Legume diversity as indicator for botanical diversity on Sundaland, South East Asia

N. Raes a,⁎, L.G. Saw b, P.C. van Welzen a, T. Yahara c

a Naturalis Biodiversity Center — Section Botany, P.O. Box 9517, NL-2300RA Leiden, The Netherlands
b Forest Research Institute Malaysia (FRIM), Kepong, 52109 Kuala Lumpur, Malaysia
c Center for Asian Conservation Ecology, Kyushu University, 6-10-1 Hakazaki, Fukuoka 812-8581, Japan

A R T I C L E   I N F O
Available online 5 July 2013
Edited by B-E Van Wyk

Keywords:
Global Legume Diversity Assessment
Biodiversity
Legumes
Fabaceae
Leguminosae
Sundaland
Dipterocarpaceae
Indicator
Species distribution model

A B S T R A C T
The Global Legume Diversity Assessment (GLDA) proposes the legume family (Fabaceae or Leguminosae) – one of the largest and economically important plant families – as a target for a global botanical diversity assessment project. Where in the Neotropics and Africa legumes dominate the rain forest in terms of diversity and abundance, the Dipterocarpaceae claim this role in South East Asia and on Sundaland in particular. This raises the question whether legumes are an indicator for overall botanical diversity on Sundaland? To answer this question we use the largest compiled database of collection records of the region and species distribution modelling techniques. As a proxy for total botanical diversity we selected seven plant families; Dipterocarpaceae, Ericaceae, Fabaceae, Lauraceae, Moraceae, Myristicaceae, and Sapindaceae. Although the legumes were the most diverse family, the predictive power of legume diversity for overall botanical diversity was poor. This related to the fact that the other seven selected families largely represent trees, whereas legume species more equally represent all different growth forms. After assigning individual legume species to different growth habits (tree, liana, herb, miscellaneous) we were able to predict 78% of the variance in botanical diversity on Sundaland. The lianas represent the single growth habit that best predicted (66%) the variance in botanical diversity. The herb- and miscellaneous growth habits had an inverse relationship to botanical diversity. Legumes can be used as a predictor of overall botanical diversity in tropical and seasonal rain forests, but the relationship should be fitted for different biogeographic regions individually.

© 2013 The Authors. SAAB. Published by Elsevier B.V. Open access under CC BY license.

1. Introduction

It is widely recognized that plants, directly and indirectly, support important ecosystem services (ES) such as purification of air and water, mitigation of droughts and floods, pollination of crops and natural vegetation, dispersal of seeds, and control of the vast majority of potential agricultural pests, to name but a few (Daily et al., 2000; Millennium Ecosystem Assessment, 2005; Carvalheiro et al., 2011; Garibaldi et al., 2013). Pivotal for many ES is high plant diversity (Iubbol et al., 2011; Hooper et al., 2012). Despite the recognized economic importance of ES (TEEB, 2009; Larigauderie and Mooney, 2010; TEEB, 2010), natural habits such as forests and wetlands that provide ES are being rapidly lost; as is the case in South East Asia (Langner et al., 2007; Sodhi et al., 2009; Miettinen et al., 2011) – the region of interest in this study. Consequently, many plant species are also being lost. It remains uncertain, however, at what rate we are losing plant species diversity. This uncertainty, and the need to assess states and trends of as many species as possible at a regional (Yahara et al., 2009) and global scale (GEO BON, 2010), has initiated the Global Legume Diversity Assessment (GLDA; Yahara et al., 2013). The GLDA proposes the legume family (Fabaceae or Leguminosae) – one of the largest and economically important plant families – as a target for a global botanical diversity assessment project.

Among the shared plant families on the tropical continents, the legumes stand out because they dominate the rain forests of both Africa and the Neotropics in terms of basal area and overall biomass (Gentry, 1988; Corlett and Primag, 2011). Throughout the Asian-Pacific region, and on Sundaland in particular, the Dipterocarpaceae claim this role. Nonetheless, 45% of the genera of the legume subfamily Papilionoideae are represented in the Asian-Pacific region. Furthermore, besides producing some of the largest canopy trees, the legumes are also an important family in terms of the number of climber-, or liana species (Corlett and Primack, 2011). Sundaland, or the Sunda Shelf, is the western part of the Malesian region (Raes and Van...
Welzen, 2009; Lohman et al., 2011) and covers the Malay Peninsula and the islands of Borneo, Sumatra, Java and Bali. The northern boundary is delimited at the Kangar–Pattani line on the Malay Peninsula (Van Steenis, 1950), which approximately coincides with the country border between Malaysia and Thailand (Fig. 1). Given the goals of the GLDA, and the current knowledge on the floristic position of legumes on Sundaland, we raised the following question: 'Is legume diversity indicative for overall botanical diversity on Sundaland, South East Asia?'

