Borneo : a quantitative analysis of botanical richness, endemicity and floristic regions based on herbarium records

Raes, N.

Citation

CHAPTER 4

Botanical richness and endemicity patterns of Borneo derived from species distribution models

⟿ Niels Raes, Marco C. Roos, J.W. Ferry Slik,
E. Emiel van Loon, and Hans ter Steege

Abstract

This study provides a Borneo-wide, quantitative assessment of botanical richness and endemicity at a high spatial resolution, and based on actual collection data. To overcome the bias in collection effort, and to be able to predict the presence and absence of species, even for areas where no collections have been made, we constructed species distribution models (SDMs) for all species taxonomically revised in Flora Malesiana. Species richness and endemicity maps were based on 1439 significant SDMs. Mapping of the residuals of the richness-endemicity relationship identified areas with higher levels of endemicity than can be expected on the basis of species richness, the endemicity hotspots. We were able to identify one previously unknown region of high diversity, the high mountain peaks of East Kalimantan; and two additional endemicity hotspots, the Müller Mountains and the Sangkulirang peninsula. The areas of high diversity and endemicity were characterized by a relatively small range in annual temperature, but with seasonality in temperatures within that range. Furthermore, these areas were least affected by El Niño Southern Oscillation drought events. The endemicity hotspots were found in areas, which were ecologically distinct in altitude, edaphic conditions, annual precipitation, or a combination of these factors. These results can be used to guide conservation efforts of the highly threatened forests of Borneo.

Introduction

Borneo, the third largest island of the world, is the botanically most diverse part of the Sundaland hotspot, one of the world's 25 biodiversity hotspots (Myers *et al.* 2000).

Southeast Asia as a whole faces an estimated loss of three quarters of its original forest area by 2100, and up to 42% of its biodiversity (Sodhi *et al.* 2004). For Borneo, currently only 57% of its land surface remains forested, and annual deforestation averages 1.7% (FAO 2006, Langner *et al.* 2007). Even more worrying is the fact that 56% of Kalimantan's (Indonesian Borneo) protected lowland forests has been lost between 1985 and 2001 (Curran *et al.* 2004, Stibig *et al.* 2007). Ca. 37% of Borneo's 15,000 vascular plant species (Roos *et al.* 2004) are thought to be endemic (van Welzen *et al.* 2005), with an estimated number of 10,000 species occurring in the WWF Borneo lowland rain forests ecoregion alone (Wikramanayake *et al.* 2002, Kier *et al.* 2005).

Considering the exceptional richness and concentration of endemic, or narrow ranged, species on Borneo, surprisingly little is known about the spatial distribution of both components. Only in 1995 the WWF and IUCN (1995) introduced the 'Centres of plant diversity' (CPD) for Australasia. In this contribution they argued that on Borneo most endemic plant species can be found in smaller areas in the north, the central mountain chain, and in the south-eastern Meratus Mountains (Fig. 4.1). A view largely supported by MacKinnon *et al.* (1996). Wong (1998) added to this list the 'Riau Pocket', which consists of two areas. One of these is similar to the north-western Sarawak biogeographical unit of MacKinnon *et al.* (1996), the other is the most western tip of Borneo (Fig. 4.1). Wong (1998) further suggests that Mount Kinabalu is a hotspot of plant diversity (Fig. 4.1), which is confirmed by its ca. 5,000 documented vascular plant species (Beaman 2005, Grytnes and Beaman 2006). Furthermore, Wong (1998) reports a comparatively lower diversity in the remaining area of Borneo, mainly consisting of the Kalimantan provinces.

