly overlapping. This might have been caused by the "curse of dimensionality" problem where, in high dimensional space an exponential increase in the space volume is experienced as the data becomes relatively small [58]. This makes it hard to find patterns in data samples shown by the ellipses overlap. Further, since the data was projected to a two-dimensional space from 58 dimensions, the similarities revealed by the NMDS plot might have been contributed by it.

To identify patterns in the volatile compounds, RF and SIMPER were employed to reduce the data dimensions. RF performance was contributed by its efficiency in recognizing data patterns, no assumptions related to data properties, user-friendly parameters and ability to flexibly address the interactions between predictive variables [27]. On the other hand, SIMPER determined the contribution of individual compounds to the separation of the three African nightshades as reflected in the NMDS plots. The performance of the data reduction techniques used in this study is supported by Muthoni [59], who used SIMPER and one way ANOSIM to compare the chemical profiles of the leaf volatiles of healthy and infected tomato plants and further visualized the clustering of the volatiles using NMDS.

Bray Curtis distance was considered as the best distance measure in obtaining the NMDS plots since the other distance measures; Euclidean, Manhattan and Kulcynski had poor ordination fit. This might have been contributed by some of the properties of the distance measures. For instance, Junker [60] highlighted that Euclidean distance often lead to high similarities between samples not sharing the same variables as it is affected by large number of zeros in the data [43] [61]. This contradicts with what Legendre and Legendre [43] stated on the properties of a good ecological distance in describing differences in species composition. Species sharing the same or most of the volatiles should have a small ecological distance than those not sharing any volatiles. Tomašev *et al.* [62] observed that Euclidean and Manhattan distances had similar trends in their results, which was in agreement with the similar NMDS plots obtained based on the two distance measures.

Bray-Curtis distance only takes the value zero for identical variables and ignores other variables having zeros [63]. Ecological distances based on Bray Curtis range from 0 to 1, with 0 indicating complete similarity and 1 indicating complete dissimilarity. Hence, interpreting the distances based on Bray Curtis is easier than the other distances which do not have an upper bound making it difficult to understand how similar two species are, as they are only understood in a

relative way [41]. This is not to say that the distance does not have its limitations; Bray Curtis and related measure such as Kulcynski tend to under estimate true ecological distances when distances become large. Therefore, the distance measure is as useful so far as it produces reasonable ecological ordinations through the ranks used for NMDS [45] [64].

Although NMDS performed well in revealing the patterns and reducing the dimension, it did not rigorously express the nature and degree of uncertainty concerning a priori hypotheses. Therefore, non-parametric methods that tested hypothesis concerning the three nightshades species were required to make probabilistic statements about the VOCs data [45]. The results obtained from the two non-parametric tests (PERMANOVA and ANOSIM) were in agreement. These two tests as discussed by Somerfield *et al.* [65] are complementary tests rather than alternative. Additionally, Rojas *et al.* [66] used NMDS to visualize the seed disperser functional types, and their relationships with fruit traits, the patterns observed were supported by PERMANOVA results.

In spite of the patterns among sampling units being visualized with an NMDS plot, use of rank orders to represent points in low dimensional space makes the solution obtained unstable and can even degenerate when applied to a small dataset [67]. In using NMDS, normality assumption is not required, however this necessitates use of intensive iterative algorithm since optimal solution may not be obtained from a single run. Therefore, multiple NMDS solutions with specified dimensionality is necessary to ensure a stable and optimal ordination configuration [67]. Further, multivariate visualization of samples by any ordination technique is not the end point of analysis but should be viewed as a framework in which patterns of individual subjects can be interpreted.

5. Conclusion

Our results showed that there was a dissimilarity between African nightshade species of *S. sarrachoides*, *S. scabrum* and *S. villosum* when the data dimension was reduced to only 16 volatile compounds as compared to using all the 58 volatile compounds depicted in the NMDS plots and outputs of PERMANOVA and ANOSIM hypothesis tests. Our study shows the merit of reducing variables using RF and SIMPER to enhance visualization and in turn increase the power of PERMANOVA or ANOSIM in analysis of high dimensional small sample dataset as encountered in chemical ecology. Based on our results, we recommend use of RF, SIMPER or any other applicable data reduction technique when dealing with small samples in high dimensional data. Although the data reduction techniques used in our study performed well in discriminating the African nightshade species, the selected variables that were identified by RF and SIMPER might not necessarily be biologically important for chemical ecologists.

