Taxonomy and Conservation: Field Identification of Dipterocarps

Chen Beijia, Kee Jing Ying, Yao Chang
Raffles Institution
Singapore, Singapore

Abstract - This paper briefly describes two methods used to differentiate leaves of five species, *Anisoptera megistocarpa, Dipterocarpus grandiflorus, Hopea sangal, Shorea pauciflora* and *Vatica maingayi,* that belong to the Dipterocarpaceae family. The two methods are: a) the traditional approach, namely, a dichotomous key and b) the computational method which consists of the use of Gray Level Co-occurrence Matrix (GLCM). Through the two approaches, we then compare the accuracy rates in identifying the leaf species by running tests using both methods. GLCM achieved an overall identification accuracy of 93.3%, while the dichotomous key achieved an average accuracy rate of 85%.

Keywords: dichotomous key, Gray Level Co-occurrence Matrix

1. Background and Purpose of Research

Asian dipterocarps constitute prominent elements of the lowland rain forest (Whitmore 1988) and are also well represented in the understorey. Most belong to the mature phase of primary forest, which contains most of the entire genetic stock (Jacobs 1988). Species diversity and structural diversity are high in this forest and canopy cover is continuously barring tree falls ("Terrestrial", 2014). In view of the ecological importance of the dipterocarps, it is crucial for forest rangers and biologists to track the health of dipterocarps for conservation purposes. To do so, they would first need to identify the species.

Hence, the aim of this project is to compare between two methods of species identification: the dichotomous key and the computational method; so as to better aid people in identifying and understanding the distribution of dipterocarps.

The flowers and fruits of dipterocarps are little use for field identification because they are frequently unavailable - majority of dipterocarps do not flower regularly (Ghazoul, J., 2016). Instead, field characters like leaf and bark, which are readily observable, are commonly used by foresters in field identification. However, bark is not used in our project as the actual colour of the bark may be affected by lichen or moss covering it; also, the tree may be in an inaccessible location. On the other hand, fallen leaves cover a larger area and can be found more easily.

2. Hypothesis

The usage of GLCM features as a computational approach to identify *Hopea sangal, Shorea pauciflora, Anisoptera megistocarpa, Dipterocarpus grandiflorus* and *Vatica maingayi* achieves a higher accuracy than the usage of a dichotomous key.

3. Method and materials

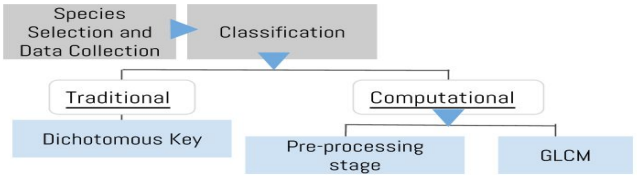


Fig 1. Overview of methodology

3.1. Species selection and Data collection

The sample species were selected based on three criteria, using the Checklist of Total Vascular Flora of Singapore. The first criterion is the native status of each species. All the chosen species are locally cultivated so that educators and amateur botanists would be able to identify the trees they see in Singapore. The second criterion is the conservation status of the species. The species we picked were critically endangered because they were most in need of identification and protection. Finally, different genera of dipterocarps were chosen so that it would be more representative of the dipterocarpaceae family.

For the collection of samples for our database, we visited the herbarium to photograph specimens of our target species using a Canon EOS 1000D camera. We also collected leaves from Lornie Trail and Petaling Trail of MacRitchie Reservoir and the rainforest trail of Botanic Gardens, under the supervision of Dr Shawn Lum from the Asian School of the Environment, NTU, who is a tropical rainforest ecologist. With his expertise, the leaves could be accurately identified. From the leaves we collected, we then chose leaves with leaf shapes that were typical of their species, to be tested using the dichotomous key and computational method. Also, natural deformation was only to a minor extent such that it does not affect the overall appearance of the leaf.

3.2. Methods of Classification

3.2.1. Dichotomous key

Firstly, we referred to the books: Trees of Tropical Asia (LaFrankie, 2010) and Foresters' manual of Dipterocarps (C.F. Symington, 2004) for a detailed explanation and

illustration of various features of the leaves from the 5 species we have selected. We constructed a table of comparison for our 5 chosen species to better visualize the differences between their features.