These findings are partly confirmed by the only quantitative Borneo-wide study of lowland
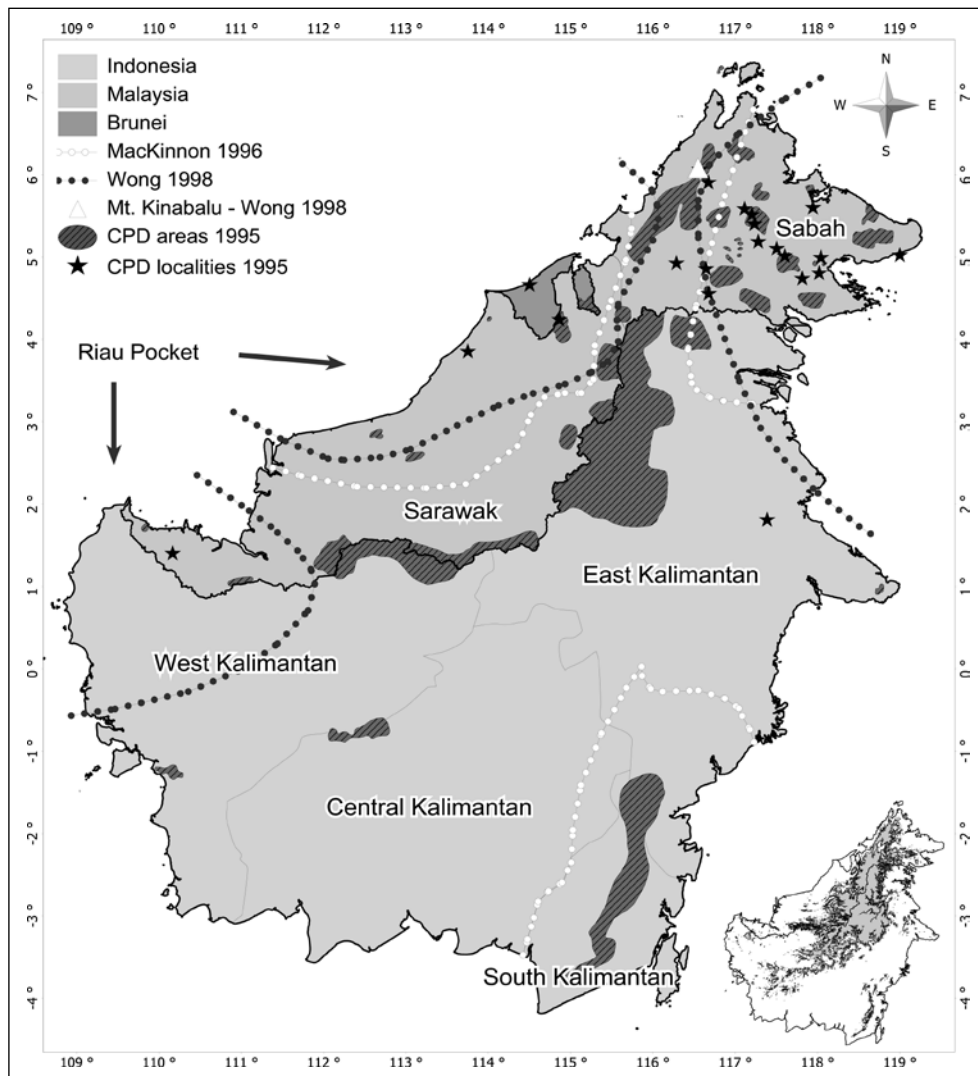
Figure 4.1. The country boundaries, the Indonesian and Malaysian provinces, the location of the centres of plant diversity (CPD) (WWF & IUCN, 1995), and the areas of high plant richness of Borneo by MacKinnon et al. (1996), and by Wong (1998). Inset shows all areas with an altitude of >500m asl.

dipterocarp forest (Slik *et al.* 2003). Based on data of 28 plots, at genus level, and for trees with a diameter of ≥ 10 cm, Slik *et al.*'s (2003) results only confirmed the biodiversity hotspots of the south-eastern Meratus Mountains and the north-western Sarawak biogeographical

units. Their analysis did not support the comparatively lower diversity in the Kalimantan provinces of Wong (1998), however.
Except for the flora of Mount Kinabalu, and the plot studies by Slik *et al.* (2003), all the identified areas of high species richness and

endemicity have boundaries that are largely based on informal expert opinions. Considering the rapid loss, and the lack of knowledge on the Borneo-wide spatial distribution of species richness and endemicity, warrants further studies.

