Journal of Manuscript Studies

RESEARCH ARTICLE

Application of clustering and ordination methods to determine the optimal plot size for vegetation sampling of a Sri Lankan dipterocarp forest

S.C.D. Wimalasena[1], P. Wijekoon[1*], S.S. Fernando[2] and I.A.U.N. Gunatilleke[3]

[1] *Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya.*

[2] *Postgraduate Institute of Science, University of Peradeniya, Peradeniya.*

[3] *Department of Botany, Faculty of Science, University of Peradeniya, Peradeniya.*

Revised: 17 August 2023; Accepted: 29 August 2023

Abstract: Using a complete inventory method to analyze the species distribution of a rainforest is a daunting task due to the extensive area to be sampled. The answer for this problem is to find the most suitable sampling plot size, which realistically represents the entire species population distribution. In this study, optimal plot size, which contains the highest number of species cover per unit area for the Sinharaja rainforest was determined using multivariate methods, namely, cluster analysis, correspondence analysis, non-metric multidimensional scaling and procrustean analysis. Plot sizes selected for the study were 5 × 5, 5 × 10, 10 × 10, 10 × 20, 20 × 20 and 20 × 40 m^2. It was revealed that the average linkage method is the most robust hierarchical clustering method according to the cophenetic correlation coefficient. The cluster solution was verified graphically using the four ordination techniques, correspondence analysis (CA), detrended correspondence analysis (DCA), principal coordinate analysis (PCoA) and non-metric multidimentional scaling (NMS), and the solutions were then compared using procrustean analysis. According to the procrustean rotations CA and NMS pair had the highest correlation, and procrustes sum of squares of pairwise rotations showed that NMS has the lowest sum of squares with CA. Since the ordination plots drawn by using NMS show clear differences among all 25 sites, it suggests that the NMS method is the most robust ordination method for forest data analyses. By assuming that the time for sampling work is constant for six plot sizes, it was noted that the 20 × 40 m^2 plot contains the highest species cover per unit area. Mantel test indicates that the species of 20 × 40 m^2 plots are highly correlated with 20 × 20 m^2 plots, and the result was verified using the average linkage method and NMS ordination. These results suggest that 20 × 40 m^2 is the most suitable rectangular shaped plot, and 20 × 20 m^2 is the most suitable square shaped plot, which contain the highest number of species cover per unit area compared to the other selected plot sizes.

Keywords: Cophenetic correlation, hierarchical clustering, ordination, procrustean analysis.

INTRODUCTION

Sinharaja forest can be described as the largest block of relatively undisturbed low land wet evergreen rainforest, which is located in the south west region of Sri Lanka. This study was conducted using a forest dynamic plot (FDP) of size 500 × 500 m^2 in Sinharaja, which was established in 1993. The FDP is located between 424 – 575 m above mean sea level (Gunatilleke *et al*., 2006). The species present in the study area were identified with the help of the National Herbarium of Sri Lanka and Dassanayake and Fosberg (1980 - 1996), and each tree was given a tag for identification purposes. As the total area of the Sinharaja forest is extensive, a complete inventory method cannot be practically used. Therefore researchers (Bormann, 1953; Ruokolainen & Salo, 2006) have paid attention to use sampling methods with fixed plots for forest data analyses.

Sampling techniques in ecology mainly focus on quadrat (a 4-sided figure) sampling techniques. However in practice, the term ‘quadrat’ is used for other sampling units having circular, hexagonal, or other irregular outline shapes. Quadrats can also be established randomly, regularly, or subjectively within a study site. In general, plants grow in clumps, and therefore long, narrow plots often include more species than square or round plots of equal area. However, the accuracy of estimates may decrease as the perimeter of the plot increases, since the

* Corresponding author (pushpaw@pdn.ac.lk)

researcher must make more subjective decisions about the placement of plants inside or outside the plot.

Quadrat counts have been used extensively on plants, and it is also suitable for animals if the person counting is a careful observer. Goldsmith *et al*. (1986) and Krebs (1999) have also discussed about quadrat sampling methodology. When sampling a forest community to estimate the abundance of a plant species, first a researcher has to make decisions on the optimal size and shape of the quadrat that best describes the species distribution of the total population. Therefore, the determination of appropriate plot size is important.