Then, we supplemented this information with our observations of the leaves collected during field trips as well as observations of the photographs taken in the herbarium. Following which, the leaves were grouped in clusters according to their similarities and differences. Since the dichotomous key involves either/or choices, we selected characteristics that isolate one species from the rest to construct our dichotomous key. Illustrations of the leaves were added in to better aid people in visualizing the various characteristics of the leaves. After deriving the dichotomous key, it was tested on 10 students who were asked to identify 7 leaf samples from *Hopea mengarawan, Shorea pauciflora, Shorea gratissima, Anisoptera megistocarpa, Dipterocarpus grandiflorus, Vatica maingayi,* and a non-dipterocarp species. We replaced *Hopea sangal* with *Hopea mengarawan* because we were unable to obtain leaf samples for *Hopea sangal*, and leaves of the same genus generally share similar characteristics.

3.2.2 Computational method

3.2.2.1. Pre-processing

For leaves collected during the field trips, they were pressed using a plant press and placed in an incubator at 60℃ to ensure that they were flat and not curled up. The flat leaves were then photographed using the Canon EOS 1000D Camera. For leaves taken at the herbarium, cropping was done to remove background noises. We also ensured that photos were taken under sufficient lighting.

3.2.2.2. Gray Level Co-occurrence Matrix

We used Matlab, a computing software, as the tool to analyse the texture of leaves through Gray Level Co-occurrence Matrix (GLCM). The GLCM is a statistical method of examining the textures that considers the spatial relationship of the pixels ("Texture Analysis", 2016). The image was first converted from RGB to grayscale. Next, four parameters of the GLCM: Contrast, correlation, homogeneity and energy, derived from the pre-processed sample images were calculated using Matlab.

A k-fold cross validation (k=5 is used in our case) was then done to estimate how accurately the predictive model would perform in practice. In k-fold cross validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k−1 subsamples are used as training data. The cross-validation process is then repeated k times (the *folds*), with each of the k subsamples used exactly once as the validation data (Hastie, 2009). In each individual test, the average values of the four parameters of the training samples were calculated, as represented by W_i, X_i, Y_i and Z_i (i=1,2,...5). Each species thus has a set of the four parameter values representative of the image features of that particular species. The Euclidean distance, d, between a matrix comprising of the four values and that of the training sample was calculated. The smallest d suggests the greatest resemblance of the testing image to that particular species. Thus, the system will classify the testing image as that species.

A matrix with columns and rows representing classified species and ground truth respectively was then constructed. This matrix translated into a confusion matrix showing True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Average accuracy, precision and recall were calculated for each of the five matrices. Results are recorded in the following section.

4. Results & Discussion

4.1. Dichotomous key

Full dichotomous key is presented in *Appendix A*.

Table 1: Survey results on students (Columns represent Classified species, Rows represent Ground Truth)

C / GT	A.M.	D.G.	H.M.	S.P.	V.M.	Non-dipterocarps
A.M.	10					
D.G.		9				1
H.M.		1	2	1	5	1
S.P.		2		7	1	
V.M.	3		1	1	4	1
Non-dipterocarps	2	6			1	1

Table 2: Confusion matrix for Dichotomous Key

	TP	TN	FP	FN
A.M.	10	45	5	0
D.G.	9	41	9	1
H.M.	2	49	1	8
S.P.	7	48	2	3
V.M.	4	43	7	6
Non-dipterocarps	1	47	3	9

In this survey, we assumed that the leaf was classified correctly if its genus could be correctly identified.

The average accuracy rate of the dichotomous key is 85%, calculated using the formula $\frac{TP+TN}{TP+TN+FN+FP}$. This high accuracy rate was achieved probably because we picked species with distinctive features, especially for *Dipterocarpus* and *Anisoptera.*

Dipterocarpus had a high species accuracy rate of 90%. The species accuracy rate was calculated by: the number of correct responses/the total number of responses for a species. This indicates that the wavy leaf edges are very distinctive of the species and can be easily observed.

However, it has the highest false positive rate due to similarities with the non-dipterocarp species in terms of leaf edge. This could possibly be due to the subjective perception of how 'wavy' is defined.

Anisoptera had the highest species accuracy rate of 100%. This is because other than the difference in secondary veins between *Anisoptera* and the other 3 species, there are other features distinctive to the species, such as its oblong leaf shape with a tapered tip and also, its hairy surface.

5 students inaccurately identified the leaf belonging to *Hopea* as belonging to *Vatica*. This suggests that the differences in their leaf blades as stated in the dichotomous key was not easily observable.