Acknowledgements

The authors gratefully acknowledge the financial support by the following or-

ganizations and agencies: UK's Foreign, Commonwealth & Development Office (FCDO); the Swedish International Development Cooperation Agency (Sida); the Swiss Agency for Development and Cooperation (SDC); the Federal Democratic Republic of Ethiopia; and the Government of the Republic of Kenya. We are grateful to Juma Meltus for creating the flow diagram **Figure 1**. The views expressed herein do not reflect the official opinion of the donors.

Authors Contributions

L.C.: Data analysis, manuscript writing and editing; D.S.: Conceptualization, manuscript review and editing. L.K.M.: Data author, manuscript review and editing, H.T.: Manuscript review and editing.

Competing Interests

The authors declare no competing interests.

References

[1] Tholl, D., Boland, W., Hansel, A., Loreto, F., Rose, U.S.R. and Schnitzler, J.P. (2006) Practical Approaches to Plant Volatile Analysis. *The Plant Journal*, **45**, 540-560. https://doi.org/10.1111/j.1365-313X.2005.02612.x

[2] Chung, S.H., Scully, E.D., Peiffer M., Geib S.M., Rosa, C., Hoover K. and Felton G.W. (2017) Host Plant Species Determines Symbiotic Bacterial Community Mediating Suppression of Plant Defenses. *Scientific Reports*, **7**, Article No. 39690. https://doi.org/10.1038/srep39690

[3] Salerno, G., Rebora, M., Piersanti, S., Gorb, E. and Gorb, S. (2020) Mechanical Ecology of Fruit-Insect Interaction in the Adult Mediterranean Fruit Fly *Ceratitis capitata* (Diptera: Tephritidae). *Zoology*, **139**, Article 125748. https://doi.org/10.1016/j.zool.2020.125748

[4] Zu, P.J., García-García, R., Schuman, M.C., Saavedra, S. and Melián, C.J. (2023) Plant-Insect Chemical Communication in Ecological Communities: An Information Theory Perspective. *Journal of Systematics and Evolution*, **61**, 445-453. https://doi.org/10.1111/jse.12841

[5] War, A.R., Paulraj, H.C.S., Gabriel, M., War, M.Y. and Ignacimuthu, S. (2011) Herbivore Induced Plant Volatiles: Their Role in Plant Defense for Pest Management. *Plant Signaling & Behavior*, **6**, 1973-1978. https://doi.org/10.4161/psb.6.12.18053

[6] Dicke, M., Van Poecke, R.M.P. and De Boer, J.G. (2003) Inducible Indirect Defence of Plants: From Mechanisms to Ecological Functions. *Basic and Applied Ecology*, **4**, 27-42. https://doi.org/10.1078/1439-1791-00131

[7] Engelberth, J., Alborn, H.T., Schmelz, E.A. and Tumlinson J.H. (2004) Airborne Signals Prime Plants Against Insect Herbivore Attack. Proc. *Proceedings of the National Academy of Sciences*, **101**, 1781-1785. https://doi.org/10.1073/pnas.0308037100

[8] Chin, S.T., Nazimah, S.A.H., Quek, S.Y., Man, Y.B.C., Rahman, R.A. and Hashim, D.M. (2007) Analysis of Volatile Compounds from Malaysian Durians (*Durio zibethinus*) Using Headspace SPME Coupled to Fast GC-MS. *Journal of Food Composition and Analysis*, **20**, 31-44. https://doi.org/10.1016/j.jfca.2006.04.011

[9] Drioiche, A., *et al.* (2022) Correlation between the Chemical Composition and the

Antimicrobial Properties of Seven Samples of Essential Oils of Endemic Thymes in Morocco against Multi-Resistant Bacteria and Pathogenic Fungi. *Saudi Pharmaceutical Journal*, **30**, 1200-1214. https://doi.org/10.1016/j.jsps.2022.06.022

[10] Paliy, O. and Shankar, V. (2016) Application of Multivariate Statistical Techniques in Microbial Ecology. *Molecular Ecology*, **25**, 1032-1057. https://doi.org/10.1111/mec.13536