Generally, the increase of plot size increases the number of detected plant species, while the variance of this number decreases. In order to maximize the accuracy in parameter estimation, it is required to minimize this variance. Also, given the constraints of time and costs, sample plot size should be as large as possible, and the sample plots should be a representative of the entire population to have unbiased parameter estimates. If the plot is too small, then the uncertainty of sample estimates arises and if the sample is too large, then the cost will be unnecessarily high. Vegetation scientists traditionally used minimal area approach (Kenkel & Podani 1991) to describe the plant community, and the objective of using minimal area is to determine a size of plot, which is necessary to represent the entire plant community. Moravec (1973) suggested similarity analysis to determine the minimal area, and it is based on presence/ absence data. Barkman (1989) has shown that the identification of minimal area through the interpretation of species-area curve contains some problems. In this research the study the area was restricted to 500 x 500 m^2, and the samples were selected relative to this total area.

In the field of ecology, hierarchical agglomerative cluster analyses is used to cluster sample units based on species distribution, and ordination methods are used to verify the results obtained by cluster analysis graphically. For different types of ecological studies, several researchers have applied the average linkage method (Clarke & Ainsworth, 1993; Quinn & Keough, 2002; Singh *et al*., 2010) as the clustering method, and NMS ordination method (Ruokolainen & Salo, 2006) as the graphical method. However, the robustness of the techniques used is seldom examined.

The main objectives of this study are to identify the most robust clustering and ordination methods, and to determine the optimal plot size, which contains the highest number of species cover per unit area for the Sinharaja rainforest by using these multivariate statistical techniques.

METHODOLOGY

The study was conducted in the Sinharaja forest dynamic plot (FDP), which is a 500 × 500 m^2 (25 ha) permanent forest dynamics study plot. The plot was established according to the methods of Hubbell and Foster (1983) and Manokaran *et al*. (1992). The first census of the plot was done over the period 1994 - 1996, and a database was established by the researchers at the Department of Botany, University of Peradeniya in 1996. For this analysis, 25 smaller plots (quadrats) of size 20 × 40 m^2 were randomly selected from the forest dynamic plot database. The size of these quadrats was decided by considering the total study area (500 × 500 m^2). Within these 25 quadrats, it was identified from the preliminary analysis that there are 180 different species types, which have DBH (diameter at breast height) ≥ 10 cm. These 25 quadrats were taken as the sampling units (n = 25), and the different species types were taken as the variables (p = 180). In each quadrat, the number of trees (DBH ≥ 10 cm) for each species type was counted and recorded. In the preliminary survey, an identification (ID) number was assigned to each quadrat. To select the optimal plot size, the analysis done for 20 × 40 m^2 quadrats have also been applied to several smaller quadrats; 5 × 5, 5 × 10, 10 × 10, 10 × 20 and 20 × 20 m^2 (Figure 1), which are nested within each selected 20 × 40 m^2 quadrat.

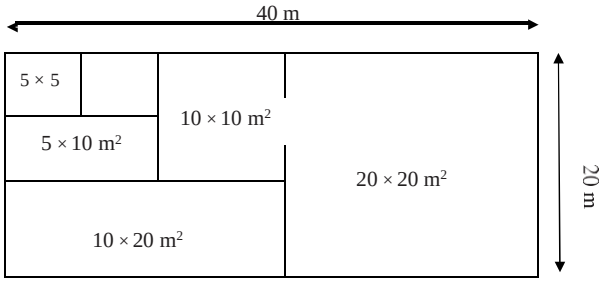


Figure 1: Nested plot configuration

The species count was determined for all six types of quadrats, and it is assumed that the time to count and tag trees in each plot size is constant. In order to choose the best clustering method, the hierarchical clustering methods, single, complete, average and Ward's methods were applied to 20 × 40 m^2 quadrats (sample units n = 25). The number of significant clusters was identified, and Adonis test was used to examine whether these clusters are significantly different. Finally, the

cophenetic correlation coefficient (CPCC) for each clustering method was calculated to decide the most suitable clustering method.