Much of the available biodiversity data is stored in herbaria and natural history museums throughout the world, and becomes available as digitized records at a still increasing rate (Graham et al., 2004). From previous studies (Hortal et al., 2007; Schulman et al., 2007; Loiselle et al., 2008), however, it has become clear that botanical collecting and the intensity of collecting, is spatially biased. Hence, plain counts of species records at a chosen raster resolution likely result in a map of collecting efforts, rather than that of biodiversity. Overcoming this problem and predicting the area of occupancy for species even for regions where no collections have been made, have led to the development of Species Distribution Models (SDMs; Guisan and Zimmermann, 2000; Franklin, 2009; Peterson et al., 2011). SDMs relate species’ occurrence data (derived from herbarium and/or natural history museum collections) to abiotic or eco-geographic predictors (such as climate and soil conditions), and interpolate the identified relationships to predict the potential distributions of species for an area of interest, here Sundaland. Although the distribution of species records might be spatially biased, this does not necessarily result in biased sampling of the explanatory (abiotic) variables; as was shown for Borneo (Raes et al., 2009), the centre of diversity and endemism of Sundaland (Roos et al., 2004), as well as for other regions in the world (Kadmon et al., 2004; Loiselle et al., 2008; Newbold, 2010). Under these conditions, SDMs can accurately estimate species abiotic niche dimensions and project these in their reciprocal geographic space resulting in species distribution maps. Stacking a large number of SDMs results in a spatial pattern of biodiversity (Raes et al., 2009; de la Estrella et al., 2012; Mateo et al., 2012; Zhang et al., 2012).

To represent overall botanical diversity in forested areas, we selected seven plant families that are characteristic for both lowland rain forest and montane forests on Sundaland. As representatives for lowland rain forest we selected the Dipterocarpaceae, Lauraceae, Moraceae, Myristicaceae and Sapindaceae families; and for montane forests the Ericaceae and Fagaceae families. The selection of plant families was based on their dominant or characteristic role in tropical lowland or montane ecosystems, and the large number of species and genera they represent. The aim of this research is to assess how accurately the diversity pattern of legumes predicts the overall botanical diversity pattern represented by the other seven selected families on Sundaland, South East Asia. Since the legumes represent a proportionally large number of herbs, weedy vines and lianas, compared to the seven selected plant families that are largely represented by trees, we also assess the fit for four different growth habits; trees, lianas, herbs, and miscellaneous (all legume species with a different habit than the former).

2. Material and methods

Assessing the biodiversity indicator status of legumes for Sundaland includes a number of steps that we describe in a workflow. First, we combined all specimen records from the databases of Naturalis Biodiversity Center (L), the Forest Research Institute Malaysia (KEP), and Singapore Botanic Gardens (SING) in one database. Second, we retrieved all records of legume species with at least one record on Sundaland from the combined database. Third, we retrieved all records of the seven other families that were chosen to represent botanical diversity. Fourth, we tested all species names for synonymy and merged all synonymised species records. Fifth, we selected all species with at least five unique records and developed Species Distribution Models (SDMs) for these species. Sixth, we stacked all thresholded SDMs of

Fig. 1. Legume diversity of Sundaland (inside polygon) based on 423 stacked SDMs.
legume species (see below), as well as the ones belonging to the seven other families, resulting in spatial patterns of legumes—and botanical diversity. Finally, we assessed the biodiversity indicator status of legumes using a linear model where botanical diversity is the dependent variable and legume diversity, or diversity of the four legumes habit classes, the explanatory variables.

2.1. Study area and collection data

Sundaland, or the Sunda Shelf, represents the western part of the Malesia (Raes and Van Welzen, 2009), which formed one continuous landmass during the glacial maxima of the Quaternary as a result of eustatic sea-levels that were 120 m lower than they are today (Hanebuth et al., 2000; Cannon et al., 2009; Woodruff, 2010). Sundaland is delimited by the Kangar–Pattani line in the North (Fig. 1) and by the Merrill-Dickerson/Huxley line in the East. The latter runs between the Philippines (incl. Palawan)/Sulawesi and Borneo, and between Bali and Lombok (Raes and Van Welzen, 2009). To prevent model errors caused by artificially defined political/geographical boundaries (Raes, 2012), we modelled the species in a wider geographical extent ranging from 11° South to 19° North, and between 92 and 127° East (see Fig. S13 for extent).