[11] Verma, S.P., Uscanga-Junco, O.A. and Díaz-González, L. (2021) A Statistically Coherent Robust Multidimensional Classification Scheme for Water. *Science of the Total Environment*, **750**, Article 141704. https://doi.org/10.1016/j.scitotenv.2020.141704

[12] Ricciardi, C., *et al.* (2020) Linear Discriminant Analysis and Principal Component Analysis to Predict Coronary Artery Disease. *Health Informatics Journal*, **26**, 2181-2192. https://doi.org/10.1177/1460458219899210

[13] Osborne, J.W. and Costello, A.B. (2004) Sample Size and Subject to Item Ratio in Principal Components Analysis. *Practical Assessment, Research, and Evaluation*, **9**, Article 11.

[14] Kocovsky, P.M., Adams, J.V. and Bronte, C.R. (2009) The Effect of Sample Size on the Stability of Principal Components Analysis of Truss-Based Fish Morphometrics. *Transactions of the American Fisheries Society*, **138**, 487-496. https://doi.org/10.1577/T08-091.1

[15] Björklund, M. (2019) Be Careful with Your Principal Components. *Evolution*, **73**, 2151-2158. https://doi.org/10.1111/evo.13835

[16] Shaukat, S.S., Rao, T.A. and Khan, M.A. (2016) Impact of Sample Size on Principal Component Analysis Ordination of an Environmental Data Set: Effects on Eigenstructure. *Ekologia* (*Bratislava*), **35**, 173-190. https://doi.org/10.1515/eko-2016-0014

[17] Sharma, A. and Paliwal, K.K. (2015) Linear Discriminant Analysis for the Small Sample Size Problem: An Overview. *International Journal of Machine Learning and Cybernetics*, **6**, 443-454. https://doi.org/10.1007/s13042-013-0226-9

[18] Austin, M.P. (2013) Inconsistencies between Theory and Methodology: A Recurrent Problem in Ordination Studies. *Journal of Vegetation Science*, **24**, 251-268. https://doi.org/10.1111/j.1654-1103.2012.01467.x

[19] Damgaard, C. (2006) Modelling Ecological Presence-Absence Data along an Environmental Gradient: Threshold Levels of the Environment. *Environmental and Ecological Statistics*, **13**, 229-236. https://doi.org/10.1007/s10651-005-0004-2

[20] Jollife, 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

[21] Belgiu, M. and Drăgu, L. (2016) Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, **114**, 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

[22] Oshiro, T.M., Perez, P.S. and Baranauskas, J.A. (2012) How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition*: 8*th International Conference*, Berlin, 13-20 July 2012, 154-168. https://doi.org/10.1007/978-3-642-31537-4_13

[23] Chang, V., Bailey, J., Xu, Q.A. and Sun, Z. (2023) Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms. *Neural Computing and Applications*, **35**, 16157-16173. https://doi.org/10.1007/s00521-022-07049-z

[24] Olden, J.D. and Jackson, D.A. (2001) Fish-Habitat Relationships in Lakes: Gaining Predictive and Explanatory Insight by Using Artificial Neural Networks. *Transactions of the American Fisheries Society*, **130**, 878-897. https://doi.org/10.1577/1548-8659(2001)130<0878:FHRILG>2.0.CO;2

[25] Qi, Y. (2012) Random Forest for Bioinformatics. In: Zhang, C. and Ma, Y., Eds., *Ensemble Machine Learning. Methods and Applications*, Springer, New York, 307-323. https://doi.org/10.1007/978-1-4419-9326-7

[26] Wang, H., Yang, F. and Luo, Z. (2016) An Experimental Study of the Intrinsic Stability of Random Forest Variable Importance Measures. *BMC Bioinformatics*, **17**, Article No. 60. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-0900-5 https://doi.org/10.1186/s12859-016-0900-5

[27] Luan, J., Zhang, C., Xu, B., Xue, Y. and Ren, Y. (2020) The Predictive Performances of Random Forest Models with Limited Sample Size and Different Species Traits. *Fisheries Research*, **227**, Article 105534. https://doi.org/10.1016/j.fishres.2020.105534