Furthermore, to compare and verify the forest assemblages obtained from hierarchical cluster analysis, robustness of the four ordination methods, namely, correspondence analysis (CA), non-metric multidimensional scaling (NMS), detrended correspondence analysis (DCA) and principal coordinate analysis (PCoA) were examined. The ordination methods CA, DCA, PCoA and NMS were applied by using the Bray and Curtis dissimilarity matrix, which was designed for quantitative data (Legendre & Legendre, 1998). The ordinations were made visually comparable with procrustean rotation (Minchin,1987; Økland, 1990). The same analysis was done by using the other five quadrat sizes 5 × 5, 5 × 10, 10 × 10, 10 × 20 and 20 × 20 m^2, to identify the optimal plot size. To test where these nested plots have the same species distribution as 20 × 40 m^2 quadrat, the chi-square goodness-of-fit test was used.

The relative variance of species count per unit area was also used as a comparison method to identify the suitable plot size, and the plot size with the minimum relative variance retain as the optimal plot size.

A researcher may also be interested to know whether he can use a square shape plot instead of a rectangular plot with the same species composition. To identify whether there is a significant correlation between the square shape plot and the 20 × 40m^2 rectangular plot, the Mantel test was used. This result can be further verified by applying the average linkage method and NMS method to the square shaped plot and 20 × 40 m^2 rectangular plot. The statistical software R 2.8 and MINITAB 16 were used to carry out the analysis.

RESULTS AND DISCUSSION

According to the preliminary analysis, 180 species were identified in the 25 quadrats of full dataset (20 × 40m^2), which have a DBH (diameter at breast height) $\geq$ 10 cm. *Humboldtia laurifolia*, *Mesua nagassarium* and *Agrostistachys intramarginalis* were the most abundant species. According to the Smithsonian Tropical Research Institute, 204 different species have been identified in the Sinharaja forest dynamic plot (Gunatilleke *et al*., 2004). This indicates that the 25 replicate quadrats, each of size 20 × 40 m^2 contain 88.23 % of the species of the entire forest dynamic plot.

Using single linkage, complete linkage, average linkage and Ward clustering methods, the dendrograms were drawn for species counts of the 25 quadrats of size 20 × 40 m^2, to identify the possible clusters in sample units (n = 25). For each method three cluster solution was considered, and the results were compared. Group separation of the 25 quadrats according to the species composition varies for different clustering methods, since hierarchical clustering methods generate different results for the same dataset.