We selected all georeferenced records of legume species that had at least one record on Sundaland from the wider geographical extent. The same was done for the species belonging to the seven families that were selected to represent overall botanical diversity; Dipterocarpaceae (includes Meranti hardwood), Ericaceae (heather family), Fagaceae (beech family), Lauraceae (laurel family), Moraceae (mulberry- or fig family), Myristicaceae (nutmeg family) and Sapindaceae (soapberry family). All records of Sundaland for these seven families and the legumes stored at the Naturalis Biodiversity Center were recently digitized. All species names were synonymised as best as possible, given inconsistencies in names used in the different merged databases (Raes, 2012), we modelled the species in a wider geographical extent ranging from 11° South to 19° North, and between 92 and 127° East (see Fig. S13 for extent).

We used two datasets to extract the environmental predictors that were used for the MaxEnt models; first, the 19 bioclimatic variables of the WorldClim dataset (Hijmans et al., 2005) and second, the ISRIC-WISE derived soil properties dataset (version 1.2; Batjes, 2012). Both datasets have a spatial resolution of 5 arc-minutes. To prevent problems with multi-collinearity and model over fitting we retained all WorldClim predictors with a |Pearson’s r correlation| ≤ 0.7. For groups of correlated variables (|r| > 0.7) we performed a principal components analysis (PCA) on standardized values, and retained the variable with the highest vector loadings. This variable has the most variation, and hence is likely to be the most informative for the selected study area; here the wider geographical extent. The selected Worldclim variables were: Bio1 — annual mean temperature, Bio2 — mean diurnal range, Bio4 — temperature seasonality, Bio12 — annual precipitation, Bio13 — precipitation of wettest month, Bio15 — precipitation seasonality and Bio19 — precipitation coldest quarter. A similar procedure was followed for the ISRIC-WISE soil data, but here we used |Spearman’s rho rank correlation| ≤ 0.7. This was done because soil property values can abruptly change over short distances, unlike the bioclimatic gradients. Again, from groups of correlated variables we retained the variable with the highest vector loadings on the PCA axes. The data are not multivariate normal distributed, but PVA is known to be robust against this violation. Furthermore, we only used the PCA to inform about which variable to retain, the PCA vector loading values were not used in further analyses or conclusions. The selected ISRIC variables were: ALSA — exchangeable Aluminium percentage, BULK — bulk density, CEC — cation exchange capacity of clay fraction, CECs — cation exchange capacity, CFRAG — coarse fragments, CNrt — C/N ration, ELCO — electrical conductivity, ESP — exchangeable Na percentage, PHAQ — pH in water, and TAWC — total available water capacity.

SDMs build on Hutchinson’s duality, i.e. the reciprocal correspondence between ecological niche space and geographic space (Hutchinson, 1957). A species occurs at a particular location because the ecological niche requirements are fulfilled, and when a species’ ecological niche requirements are defined, it is possible to identify where in the landscape these conditions occur, hence reciprocal. However, whether a species actually occupies its entire ecological niche space also depends on the accessibility of suitable regions (Soberón and Nakamura, 2009). Dispersal barriers such as mountain ranges and seaways prevent species from colonizing all regions that have suitable abiotic conditions. Although the islands of Sundaland were connected during glacial maxima, the abiotic conditions on the surfaced Java Sea floor were relatively dry and consist of poor, coarse sandy soils, which consequently posed an effective dispersal barrier for most organisms (Bird et al., 2005; but see Cannon et al., 2009; Slik et al., 2011). This separation through time of the Sundaland islands has resulted in considerable levels of endemism (Roos et al., 2004; Van Welzen and Slik, 2009; Slik et al., 2011). We therefore included a variable that prevents island endemic species from being predicted to be present on islands where they were not collected, but where abiotic conditions are suitable. We constructed a presence/absence matrix of 8110 species treated in the Flora Malesiana I series (Anon., 1950–2010) and nine biogeographic units (Malay Peninsula, Sumatra, Borneo, Java, Philippines, Sulawesi, Lesser Sunda Islands, Moluccas and New Guinea) representing the major Malesian island groups (Fig. S1; Van Welzen and Slik, 2009). We assessed the floristic similarity between the nine biogeographic units with a detrended correspondence analysis (DCA) implemented in the function decorana of the R-package vegan (Oksanen et al., 2012). We used DCA axis 1 scores to effectuate dispersal limitation; i.e. all raster cells of one biogeographical unit receive the same DCA axis 1 score, i.e. all raster cells of Borneo receive value -1.3525, for the Malay Peninsula -1.2094, etc. (Table S1 and Fig. S1). This layer effectively prevents species with collections of only one island from being predicted to be present on the other islands, even if abiotic conditions are suitable. Furthermore, the degree of dispersal limitation few collection records are available (Wisz et al., 2008) and was therefore selected to develop the SDMs.
is determined by overall floristic similarity between biogeographical units.