With the ongoing digitization of natural history museum collections and herbarium specimens (Graham *et al.* 2004), much data has become available that allows for a quantitative analysis of species richness and endemicity patterns. Despite all collecting efforts of the last centuries, however, there is no complete inventory of all organisms inhabiting any single locality (Hortal *et al.* 2004). Moreover, collection localities are often biased towards easily accessible areas (Reddy and Davalos 2003, Kadmon *et al.* 2004, Hortal *et al.* 2007). Hence, the collecting effort on Borneo (and elsewhere) has been biased, as is evident from the spatial distribution of Bornean collection records (Raes and ter Steege 2007; Chapter 3, Fig.3.3). To overcome incomplete and biased sampling, and to be able to predict the geographic distributions of species even for areas where no collections have been made, has contributed to the development of species distribution modelling techniques (Araújo and Guisan 2006, Elith *et al.* 2006, Peterson 2006). A species distribution model (SDM) predicts the potential distribution of a species by interpolating identified relationships between presence/ absence, or presence-only data of a species on one hand, and environmental predictors on the other hand, across an area of interest (Elith *et al.* 2006). Only few studies have attempted to develop species richness and endemicity patterns derived from SDMs (Zaniewski *et al.* 2002, Schmidt *et al.* 2005, Küper *et al.* 2006, Costa *et al.* 2007). Although these studies clearly contribute to the identification of possible centres of high richness, endemicity, and data deficiency; none of the underlying SDMs were statistically tested.

The importance of assessing whether an SDM differs from what would be expected on the basis of chance alone was first recognized by Olden *et al.* (2002). Recently, Raes and ter Steege (2007; Chapter 3) developed a methodology to test the significance of SDMs developed with presence-only data, the principal data of herbaria and natural history museums (Graham *et al.* 2004). Their methodology tests whether the correlations found between species' presence localities and the environmental predictors deviate from random chance expectation with a null-model (Raes and ter Steege 2007; Chapter 3).
To contribute to the conservation of botanical diversity of Borneo, we set out to model the patterns of botanical richness and endemicity, based on all significant SDMs at 5 arc-minute (~10x10 km at the equator) spatial resolution for all species treated in Flora Malesiana (Anon. 1959-2007) occurring on Borneo. Then, based on these patterns, we identify areas with higher levels of endemicity than can be expected on the basis of species richness. Finally, we analyse which environmental factors best explain the botanical richness and endemicity patterns.

Methods

Species data

We extracted all georeferenced species records from Borneo belonging to families treated in Flora Malesiana (Anon. 1959-2007) from the BRAHMS database of the National Herbarium of the Netherlands. To this dataset we added the georeferenced records of revised genera of the Annonaceae, Euphorbiaceae, and Orchidaceae. In total this dataset comprised 66,262 georeferenced records belonging to 102 plant families representing 4674 species. From this set of georeferenced records, we recorded

species presences for each 5 arc-minute grid cell, avoiding duplicate species records in one grid cell. We used a 5 arc-minute spatial resolution because this is the available resolution of the FAO soil property predictors (see below), and because georeferencing at a higher spatial resolution is not realistic. Furthermore, a species had to be represented in at least five grid cells to be modelled. These requirements were met for 2273 species represented by 44,106 unique records, ranging from 5 to 202 unique records per species.

Environmental predictors
To model the species distributions, we initially selected 37 environmental predictors. We downloaded the digital elevation model (DEM (in m)) and the 19 bioclimatic predictors (1950-2000) of the WORLDCLIM dataset (<http://www.worldclim.org>) for Borneo at 5 arc-minute resolution (Hijmans et al. 2005). Additionally, we selected 15 soil predictors from the FAO database for poverty and insecurity mapping (FAO 2002), shown in table 4.1. To this dataset

we added a layer with the elevation range, defined as the difference between the lowest and highest altitude within a 5 arc-minute grid cell based on the 90m SRTM altitude data (<http://srtm.csi.cgiar.org>). Finally, a data-layer, reflecting the El Niño Southern Oscillation (ENSO) event drought impact, was added. ENSO drought impact was defined as the relative average annual difference in 'normalized difference vegetation index' (NDVI) values (<ftp://ftp.glcf.umiacs.umd.edu/glcf/GIMMS/Geographic/>) between months of a severe ENSO (07/1982-06/1983), and a non-ENSO year (07/1981-06/1982). These NDVI data were the oldest data available, and are therefore least affected by deforestation and land use change. We retained only grid cells with data for all data-layers, resulting in 8577 grid cells for Borneo. Records on the coast-line just falling outside the grid cells due to the 5 arc-minute resolution were shifted to their closest grid cell. Data-layer manipulations were performed with Manifold GIS (Manifold Net Ltd).