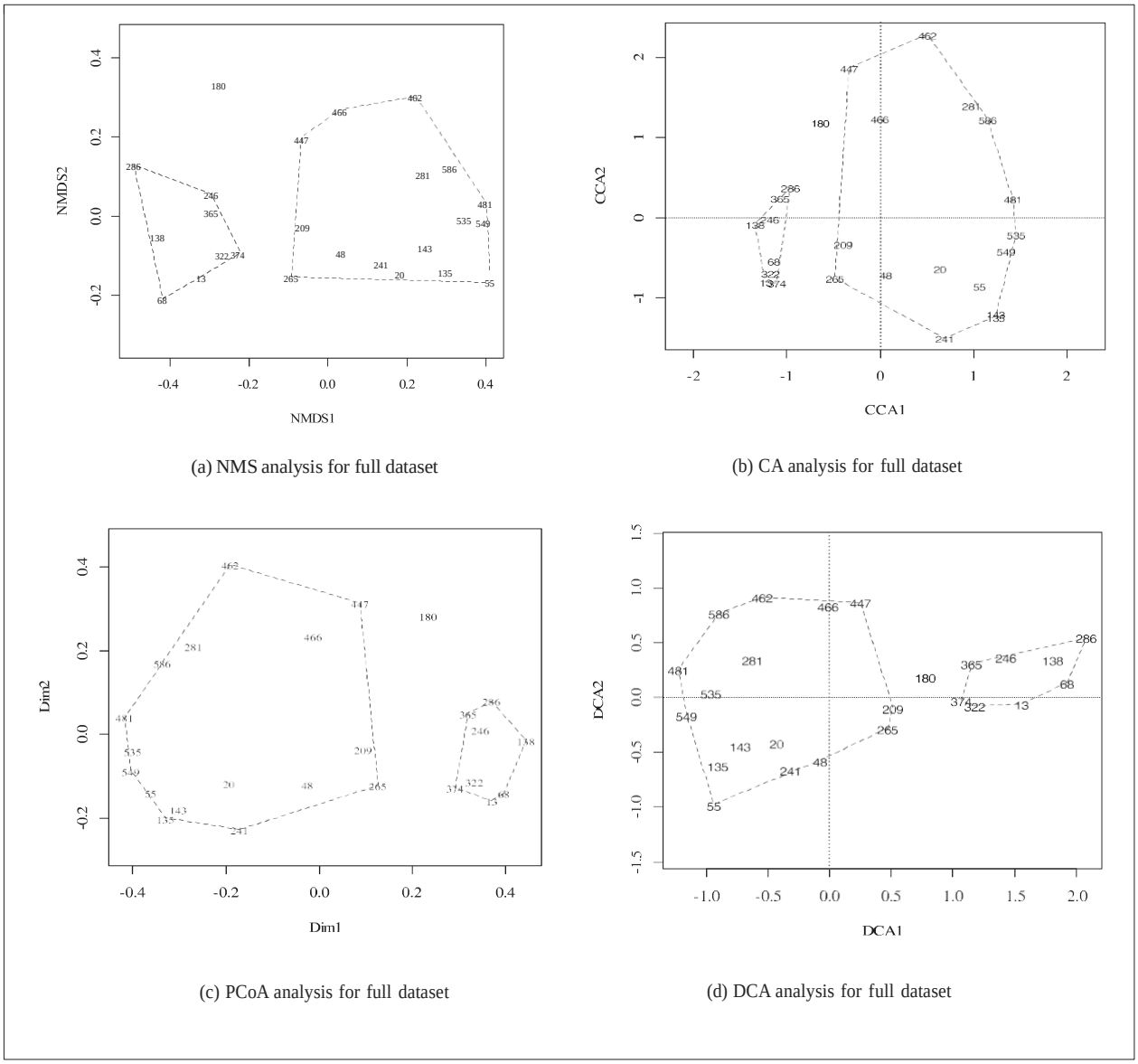
Since species counts in the quadrats sorted according to the three clusters obtained from the dendrograms were not normal (p value < 0.05), non-parametric MANOVA test (Adonis test) was used to determine the significant difference between the three clusters. Adonis test showed that there is a significant difference among the tree count of the quadrats in the selected three clusters for complete, Ward and average linkage method (p value < 0.05) except for single linkage method (p value > 0.05).

To find the most suitable clustering method among the complete linkage, Ward clustering and average linkage methods, the cophenetic correlation coefficient (CPCC) was obtained for quadrats of size 20 × 40m^2, and the average linkage method showed the highest CPCC (Table 1). The same procedure was applied to the other smaller nested quadrats to determine the validity of this result.

Since the average linkage method provides higher cophenetic correlations for all selected plot sizes than the other clustering methods (Table 1), it is selected as the most robust clustering method for forest data analysis. According to the species composition, the two-dimensional graphical representation of the proximities of the 25 quadrats of size 20 × 40 m^2 can be represented

Table 1: CPCC for nested plots

Plot size	5 × 5 m^2	5 × 10 m^2	10 × 10 m^2	10 × 20 m^2	20 × 20 m^2	20 × 40 m^2
Complete	0.744	0.717	0.856	0.844	0.832	0.763
Ward	0.617	0.632	0.683	0.747	0.755	0.661
Average	0.890	0.891	0.878	0.872	0.867	0.808
Single	0.778	0.849	0.796	0.786	0.678	0.627



(a) NMS analysis for full dataset

(b) CA analysis for full dataset

(c) PCoA analysis for full dataset

(d) DCA analysis for full dataset

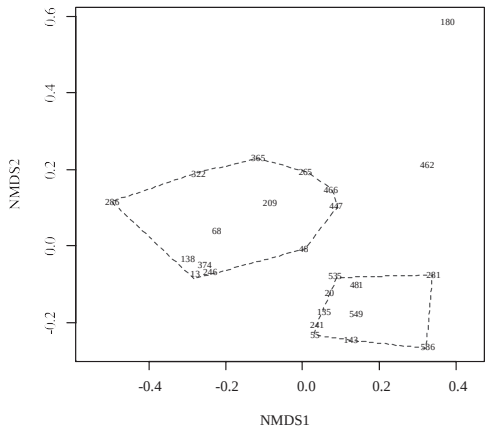
Figure 2: Four different ordination plots representing NMS, CA, PCoA and DCA ordinations