We accounted for possible effects of sampling bias by adding a target-background sample layer (Phillips et al., 2009; Elith et al., 2011). As target-background sample we used all plant collection sites of the three databases (also including collection sites from other plant families) from the wider geographical extent (Fig. S13). This layer contained 7298 collection sites. SDMs were fitted making use of the target-background sample and were subsequently projected to the wider geographical extent, covering 59,388 raster cells.

All MaxEnt models were run using the default settings (version 3.3.3k; Phillips et al., 2006). The only deviation from default is that we used all records for model training, and did not partition the data into train- and test partitions. We calculated in-sample model performance by using the models to predict the training data used for fitting the models (Wenger and Olden, 2012). This allowed us to make use of all available information from the many species with scant collection records. As measure of model performance we used the ‘area under the receiver operating characteristic curve’ or simply the AUC value, a widely used and unbiased summary metric of model performance for binary data. AUC values run from zero to one when applied to presence/absence data, where a value of 0.5 indicates no better than random prediction, and value 1 perfect model fit. It should be noted, however, that the use of AUC values is unreliable when applied to presence-only data and a background sample. The reason for this is that the maximum AUC value is no longer 1, but $1 - \frac{\alpha}{2};$ where $\alpha$ stands for the true species distribution, which is typically not known (for details see Raes and ter Steege, 2007; Lobo et al., 2008; Phillips et al., 2009). The solution of testing against a null-distribution (Raes and ter Steege, 2007) was too computationally intensive. For every species the spatial distribution of records over the four Sundaland islands (and beyond for the wider geographical extent) should be mimicked by random draws and replicated 99 (or 999) times. Next, the 99 random sets should be modelled with MaxEnt in order to generate a null-distribution of AUC values against which to test the true species’ model AUC value for significance. We recognize that this is a compromise, but using 100% of the records for modelling has been applied before (de la Estrella et al., 2012; Mateo et al., 2012; Wenger and Olden, 2012). Moreover, the aim is to test how well legume diversity is a predictor for total botanical diversity. Both legume SDMs and the SDMs of the seven selected families suffer from the same compromise. We do report average AUC values and their standard deviations of the eight families for comparative purposes.

Finally, we converted the continuous MaxEnt predictions to discrete presence/absence values using the 10 percentile training threshold. This is a conservative threshold and predicts absence the 10% most extreme presence observations, or forces 10 percent of the collection presence records outside the predicted presence area. This procedure accounts for errors in taxonomic identifications and in georeferencing of the collection localities.

2.3. Data analyses

To assess the indicator status of the legumes for botanical diversity on Sundaland we stacked all thresholded legume SDMs, as well as the SDMs of the other seven families that represent botanical diversity to generate two patterns of ‘diversity’. We clipped the Sundaland extent from the wider geographical extent at the Kangar–Pattani line and excluded the Philippines, Sulawesi, the Moluccas, and Lesser Sunda Islands (Fig. 1). Sundaland covers 18,094 raster cells. We used an ordinary least squares (OLS) regression model to assess how well the legume diversity predicts botanical diversity. Additionally, we established the indicator status for ‘four’ different growth habits; trees, lianas, herbs, or miscellaneous. We assessed the indicator status of all legumes, for the different legume growth habits individually, and for the different legume growth habits simultaneously with a forward-backward stepwise multiple regression analysis and the Akaike Information Criterion (AIC) best model selection procedure. All analyses were performed in R (R Development Core Team, 2012). For the stepwise multiple regression we used the function stepAIC from the R-library MASS (Venables and Ripley, 2002). In the ecological literature, however, claims are made that residual spatial autocorrelation in models of spatially structured data (as is the case here) result in shifts in the partial coefficients, which bias the interpretation of factors influencing geographical patterns (Dormann et al., 2007). Yet, Hawkins et al. (2007) and Hawkins (2012) found that “OLS coefficients generated from data containing residual spatial autocorrelation were statistically indistinguishable from coefficients generated from the same data sets in which short-distance spatial autocorrelation was not present in all 22 coefficients tested. Consistent with the statistical literature on this subject, we conclude that coefficients estimated from OLS regression are not seriously affected by the presence of spatial autocorrelation in gridded geographical data”. These findings led us to plot the residuals of the best OLS model to identify regions where legumes under- and over-predict botanical diversity.

3. Results

The total dataset of legumes and the seven selected families contained 3004 species (742 legumes) of which 1720 species (423 legumes) were represented by at least five spatial unique records (Table 1). The latter represents the species that were modelled. From this group the most species rich family were the legumes, followed by the Dipterocarpaceae and Lauraceae. The average in-sample AUC values were all high, indicating that the SDMs accurately fit the collection data (Table 1).