[28] Janitza, S., Celik, E. and Boulesteix, A.L. (2018) A Computationally Fast Variable Importance Test for Random Forests for High-Dimensional Data. *Advances in Data Analysis and Classification*, **12**, 885-915. https://doi.org/10.1007/s11634-016-0276-4

[29] Clarke, K.R. (1993) Non-Parametric Multivariate Analyses of Changes in Community Structure. *Australian Journal of Ecology*, **18**, 117-143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

[30] Hattas, D., Hjältén, J., Julkunen-Tiitto, R., Scogings, P.F. and Rooke, T. (2011) Differential Phenolic Profiles in Six African Savanna Woody Species in Relation to Antiherbivore Defense. *Phytochemistry*, **72**, 1796-1803. https://doi.org/10.1016/j.phytochem.2011.05.007

[31] Gibert C. and Escarguel, G. (2019) PER-SIMPER—A New Tool for Inferring Community Assembly Processes from Taxon Occurrences. *Global Ecology and Biogeography*, **28**, 374-385. https://doi.org/10.1111/geb.12859

[32] Torok, V.A., Ophel-Keller, K., Loo, M. and Hughes, R.J. (2008) Application of Methods for Identifying Broiler Chicken Gut Bacterial Species Linked with Increased Energy Metabolism. *Applied and Environmental Microbiology*, **74**, 783-791. https://doi.org/10.1128/AEM.01384-07

[33] Murungi, L.K., Kirwa, H., Salifu, D. and Torto, B. (2016) Opposing Roles of Foliar and Glandular Trichome Volatile Components in Cultivated Nightshade Interaction with a Specialist Herbivore. *PLOS ONE*, **11**, e0160383. https://doi.org/10.1371/journal.pone.0160383

[34] Kulkarni, V.Y. and Sinha, P.K. (2012) Pruning of Random Forest Classifiers: A Survey and Future Directions. 2012 *International Conference on Data Science and Engineering* (*ICDSE*), Cochin, 18-20 July 2012, 64-68. https://doi.org/10.1109/ICDSE.2012.6282329

[35] Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. and Zeileis, A. (2008) Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, **9**, Article No. 307. https://doi.org/10.1186/1471-2105-9-307

[36] Ramette, A. (2007) Multivariate Analyses in Microbial Ecology. *FEMS Microbiology Ecology*, **62**, 142-160. https://doi.org/10.1111/j.1574-6941.2007.00375.x

[37] Van Der Gucht, K., *et al.* (2005) Characterization of Bacterial Communities in Four Freshwater Lakes Differing in Nutrient Load and Food Web Structure. *FEMS Microbiology Ecology*, **53**, 205-220. https://doi.org/10.1016/j.femsec.2004.12.006

[38] Salido, J.A. and Clemente, J. (2012) Non-Metric Multidimensional Scaling for Biological Characterization of Reduced Yeast Cell Cycle. 2012 *International Conference on Biological and Life Sciences*, Singapore, 23-24 July 2012, 104-108.

[39] Dexter, E., Rollwagen-Bollens, G. and Bollens, S.M. (2018) The Trouble with Stress: a Flexible Method for the Evaluation of Nonmetric Multidimensional Scaling. *Limnology and Oceanography*: *Methods*, **16**, 434-443. https://doi.org/10.1002/lom3.10257

[40] San Segundo, E., Tsanas, A. and Gómez-Vilda, P. (2017) Euclidean Distances as Measures of Speaker Similarity Including Identical Twin Pairs: A Forensic Investigation Using Source and Filter Voice Characteristics. *Forensic Science International*, **270**, 25-38. https://doi.org/10.1016/j.forsciint.2016.11.020

[41] Legendre, P. and Gallagher, E.D. (2001) Ecologically Meaningful Transformations for Ordination of Species Data. *Oecologia*, **129**, 271-280. https://doi.org/10.1007/s004420100716

[42] Gomathi, V.V. and Karthikeyan, S. (2014) An Efficient Clustering Segmentation Algorithm for Computer Tomography Image Segmentation. *Journal of Biomedical Engineering and Medical Imaging*, **1**, 1-11. https://doi.org/10.14738/jbemi.13.267

[43] Legendre, P. and Legendre, L. (2012) Numerical Ecology, Developments in Environmental Modelling. 3rd Edition, Elsevier, Amsterdam, 419.