Table 2: Correlation between ordination methods in procrustes rotations, based on 999 permutations

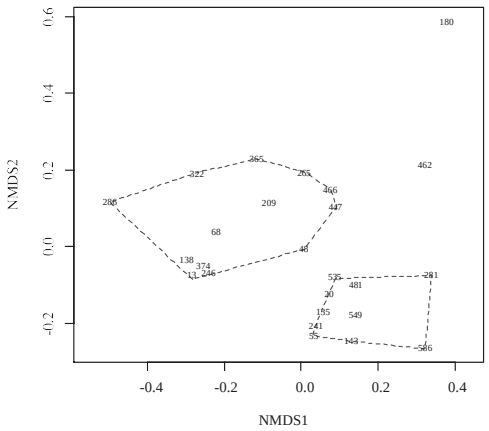
Configuration	Correlation coefficient
PCoA × NMS	0.710
CA × NMS	0.901
DCA × NMS	0.896
CA × PCoA	0.694
DCA × PCoA	0.753
CA × DCA	0.829

Table 3: Procrustes sum of squares of pair-wise rotations. Read left to right

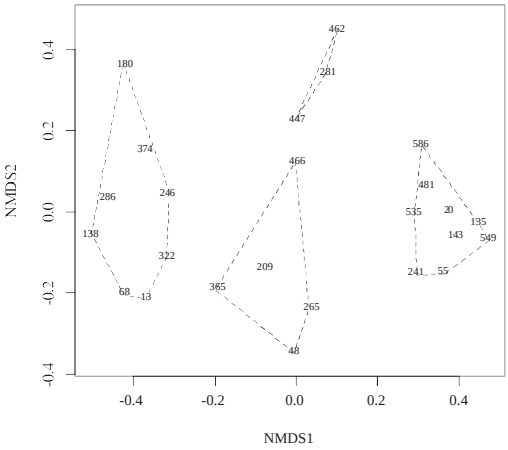
	NMS	CA	DCA	PCoA
NMS		0.504	0.529	1.324
CA	9.422		15.600	25.900
DCA	7.591	11.980		16.590
PCoA	1.203	1.258	1.049	