The stacked legume SDMs indicate that the highest diversity is predicted on Java, with 210 species co-occurring in one raster cell (Fig. 1). The highest overall botanical diversity is found on Borneo, where 639 species are predicted to co-occur in one raster cell (Fig. 2). The diversity patterns of the different legume growth habits were very different; trees have a centre of diversity on east Borneo (Fig. S2), lianas in north-west Borneo (Fig. S3), and the herbs (Fig. S4) and miscellaneous growth habit (Fig. S5) both on Java. The separate diversity patterns of the seven families that represent overall botanical diversity are shown in Figs. S6–S12. These show a distinct pattern of diversity for each of the other seven families.

The OLS model between the dependent variable ‘overall botanical diversity’ and the independent variable ‘legume diversity’ had a poor fit ($R^2$-adj. 0.012; Table 2). The individual legume growth habits performed better, with the highest fit for the legume lianas ($R^2$-adj. 0.618; Table 2). The forward-backward stepwise multiple regression analysis indicated that the contribution of legume tree diversity became redundant when the other habits were added to the model (AIC 149,181 vs. 149,182 for the full model). This best model explained 78% of the variance in overall

<table>
<thead>
<tr>
<th>Family</th>
<th># species</th>
<th># species ≥ 5</th>
<th>Avg. AUC (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legumes</td>
<td>742</td>
<td>423</td>
<td>0.905 (0.057)</td>
</tr>
<tr>
<td>Tree</td>
<td>120</td>
<td></td>
<td>0.904 (0.056)</td>
</tr>
<tr>
<td>Liana</td>
<td>124</td>
<td></td>
<td>0.910 (0.053)</td>
</tr>
<tr>
<td>Herb</td>
<td>83</td>
<td></td>
<td>0.899 (0.053)</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>59</td>
<td></td>
<td>0.901 (0.065)</td>
</tr>
<tr>
<td>Dipterocarpaceae</td>
<td>414</td>
<td>335</td>
<td>0.938 (0.037)</td>
</tr>
<tr>
<td>Ericaceae</td>
<td>316</td>
<td>130</td>
<td>0.965 (0.029)</td>
</tr>
<tr>
<td>Fagaceae</td>
<td>229</td>
<td>122</td>
<td>0.928 (0.041)</td>
</tr>
<tr>
<td>Lauraceae</td>
<td>552</td>
<td>262</td>
<td>0.916 (0.044)</td>
</tr>
<tr>
<td>Moraceae</td>
<td>433</td>
<td>233</td>
<td>0.901 (0.054)</td>
</tr>
<tr>
<td>Myristicaceae</td>
<td>200</td>
<td>141</td>
<td>0.928 (0.046)</td>
</tr>
<tr>
<td>Sapindaceae</td>
<td>118</td>
<td>74</td>
<td>0.893 (0.071)</td>
</tr>
<tr>
<td>Total</td>
<td>3004</td>
<td>1720</td>
<td></td>
</tr>
</tbody>
</table>
botanical diversity (Table 2; gray). Notably, the growth habits herbs and miscellaneous have an inverse relationship to botanical diversity. This is also obvious when Fig. 2 is compared with Figs. S4 and S5.

After plotting the residuals of the best ‘forward–backward stepwise multiple regression’ model between legume diversity and overall botanical diversity (Fig. 3) it became clear that certain re-

Table 2

Regression coefficients of the different OLS models. ***p ≤ 0.001; **p ≤ 0.01; *p ≤ 0.05; ns — not significant.