[44] Gagné, S.A. and Fahrig, L. (2011) Do Birds and Beetles Show Similar Responses to Urbanization? *Ecological Applications*, **21**, 2297-2312. https://doi.org/10.1890/09-1905.1

[45] Anderson, M.J. (2001) A New Method for Non-Parametric Multivariate Analysis of Variance. *Austral Ecology*, **26**, 32-46.

[46] R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

[47] Liaw A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News*, **2**, 18-22. https://cran.r-project.org/doc/Rnews/

[48] Greenwell, B.M. and Boehmke, B.C. (2020) Variable Importance Plots—An Introduction to the Vip Package. *The R Journal*, **12**, 343-366. https://doi.org/10.32614/RJ-2020-013

[49] Oksanen, J., *et al.* (2022) Vegan: Community Ecology Package. https://cran.r-project.org/package=vegan

[50] Von Lampe, F. and Schellenberg, J. (2023) Goeveg: Functions for Community Data and Ordinations. https://cran.r-project.org/package=goeveg

[51] Wickham, H. (2016) Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. https://ggplot2.tidyverse.org https://doi.org/10.1007/978-3-319-24277-4

[52] Pedersen, T.L. (2022) Ggforce: Accelerating 'Ggplot2'. https://cran.r-project.org/package=ggforce

[53] Konietschke, F., Schwab, K. and Pauly, M. (2020) Small Sample Sizes: A Big Data Problem in High-Dimensional Data Analysis. *Statistical Methods in Medical Research*, **30**, 687-701. https://doi.org/10.1177/0962280220970228

[54] Chang, J., Zheng, C., Zhou, W. and Zhou, W. (2017) Simulation-Based Hypothesis Testing of High Dimensional Means under Covariance Heterogeneity. *Biometrics*, **73**, 1300-1310. https://doi.org/10.1111/biom.12695

[55] Hufnagel, M.J. (2015) Chemical Ecology of Wild *Solanum spp* and Their Interac-

tion with the Colorado Potato Beetle. Master's Thesis, Michigan State University, East Lansing.

[56] Suinyuy, T.N., Donaldson, J.S. and Johnson, S.D. (2012) Variation in the Chemical Composition of Cone Volatiles within the African Cycad Genus *Encephalartos*. *Phytochemistry*, **85**, 82-91. https://doi.org/10.1016/j.phytochem.2012.09.016

[57] Ruokolainen L. and Salo, K. (2006) Differences in Performance of Four Ordination Methods on a Complex Vegetation Dataset. *Annales Botanici Fennici*, **43**, 269-275.

[58] Wang, J., Liu, X. and Shen, H. (2019) High-Dimensional Data Analysis with Subspace Comparison Using Matrix Visualization. *Information Visualization*, **18**, 94-109. https://doi.org/10.1177/1473871617733996

[59] Muthoni, K.R. (2023) Identification and Mechanisms of Allelochemicals Regulating Root-Knot Nematode Parasitism. Ph.D. Thesis, Kenyatta University, Kahawa.

[60] Junker, R.R. (2018) A Biosynthetically Informed Distance Measure to Compare Secondary Metabolite Profiles. *Chemoecology*, **28**, 29-37. https://doi.org/10.1007/s00049-017-0250-4

[61] Roberts, D.W. (2017) Distance, Dissimilarity, and Mean-Variance Ratios in Ordination. *Methods in Ecology and Evolution*, **8**, 1398-1407. https://doi.org/10.1111/2041-210X.12739

[62] Tomašev, N., Radovanović, M., Mladenić, D. and Ivanović, M. (2014) The Role of Hubness in Clustering High-Dimensional Data. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 739-751. https://doi.org/10.1109/TKDE.2013.25

[63] Ricotta, C. and Podani, J. (2017) On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning. *Ecological Complexity*, **31**, 201-205. https://doi.org/10.1016/j.ecocom.2017.07.003

[64] Faith, D.P., Minchin, P.R. and Belbin, L. (1987) Compositional Dissimilarity as a Robust Measure of Ecological Distance. *Vegetatio*, **69**, 57-68. https://doi.org/10.1007/BF00038687