It is noted that the species accuracy rate of the identification of *Hopea* and *Vatica*, which was 20% and 40% respectively, was much lower than that compared to *Anisoptera*, *Shorea*, and *Dipterocarpus*. This could be due to *Hopea* and *Vatica* being located the lowest in the key, thus increasing the probability of errors made when reading the previous levels of the key. Another reason is because as the user progresses down the dichotomous key, the similarities between the species increases.

The test on *Shorea gratissima* only achieved 10% accuracy rate (*Appendix B*) , as compared to 70% for *Shorea pauciflora*. This highlights that even within the same genus, there could be variation among the species. Therefore, the dichotomous key would be less accurate for genera with characteristics which are not uniform across all species.

4.2. Gray Level Co-occurrence Matrix

Table 3. Results for average accuracy, error rate, precision and recall in the Top 1 Test

K-fold	Accuracy	Precision	Recall
Set A	0.725	0.183	0.300
Set B	0.775	0.400	0.530
Set C	0.812	0.517	0.500
Set D	0.694	0.127	0.250
Set E	0.617	0.113	0.188
Average	0.725	0.268	0.354

Low precision and recall rates were observed in the Top 1 Test. We also noticed that when calculating the Euclidean distance, *d*, in a few cases, the images belong to species which rendered the second smallest value of *d*, suggesting a great similarity between the correct species (groundtruth) and the classified species. Thus, the Top 2 Test and Top 3 Test were then carried out, meaning that as long as the testing image belongs to one of the two or three classes (in the Top 2 and Top 3 Test respectively) that the image is most likely to fall under, one count will be added to the True Positive category. The new values for average accuracy, precision and recall can be found in *Appendix C*.

A 14% increase in accuracy rate was observed in the Top 2 Test while another 7.4% increase was seen in the Top 3 Test. The sharp rise in the Top 2 Test is likely to be due to the closer similarity between the top two species that the selected leaf is most likely to fall under. In the Top 2 Test and the Top 3 Test, there are also increments in both Precision and Recall, which suggests a larger fraction of retrieved images that are relevant and a larger fraction of relevant images that are retrieved (Powers, 2011).

Comparing our computational method with Leafsnap, the first mobile app for identifying plant species using automatic visual recognition, the highest accuracy achieved in our test, 93.9%, is lower than that using Leafsnap, which is 96.8%. The Leafsnap system uses the distinctive shapes of leaves as the sole recognition cue to identify species from a dataset of 184 trees in the Northeastern United States (Neeraj Kumar, 2012). However, this is unable to be replicated for dipterocarps as their leaves are described to be generally oblong. Thus, to differentiate dipterocarps' leaves, GLCM was utilized to analyze the texture of the leaves, which could possibly lower the accuracy rate as texture is a subtler feature than curvature and thus the differences in texture among species is less distinct than the differences in shape.

4.3. Comparison between Dichotomous key and Computational method

The accuracy using the dichotomous key is 85% while the top 3 accuracy for GLCM is 93.9%. While for the dichotomous key, *Dipterocarpus* can be easily differentiated from *Anisoptera* due to its distinctive corrugated surface and wavy edges, the computational method has difficulty in differentiating these two species as majority of the discrepancies were due to the confusion between these two species (about 60%).

When using the dichotomous key, many aspects of the leaves can be considered, for example: texture, leaf shape, venation, etc. and participants are able to synthesize the different kinds of information about the leaf to make a decision. However, GLCM used in the computational method is a global feature which is representative of the texture only.

For the dichotomous key, student respondents took a long time to complete their identification of the leaves, and this would be inconvenient when they are out in the field. This can be overcome by the computational method, which would generate results at a faster speed.

In addition, respondents using the dichotomous key would automatically try to classify an unknown leaf as one of the five species if no instruction was given, as they can only

make choices present within the dichotomous key. This can be overcome in GLCM by setting a threshold value for the Euclidean distance, *d* such that when the testing image gives a *d* value larger than the threshold number, the image will then become an 'outlier' and classified as unknown.

5. Conclusion and Future Works

This project compares two methods of identification in field taxonomy of dipterocarps, namely, the dichotomous key and the computational method. The project itself has the potential to aid researchers in their work, as well as generate interest and increase awareness among the public about dipterocarps. It could also allow for better conservation of dipterocarps by tracking their distribution in Singapore.