| Coefficients                          | Estimate | Std. Error | t value | Pr(>|t|) | R^2 adj | Sign. |
|---------------------------------------|----------|------------|---------|----------|---------|-------|
| diversity – legumes                   |          |            |         |          |         |       |
| (Intercept)                           | 265.92854| 3.4191     | 77.78   | <2.2e-16 | 0.012   | ***   |
| legumes                               | 0.44766  | 0.03015    | 14.85   | <2e-16   |         | ***   |
| diversity – legumes.trees             |          |            |         |          |         |       |
| (Intercept)                           | 115.09337| 2.26141    | 50.9    | <2.2e-16 | 0.330   | ***   |
| legumes.trees                         | 5.62908  | 0.05963    | 94.4    | <2e-16   |         | ***   |
| diversity – legumes.lianas            |          |            |         |          |         |       |
| (Intercept)                           | -0.92767 | 1.943      | -0.477  | 0.633    | ns      |       |
| legumes.lianas                        | 10.56167 | 0.06177    | 170.979 | <2e-16   |         | ***   |
| diversity – legumes.herbs             |          |            |         |          |         |       |
| (Intercept)                           | 409.43386| 1.36427    | 300.11  | <2e-16   | 0.30    | ***   |
| legumes.herbs                         | -6.67024 | 0.05347    | -87.35  | <2e-16   |         | ***   |
| diversity – legumes.misc              |          |            |         |          |         |       |
| (Intercept)                           | 431.30357| 2.37608    | 181.52  | <2e-16   | 0.135   | ***   |
| legumes.misc                          | -5.07019 | 0.09523    | -53.24  | <2e-16   |         | ***   |
| diversity – legumes.trees + legumes.lianas + legumes.herbs + legumes.misc |          |            |         |          | 0.783   | ***   |
| (Intercept)                           | 120.04031| 1.80934    | 66.345  | <2e-16   |         | ***   |
| legumes.trees                         | 0.03846  | 0.06522    | 0.59    | 0.555    | ns      |       |
| legumes.lianas                        | 10.6359  | 0.08322    | 127.807 | <2e-16   |         | ***   |
| legumes.herbs                         | -0.35309 | 0.08556    | -4.127  | 3.70e-05 |         | ***   |
| legumes.misc                          | -5.09784 | 0.13902    | -36.67  | <2e-16   |         | ***   |
| diversity – legumes.lianas + legumes.herbs + legumes.misc |          |            |         |          |         |       |
| (Intercept)                           | 119.90008| 1.79361    | 66.848  | <2e-16   | 0.783   | ***   |
| legumes.lianas                        | 10.67156 | 0.05716    | 186.688 | <2e-16   |         | ***   |
| legumes.herbs                         | -0.36093 | 0.08065    | -4.587  | 4.50e-06 |         | ***   |
| legumes.misc                          | -5.06396 | 0.12659    | -40.004 | <2e-16   |         | ***   |

Fig. 2. Botanical diversity of Sundaland (inside polygon) based on 1297 stacked SDMs.
regions of Sundaland have higher (green) overall botanical diversity than predicted by the legumes, and others have lower (red) overall botanical diversity than predicted.

4. Discussion

The objective of the Global Legume Diversity Assessment is to use legume diversity as a proxy for overall angiosperm diversity at a global scale (Yahara et al., 2013). In the Neotropics and Africa legumes are the most dominant and diverse family lowland rainforest Angiosperm plant family. In Southeast Asian forests this role, at least in terms of dominance, is claimed by the Dipterocarpaceae (Gentry, 1988; ter Steege et al., 2006; Corlett and Primack, 2011). Besides this exception, pantropical composition of lowland forests is remarkably similar (Gentry, 1988). Our results show, however, that the legumes, in terms of number of species occurring on Sundaland, also are the most diverse family with 742 species. They are followed by the Lauraceae and Moraceae, and the Dipterocarpaceae only rank fourth in our dataset (Table 1). Similarly, in terms of number of species represented by at least five unique records the legumes rank on the first position.

When we compare the spatial patterns of legume diversity (Fig. 1) and botanical diversity (Fig. 2) it is obvious that the legumes are a poor predictor for overall botanical diversity. This is confirmed by the OLS model between legumes and botanical diversity, which indicated that legume diversity only explained 1.2% of the variance in overall botanical diversity (Table 2). The lack of power of the legumes to predict botanical diversity can, at least partly, be ascribed to the negative relation between botanical diversity and the herbaceous- and miscellaneous growth habits of the legumes (Table 2; Fig. S4 & S5). These two habits have their centre of diversity on east Java, whereas the centre of overall botanical diversity is located in northwest Borneo (Fig. 2; Raes et al., 2009). The best predictor of botanical diversity are the legume lianas, which pattern explained 61.8% of the variance in overall botanical diversity (Table 2); followed by the diversity pattern of trees, which explained 33%. However, in the multiple regression the contribution of trees is non-significant and therefore omitted in the forward–backward stepwise multiple regression. Both multiple regression models explained 78.3%, which makes the legumes a good predictor of botanical diversity after all.

The spatial pattern of the residuals of the relationship between legumes and overall botanical diversity (Fig. 3) indicated that the legumes under predict botanical diversity in the montane regions of Sundaland. These regions are mainly found in central Borneo, on the west side of the Malay Peninsula, the mountain ridge along the west coast of Sumatra, and at the volcanoes of Java (see Fig. S13 for altitude map). Regions that are over predicted by the legumes are found in the peat swamp — and kerangas forests of southern Borneo and Sumatra, the lowland regions of Java, sections of East Kalimantan and East Sabah on Borneo. The under prediction of botanical diversity in montane regions likely relates to the relative absence of legumes from higher altitudes, where the botanical diversity pattern included two montane plant families, knowingly the Ericaceae (Fig. S7) and Fagaceae (Fig. S8). Furthermore, both the Lauraceae (Fig. S9) and the Moraceae (Fig. S10) have a reasonable representation on the montane regions. It seems that legumes are generally not diverse in tropical montane forests worldwide (e.g., cloud forests), so the findings here apply generally.