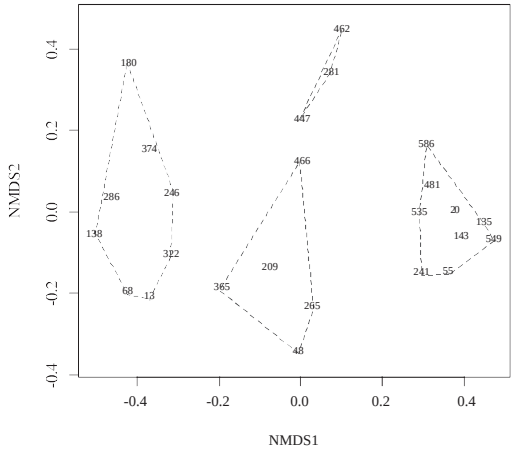
(a) NMS analysis of 5 × 5 m² plot



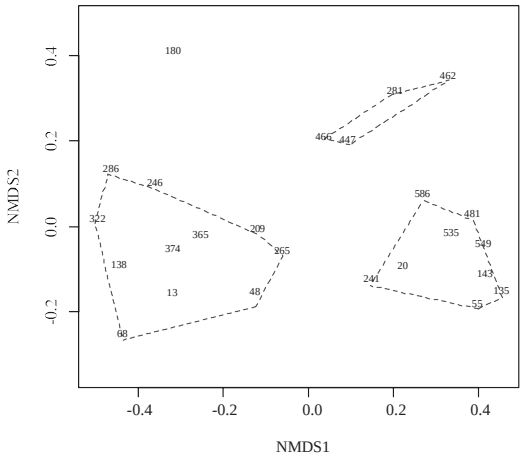
(b) NMS analysis of 5 × 10 m² plot



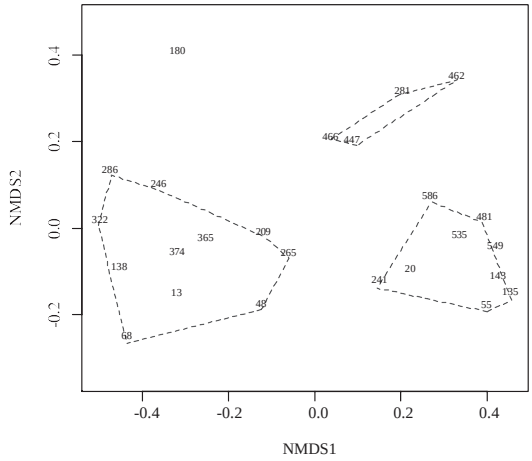
(c) NMS analysis of 10 × 10 m² plot



(d) NMS analysis of 10 × 20 m² plot



(e) NMS analysis of 20 × 20 m² plot



(f) NMS analysis of 20 × 40 m² plot

Figure 3: Comparison of similarities between six plots using NMS ordinations

by ordination plots. These ordination plots are drawn by using NMS, CA, PCoA and DCA techniques to the solution given by average linkage method to verify the results graphically (Figure 2).

The axes are the first and second ordination axes for each method, and the numbers shown in the ordination plots are the ID numbers of the quadrats. Visual comparison of ordination plots revealed that the ordination methods generate different results for the same dataset.

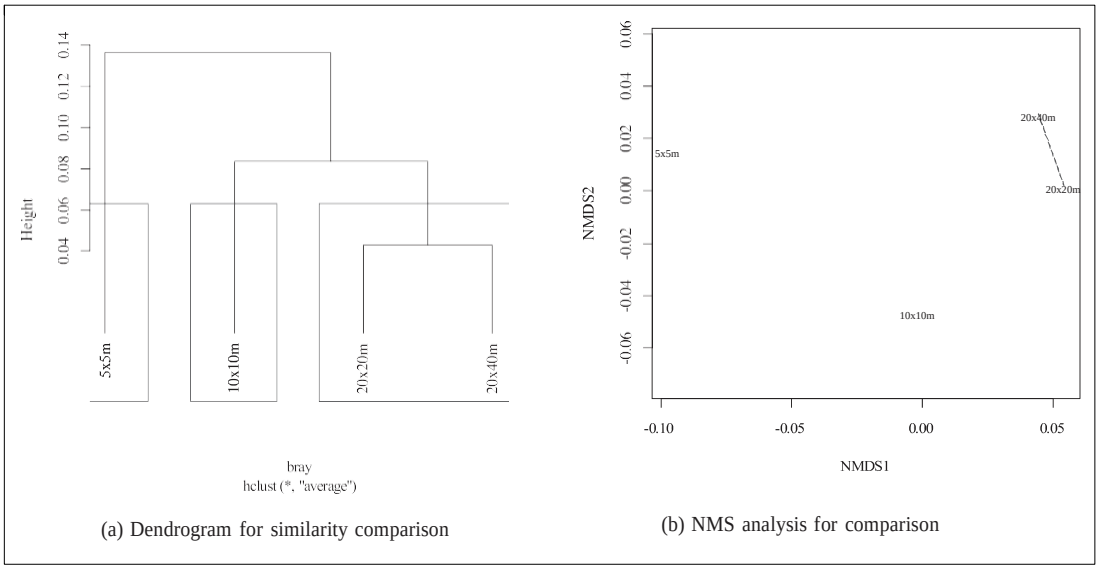
To identify the most suitable ordination plot, procrustean rotation was applied. The procrustean rotation rotates a solution to maximize the similarity with another solution, and it also provides the correlation coefficient between the solutions. According to the procrustean rotations CA and NMS pair has the highest correlation (Table 2), and the procrustean sum of squares of pairwise rotations show that NMS has the lowest sum of squares with CA (Table 3). When comparing the ordination plots drawn by CA and NMS methods, CA arranges data into a compact arc (Figure 2b). However the ordination plots drawn by using NMS show clear differences (Figure 2a) among all 25 quadrats compared to the ordination plots drawn by using CA. Based on these results it is noted that NMS is the most robust ordination method for forest data analysis. Due to the limitation of time and cost, if the nested plots have a species distribution similar to 20 × 40 m^2 plot (homogeneities between sites), a smaller plot size is preferred as the optimal plot size. The chi-square goodness-of-fit test (Table 4) is used to test whether the nested plots have the same species distribution.