This result affirms our hypothesis that GLCM as a computational method is more accurate than the dichotomous key.

For future works, we could expand our database by including more genera belonging to the dipterocarpaceae family to broaden our scope, or more species within the same genus to improve specificity. In addition, Artificial Neural Networks (ANN) can be utilized to train the programme in classification instead of K-fold cross validation to achieve higher accuracy.

Finally, we propose a semi-automated approach to capitalize on the merits of both methods. We can create a digitized version of the dichotomous key, where the user will be brought to the next level of the key automatically after answering either/or questions. Their results will be double-checked by another computational system involving the use of GLCM as well as other feature extraction methods such as Centroid Contour Distance (CCD) in order to improve the accuracy rate.

ACKNOWLEDGMENT

We would like to thank Dr Tan Guoxian and Dr Jeffrey Lee, our mentors, for their constant guidance and support throughout this project.

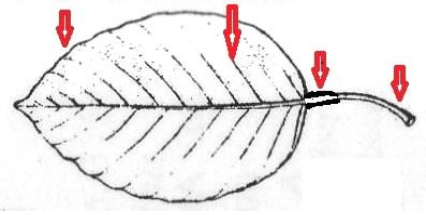
REFERENCES

1. LaFrankie, J. V. (2010). *Trees of tropical Asia: an illustrated guide to diversity*. Philippines: Black Tree Publications.
2. Hastie, & Tibshirani. (2009, February 25). K-Fold Cross-Validation. Retrieved August 2, 2016, from http://statweb.stanford.edu/~tibs/sta306bfiles/cvwrong.pdf
3. Appanah, S. & Turnbull, J. (1998). *A review of dipterocarps* (1st ed.). Kuala Lumpur: Center for International Forestry
4. *Terrestrial*. *National Parks Board*. Retrieved 19 November 2016, from https://www.nparks.gov.sg/biodiversity/our-ecosystems/terrestrial
5. Ghazoul, J. *Dipterocarp biology, ecology, and conservation* (1st ed.). United Kingdom: Oxford University Press.
6. Stork, D., Duda, R., & Yom-Tov, E. (2004). *Computer manual in MATLAB to accompany Pattern classification* (1st ed.). Hoboken, NJ: Wiley-Interscience
7. *Properties of gray-level co-occurrence matrix - MATLAB graycoprops*. *Mathworks.com*. Retrieved 11 October 2016, from https://www.mathworks.com/help/images/ref/graycoprops.html
8. *Texture Analysis Using the Gray-Level Co-Occurrence Matrix (GLCM)*. Retrieved 3 January 2017, from https://www.mathworks.com/help/images/gray-level-co-occurrence-matrix-glcm.html
9. Yan Chong, K., T. W. Tan, H., & T. Corlett, R. (2009). *A CHECKLIST OF THE TOTAL VASCULAR PLANT FLORA OF SINGAPORE* (1st ed.). Singapore: Raffles Museum of Biodiversity Research, National University of Singapore. Retrieved from https://lkcnhm.nus.edu.sg/nus/pdf/PUBLICATION/LKCNH%20Museum%20Books/LKCNHM%20Books/flora_of_singapore_tc.pdf
10. Symington, C. (1974). *Foresters' manual of dipterocarps* (1st ed.). Kuala Lumpur: Penerbit Universiti Malaya.
11. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). *Journal of Machine Learning Technologies*. **2** (1): 37–63

Appendix A

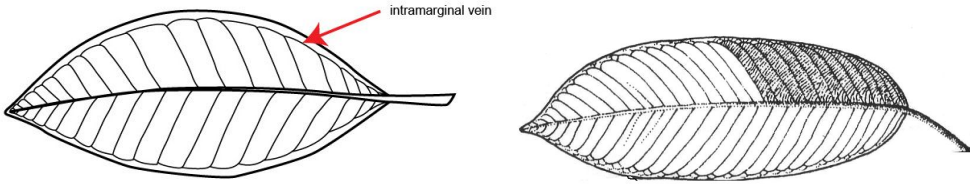
Dichotomous key to five species of Dipterocarpaceae based on leaf characters

a. Leaf edges are wavy, leaf surface folded, ~23 strong raised secondary veins nearly touch margin, leaf stalk long (~7 cm), strongly swollen and bent at top ……..... *Dipterocarpus grandiflorus*