The over prediction in the lowlands of Java can be related to the high diversity of herbaceous and ‘miscellaneous’ legumes on Java. Java is the densest populated region in South East Asia, and has consequently largely been deforested (Miettinen et al., 2012). The deforested areas were already converted to rice fields a long time ago (Junghuhn, 1850). The deforested and converted regions provide a perfect habitat for legume herbs and shrubs (miscellaneous). Furthermore, the legumes include a large number of herbaceous species, whereas the other seven families hardly include any herbs. The inclusion of families that represent herbaceous species to the analysis might result in a better prediction of botanical diversity by the legumes. The over prediction in East Kalimantan and east Sabah on Borneo relates to the strong representation of legume lianas and trees in those regions. These areas have slightly drier climatic conditions which are favoured by legumes in general (Schrire et al., 2005). The over prediction in the peat swamps and kerangas forests of southern Borneo and Sumatra may be because...
those forests are more stable, and thus species have larger seeds and more shade-tolerant seedlings (pers. obs. TY), conditions that favour legume liana and tree species (ter Steege et al., 2006).

Much of our analyses build on the reliability of SDMs. The advantage of the use of SDMs is that they predict the presence and absence of species, even for areas where no collections have been made, at an unprecedented high spatial resolution. At the same time SDMs build on the ecological niche theory (Colwell and Rangel, 2009; Soberón and Nakamura, 2009; Araújo and Peterson, 2012). We accounted for possible collection bias by using the target-background sample approach, suggested by Phillips et al. (2009). The target-background sample approach was enforced by the use of a mask layer, and the calibrated models were subsequently projected to the entire region, here the wider geographical extent (Fig. S13). When projecting SDMs in MaxEnt, a clamping file is automatically produced. A clamping file identifies regions with predictor conditions that are not represented in the training data, here the target-background sample. Visual inspection of a large number of MaxEnt models revealed negligible clamping, which indicated that the collection sites are not environmentally biased. For Borneo this finding was confirmed previously (Ras et al., 2009). To objectively test the accuracy of SDMs requires independent test results, which are hardly available. Among the few useful studies that qualify to compare our results on overall biodiversity patterns of Sundaland are the plot studies of Slik et al. (2009, 2011).

These studies confirm the higher predicted diversity on Borneo compared to the Malay Peninsula and Sumatra (Fig. 2). Furthermore, our botanical diversity map clearly delineates the less diverse peat swamps of southern Borneo and east Sumatra. Additionally, the east–west differentiation on Java is also reflected in the WWF ecoregions (Wikramanyake et al., 2002). Finally, if higher collection densities on northwest Borneo and the Malay Peninsula (Fig. S13) would have resulted in systematic bias, then both the botanical diversity as well as the legume diversity pattern would suffer from the same bias error. However, the predicted botanical diversity on Borneo is much higher than on the Malay Peninsula, while collection density on the Malay Peninsula is arguably higher than on northwest Borneo. We are therefore confident that both the legume diversity pattern and the overall botanical diversity pattern reflect reality. This leads to the conclusion that legumes as a whole are a poor predictor of overall botanical diversity on Sundaland. However, when species are differentiated to different growth habits a model can be fitted which quite accurately predicts overall botanical diversity. This implicates that for different biogeographic regions another model should be fitted to make the legumes a reliable predictor for overall botanical diversity.

Acknowledgements

We like to thank Serena Lee for contributing the Singapore Botanical Gardens collections database, and Matt Lavin and one anonymous referee for useful comments on a previous version. NR was supported by the Netherlands Research Council NWO-ALW with grant 819.01.014. TY was supported by the Environment Research and Technology Development of the Netherlands, Leiden University branch.