Table 4: Chi-square goodness-of-fit test

Plot size	5 × 5 m^2	5 × 10 m^2	10 × 10 m^2	10 × 20 m^2	20 × 20 m^2
Chi-square	82.01	212.98	376.58	829.77	1439.81
p value	2.90×10^{-8}	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$

Table 5: Relative variance per unit area (25 m^2)

Plot size	5 × 5 m^2	5 × 10 m^2	10 × 10 m^2	10 × 20 m^2	20 × 20 m^2	20 × 40 m^2
Mean	17.16	18.52	19.19	19.45	19.86	20.79
Variance	64.00	82.08	75.52	84.01	74.30	43.43
Relative variance	1.47	1.89	1.74	1.94	1.71	1.00



(a) Dendrogram for similarity comparison

(b) NMS analysis for comparison

Figure 4: Dendrogram and NMS analysis for similarity between 20 × 20 m^2 and 20 × 40 m^2 size plots

Table 6: Mantel correlation

Plot size	Mantel correlation	p value
5 × 5 m^2	0.66	0.001
10 × 10 m^2	0.82	0.001
20 × 20 m^2	0.89	0.001

The goodness-of-fit test shows that the nested plots do not follow a species distribution similar to 20 × 40 m^2 plot (p value < 0.05). Since we have already identified the average linkage method as the most robust clustering method and the NMS method as the most robust ordination method, we apply these techniques to all nested plots to verify the results obtained by the chi-square goodness-of-fit test.

Visual comparison of the average clustering results and the corresponding ordination results (Figure 3) verified the heterogeneity between sites. These results indicate that the optimal plot size for sampling a tropical low land rainforest is 20 × 40 m^2 compared to the other selected plot sizes.

The relative variance of species cover per unit area (Wiegert, 1962) is used to compare the species distribution of different plot sizes. The lower the relative variance, the higher the species richness. Since the plot sizes are different, it is essential that the data from all plot sizes be standardized to a single unit area (per m^2). In this study since the minimum plot size is 5 × 5 m^2 (25 m^2), the relative variance (variance/minimum variance) is calculated for 25 m^2. The relative variance of species cover per unit area for all plot sizes are given in Table 5. Note that the relative variance of species cover per unit area is the lowest for 20 × 40 m^2 plot, and hence this comparison further validates that the 20 × 40 m^2 plot is the most suitable plot size compared to the other selected plot sizes to analyze the species composition of forest data.

Since 20 × 40 m^2 is a rectangular plot, a researcher may be interested in knowing whether he can use a square shaped plot with the same species composition. The relationship between the square shape plots; 5 × 5 m^2, 10 × 10 m^2 and 20 × 20 m^2 with the 20 × 40 m^2 plot can be identified by applying the Mantel test (Table 6). The Mantel test showed that there is a significant relationship (p value = 0.001) between the 5 × 5 m^2, 10 × 10 m^2 and 20 × 20 m^2 plots with 20 × 40 m^2 size plot.

Mantel correlation coefficients indicated that the species distribution of 20 × 40 m^2 plots highly correlate with 20 × 20 m^2 plots (Mantel correlation = 0.89). Visual comparison of average clustering results and the ordination results verified the similarity between 20 × 20 m^2 plots and 20 × 40 m^2 plots (Figure 4).

CONCLUSION

The most suitable hierarchical clustering method and the ordination method for analyzing the species in a tropical low land rainforest is the average linkage method and non-metric multidimensional scaling. The optimal plot size, which represents the highest number of species cover per unit area is 20 × 40 m^2 compared to the other selected plot sizes. When considering the optimal plot size for a tropical low land rainforest based on the quadrat shape, 20 × 20 m^2 is the most suitable square plot size, and 20 × 40 m^2 is the most suitable rectangular plot size.