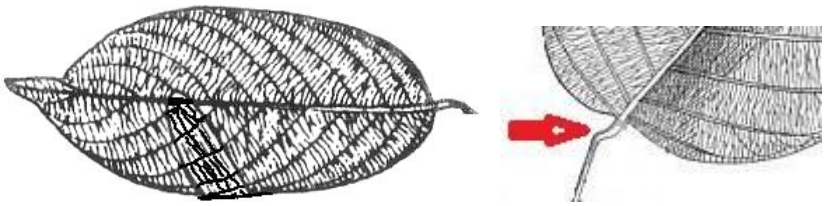
aa. Leaf edges are smooth not wavy, leaf surface smooth not folded, ≤ 12 or ≥ 30 secondary veins curve away a distance from margin, leaf stalk < 3 cm and not strongly swollen or bent at top

b. Secondary veins form looped intramarginal vein, ~30-31 secondary veins, leaf shape oblong with tapered tip. Only strong raised regular spreading main secondary veins looped, without shorter looped intermediary veins, coarsely hairy …...…….…… *Anisoptera megistocarpa*



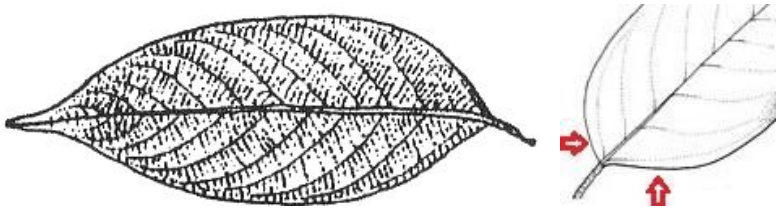
bb. Secondary veins curved but do not converge into loops to form intramarginal veins, ~9-12 secondary veins, leaf shape range from oblong, elliptic, ovate to lanceolate, with tapered tip

c. Petiole strongly twisted, ~1.8 cm long, 9 secondary veins, tertiary veins inconspicuous ……...……..…….. *Shorea pauciflora*

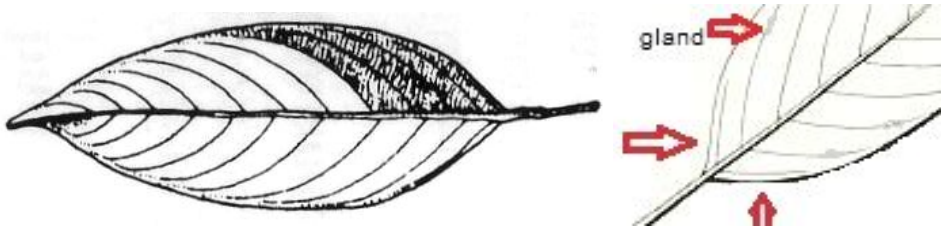


cc. Petiole not or only somewhat twisted, 1.2 - 1.5 cm long, ≥ 10 secondary veins, tertiary veins conspicuous

d. Leaf blade at leaf base unequal in size on either side of midrib, both midrib and 12 secondary veins very strongly raised beneath, tertiary veins join secondary veins in parallel wavy ladder-like (scalariform) venation pattern, petiole ≤ 1.2 cm, slender and somewhat twisted ……...…...…….…. *Hopea sangal*



dd. Leaf blade at leaf base about equal on either side of midrib, midrib prominent beneath but 10 secondary veins (with glands near ends) more faint, tertiary veins form netlike (reticulate) venation pattern, petiole ~1.5 cm, slender but slightly thickened on upper half ..……...……..…… *Vatica maingayi*



Appendix B

Table 1. Results of test for Shorea Gratissima

GT \ C	A.M.	D.G.	H.M.	S.P.	V.M.	Non-dipterocarp
S.G.		1	3	1	3	2

Appendix C

Table 1. Accuracy, Precision and Recall in Test 2

K-fold	Accuracy	Precision	Recall
Set A	0.850	0.383	0.533
Set B	0.850	0.680	0.713
Set C	0.918	0.760	0.700
Set D	0.906	0.631	0.650
Set E	0.800	0.604	0.675
Average	0.865	0.611	0.654

Table 2. Accuracy, Precision and Recall in Test 3

K-fold	Accuracy	Precision	Recall
Set A	0.875	0.533	0.720
Set B	0.900	0.770	0.833
Set C	0.976	0.960	0.967
Set D	0.976	0.760	0.800
Set E	0.967	0.917	0.950
Average	0.939	0.788	0.854