Ibáñez, E, Calcagno, Y, Hector, A, Connolly, J, Harpole, W.S, Reich, P.B, Scherer-Lorenzen, M, Schmidt, B, Tilman, D, Van Ruijven, J, Weigelt, A, Wisley, B.J,
Zavaleta, E.S., Loreau, M., 2011. High plant diversity is needed to maintain ecosyste-


Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predic-
tive maps produced by bioclimatic models. Ecological Applications 14, 401–413.

the role of fire derived from MODIS imagery. Global Change Biology 13, 2329–2340.

on Biodiversity and Ecosystem Services: moving a step closer to an IPCC-like mech-

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the per-

Lohman, D.J., de Bruyn, M., Page, T., von Rintelen, K., Hall, R., Ng, P.K.L., Shih, H.-T.,
Yahara, T., Javadi, F., Onoda, Y., Queiroz, L.P.d, Faith, D.P., Prado, D.E., Akasaka, M.,
sample size on the performance of species distribution models. Diversity and Dis-
tributions 14, 763–776.

Island Press, Washington, DC.

Newbold, T., 2010. Applications and limitations of museum data for conservation and
ecology, with particular attention to species distribution models. Progress in Physical
Geography 34, 3–22.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O’Hara, R.B., Simpson,

University Press, Princeton and Oxford.


2009. Sample selection bias and presence-only distribution models: implications for

R Development Core Team, 2012. R: A Language and Environment for Statistical Com-


Raes, N., ter Steege, H., 2007. A null-model for significance testing of presence-only

Raes, N., Van Welzen, P.C., 2009. The demarcation and internal division of Flora


C.O., Wilkie, P., Wulffraat, S., 2009. Environmental correlates for tropical tree diver-
sity and distribution patterns in Borneo. Diversity and Distributions 15, 523–532.

Slik, J.W.F., Alba, S.-t., Bastian, M., Brebny, F.Q., Cannon, C.H., Eichhorn, K.A.O., Fredriksson,
C., Kartawinata, K., Laumonier, Y., Mansor, A., Marjokori, A., Meijaard, E., Morley, R.J.,
Nagamasu, H., Nils, R., Nurjihaya, E., Payne, J., Permana, A., Poulsen, A.D., Raes, N.,
Sunda Shelf shaped biogeographic patterns in the equatorial forests of Southeast

Soberón, J., Nakamura, M., 2009. Niches and distributional areas: concepts, methods,

Sodhi, N.S., Lee, T.M., Koh, L.P., Brook, B.W., 2009. A meta-analysis of the impact of an-
thropogenic forest disturbance on Southeast Asian forests. Biotropica 41, 103–109.

TEEB, 2009. The Economics of Ecosystems and Biodiversity for National and Interna-

TEEB, 2010. The Economics of Ecosystems and Biodiversity: Mainstreaming the Eco-
nomics of Nature: A Synthesis of the Approach, Conclusions and Recommendations
of TEEB. 39.

Continental-scale patterns of canopy tree composition and function across Amazonia.


Tropicalis.org, Missouri Botanical Garden http://www.tropicalis.org (accessed March 15,
2013).

Van Steenis, C.G.G.J., 1950. The delimitation of Malaysia and its main plant geographical
divisions. In: Van Steenis, C.G.G.J. (Ed.), Flora Malesiana series 1, vol. 1. Noordhoff-
Kolff N.V., Djakarta, pp. LXX–LXXX.

Van Welzen, P.C., Silk, J.W.F., 2009. Patterns in species richness and composition of


Wenger, S.J., Olden, J.D., 2012. Assessing transferability of ecological models: an under-
appreciated aspect of statistical validation. Methods in Ecology and Evolution 3,
260–267.

Wakamanyake, E., Dinerstein, E., Loucks, C.J., Olson, D.M., Morrison, J., Lamoureux, J.,
Press, Washington, D.C.

sample size on the performance of species distribution models. Diversity and Dis-
tributions 14, 763–773.

Woodruff, D., 2010. Biogeography and conservation in Southeast Asia: how 2.7 million
years of repeated environmental fluctuations affect today’s patterns and the future of
the remaining refugial-phase biodiversity. Biodiversity and Conservation 19, 919–941.

Strategies to observe and assess changes of terrestrial biodiversity in the Asia-
Pacific regions. In: Nakano, S.-i, Yahara, T., Nakashizuka, T. (Eds.), The Biodivi-
sity Observation Network in the Asia-Pacific Region. Springer, Japan, pp. 3–19.

Yahara, T., Javadi, F., Onoda, Y., Queiroz, L.P.d, Faith, D.P., Prado, D.E., Aksaka, M.,
Kadoya, T., Ishihama, F., Davies, S., Silk, J.W.F., Yi, T., Ma, K., Bin, C., Darnai, D.,
Pennington, R.T., Tada, M., Shimada, M., Ito, M., Egan, A.N., Uecki, R., Raes, N.,
Kajita, T., Mimura, M., Tachida, H., Iwasa, Y., Smith, C.F., Victor, J.E., Nikonov, T.,
Taxon 62, 249–266.

species distribution modeling to improve conservation and land use planning of