[65] Somerfield, P.J., Clarke, K.R. and Gorley, R.N. (2021) Analysis of Similarities (ANOSIM) for 2-Way Layouts Using a Generalised ANOSIM Statistic, with Comparative Notes on Permutational Multivariate Analysis of Variance (PERMANOVA). *Austral Ecology*, **46**, 911-926. https://doi.org/10.1111/aec.13059

[66] Rojas, T.N., Zampini, I.C., Isla, M.I. and Blendiger, P.G. (2022) Fleshy Fruit Traits and Seed Dispersers: Which Traits Define Syndromes? *Annals of Botany*, **129**, 831-838. https://doi.org/10.1093/aob/mcab150

[67] Kenkel, N.C. (2006) On Selecting an Appropriate Multivariate Analysis. *Canadian Journal of Plant Science*, **86**, 663-676. https://doi.org/10.4141/P05-164

Supplementary Information

Table S1. Average contribution of the three nightshade species based on pairwise comparison of the three African nightshade species, *S. sarrachoides* (sarr) and *S. scabrum* (sca), and *S. villosum* (vill) using similarity percentage (SIMPER).

Volatiles	Average contribution	Standard deviation	P-value	Pairwise comparison
c54	0.0104	0.0087	0.027	sarr vs villo
c3	0.0063	0.0054	0.019	sarr vs villo
c40	0.0040	0.0043	0.027	sarr vs villo
c46	0.0129	0.0063	0.012	sarr vs sca
c45	0.0085	0.0035	0.011	sarr vs sca
c31	0.0001	0.0000695	0.035	sarr vs sca
c12	0.0471	0.0272	0.018	sarr vs sca
c32	0.0153	0.0132	0.022	sarr vs sca
c461	0.0143	0.0052	0.006	villo vs sca
c39	0.0103	0.0093	0.041	villo vs sca
c451	0.0094	0.0027	0.006	villo vs sca
c33	0.0065	0.0053	0.022	villo vs sca
c6	0.0053	0.0038	0.038	villo vs sca
c38	0.0028	0.0021	0.047	villo vs sca
c10	0.0020	0.0016	0.034	villo vs sca

Table S2. All 58 volatile organic compounds with their actual name and code as used in this study.

Chemical volatile	volatile code
hexanal	c1
2-hexenal	c2
(Z)-3-hexen-1-ol	c3
heptanal	c4
3-Methyl-2-butenal	c5
1R-a-Pinene	c6
Benzaldehyde	c7
b-Pinene	c8
6-Methyl-5-hepten-2-one	c9
Beta Myrcene	c10
Octanal	c11
Limonene	c12
Benzyl alcohol	c13
Dihydromyrcenol	c14
3,7-Dimethyl-1-octanol (Geraniol tetrahydride)	c15
Methyl benzoate	c16

Continued

Linalool	c17
Nonanal	c18
1,2,4,5-Tetramethylbenzene (Durol)	c19
Isophorone	c20
2-Ethylhexanoic acid	c21
Octanoic acid	c22
à-Terpineol	c23
Decanal	c24
2,3,3-Trimethyl-2-(3-methylbutyl)-cyclohexanone	c25
Isothymol methyl ether (anisole)	c26
Nonanoic acid	c27
Isobornyl acetate	c28
Isobutyl butanoate	c29
Butyl butanoate	c30
Copaene	c31
B-Elemene	c32
7-Epi-Sesquithujene	c33
Longifolene	c34
a-Cedrene	c35
Caryophyllene	c36
B-Cedrene	c37
Gemacrene B	c38
Geranyl acetone	c39
Humulene	c40
2,5-Di-tert-butylbenzoquinone	c41
Epizonarene	c42
Butylated hydoxytoluene	c43
D-Cadinene (+)-	c44
Geranyl linalool	c45
4,8,12-Trimethyl-1,3 (E), 7 (E)-11-Tridecatetraene	c46
Caryophyllene oxide	c47
Cedrol	c48
Humulene epoxide II	c49
Methyl dihydrojasmonate	c50
7-Methyl-Z-tetradecen-1-ol acetate	c51
Ethylhexyl benzoate	c52
Isopropyl myristate	c53
Hexadrofarnesyl acetone	c54
Hexadecanoic acid	c55
Isopropyl palmitate	c56
10,18-Bisnorabieta-8,11-triene	c57
(Z)-9-Octadecanoic acid	c58