REFERENCES

1. Barkman J.J. (1989). A critical evaluation of minimal area concepts. *Vegetatio* **85**: 89 – 104. DOI: http://dx.doi.org/10.1007/BF00042259
2. Bormann F.H. (1953). The statistical efficiency of sample plot size and shape in forest ecology. *Ecology* **34**: 474 – 487. DOI: http://dx.doi.org/10.2307/1929720
3. Clarke K. & Ainsworth M. (1993). A method of linking multivariate community structure to environmental variables. *Marine Ecology Progress Series* **92**: 205 – 219. DOI: http://dx.doi.org/10.3354/meps092205
4. Dassanayake M.D. & Fosberg F.R. (1980 – 2000). *A Revised Hand Book to the Flora of Ceylon*, volumes 1 – 12. Amarind Publishing, New Delhi, India.
5. Goldsmith F.B., Harrison C.M. & Morton A.J. (1986). Description and analysis of vegetation. *Methods in Plant Ecology* (eds. P.D. Moore & S.B. Chapman), pp. 437 – 524. Blackwell Scientific Publications, Oxford, UK.
6. Gunatilleke C.V.S., Gunatilleke I.A.U.N., Athugala A.U.K. & Esufali S. (2004). *Ecology of Sinharaja Rain Forest and the Forest Dynamics Plot*, pp. 221. WHT Publications (Pvt.) Ltd., India.
7. Gunatilleke C.V.S., Gunatilleke I.A.U.N., Esufali S., Harms K.E., Ashton P.M.S., Burslem D.F.R.P. & Ashton P.S. (2006). Species-habitat associations in a Sri Lankan dipterocarp forest. *Journal of Tropical Ecology* **22**: 371 – 384. DOI: http://dx.doi.org/10.1017/S0266467406003282
8. Hubbel S.P. & Foster R.B. (1983). Diversity of canopy trees in a neotropical forest and implications for conservation. *Tropical Rain Forest: Ecology and Management* (eds. S.J. Sutton, T.C. Whitmore & A.C. Chadwick), pp. 25 – 41. Blackwell Scientific, Oxford, UK.
9. Kenkel N.C. & Podani J. (1991). Plot size and estimation efficiency in plant community studies. *Journal of Vegetation Science* **2**: 539 – 544.

DOI: http://dx.doi.org/10.2307/3236036
10. Krebs C.J. (1999). *Ecological Methodology*, pp. 620. Addison-Wesley Educational Publishers, Inc., Menlo Park, California, USA.
11. Legendre L. & Legendre P. (1998). *Numerical Ecology*. Elsevier Science, Amsterdam, The Netherlands.
12. Manokaran N., Lafankie J.V., Kochummen K.M., Quah E.S., Elahn J.E., Ashton P.S. & Hubbell S.P. (1992). *Stand Table and Distribution of Species in the 50 ha Research Plot at Pasoh Forest Reserve*, pp. 454. Forest Research Institute of Malaysia, Kepong, Malaysia.
13. Minchin P.R. (1987). An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio* **69**: 89 – 107.
DOI: http://dx.doi.org/10.1007/BF00038690
14. Moravec J. (1973). The determination of the minimal area of phytocoenises. *Folia Geobotanical and Phytotaxonomica* **8**: 23 – 47.
15. Økland R.H. (1990).Vegetation ecology: theory, methods and applications with reference to Fennoscandia. *Sommerfeltia Supplement* **1**: 1 – 233.
16. Quinn G. & Keough M. (2002). *Experimental Design and Data Analysis for Biologists*, pp. 509. Cambridge University Press, Cambridge, UK.
17. Ruokolainen L. & Salo K. (2006). Differences in performance of four ordination methods on a complex vegetation dataset. *Annales Botanical Fennici* **43**: 269 – 275.
18. Singh W., Hjorleifsson E. & Stefansson G. (2010). Robustness of fish assemblages derived from three hierarchical agglomerative clustering algorithms performed on Icelandic ground fish survey data. *ICES Journal of Marine Science* **10**: 1 – 12.
19. Wiegert R.G. (1962). The selection of an optimum quadrat size for sampling the standing crop of grasses and forbs. *Ecology* **43**: 125 – 129.