Table 4.1. Pearson's correlation coefficients (r) for the 15 FAO soil predictors and the five PCA soil axes. Values in italic r>0.4 or r<-0.4 (modest correlation) and in bold italic r>0.7 or r<-0.7 (strong correlation). Correlations are significant at 0.05 level (*), 0.01 Level (**), at 0.001 level (***), or not significant at 0.05 level (ns) (2-tailed). CEC = Cation Exchange Capacity; C:N = Carbon:Nitrogen.

	PCA01	PCA02	PCA03	PCA04	PCA05
Base saturation % topsoil	-0.1013 ns	*-0.8429* ***	0.0974 ns	0.2864 ns	0.2268 ns
CEC clay topsoil	*0.5712* ***	-0.3342 *	*-0.5284* ***	0.4161 *	-0.1366 ns
CEC soil topsoil	*0.7449* ***	-0.1708 ns	0.2366 ns	0.0295 ns	0.0806 ns
C:N-ratio class topsoil	*0.5083* ***	0.3100 ns	*-0.4183* **	-0.2982 ns	*0.5314* ***
Easy available water	*-0.7886* ***	*0.4747* **	0.0332 ns	-0.1926 ns	0.1242 ns
Effective soil depth	0.2428 ns	-0.3322 *	-0.1498 ns	*-0.8224* ***	-0.2733 ns
Nitrogen % topsoil	*0.7360* ***	0.2317 ns	0.3555 *	0.1245 ns	-0.1529 ns
Organic carbon % topsoil	*0.5523* ***	*0.5221* ***	0.3205 *	0.0227 ns	-0.2646 ns
Organic carbon pool	*0.7626* ***	0.3883 *	0.2412 ns	-0.1427 ns	0.1172 ns
pH topsoil	*-0.4389* **	*-0.6870* ***	0.1403 ns	0.2410 ns	-0.0953 ns
Soil drainage class	*0.8323* ***	-0.2111 ns	0.1628 ns	0.1071 ns	0.2241 ns
Soil moisture storage capacity	*-0.7108* ***	*0.5545* ***	-0.0116 ns	-0.0222 ns	0.1731 ns
Soil production Index	-0.0489 ns	*-0.8584* ***	0.0836 ns	-0.1758 ns	-0.2444 ns
Textural class subsoil	-0.2891 ns	-0.0161 ns	*0.8747* ***	-0.1090 ns	0.1550 ns
Textural class topsoil	-0.0382 ns	-0.1762 ns	*0.9153* ***	-0.1784 ns	0.1307 ns

To avoid problems such as multi-collinearity (Graham 2003), which can result in model over-fitting (Peterson et al. 2007), we reduced the number of environmental predictors. From the bioclimatic predictors, together with the DEM, elevation range, and ENSO, we selected only those predictors which were least correlated (highest Pearson's r = 0.737; Table S4.1): DEM (correlated with and proxy for elevation range, and strongly negatively correlated with the mean annual temperature, maximum temperature warmest month and -quarter, minimum temperature coldest month and -quarter, mean temperature wettest- and driest quarter); Bio04 – Temperature seasonality; Bio07– Temperature annual range (correlated with diurnal temperature range); Bio12 – Annual precipitation (correlated with and proxy for precipitation in the wettest month and -quarter, driest month and - quarter, warmest - and coldest quarter); Bio15 – Precipitation seasonality; and ENSO (table S4.1). To reduce the number of soil predictors we used a principal component analysis (PCA). We performed the PCA on the 41 unique combinations of the 15 soil predictors values observed for the 8577 grid cells of Borneo. We selected the first five PCA-axes as our soil property predictors (PCA01-05), which together explained 83 % of the variance in the soil data. Pearson's correlation was used to determine which of the original 15 FAO soil predictors were significantly correlated to each of the five PCA axes (Table 4.1). This resulted in a reduction from 37 to 11 uncorrelated predictors, which were used to model the species distributions (Table S4.1; Fig. S4.1).

Species distribution model (SDM) building and testing with a bias corrected null-model
To model the species distributions we selected the modelling application Maxent (ver. 3.0.4; <http://www.cs.princeton.edu/~shapire/maxent/>) (Phillips et al. 2006). Maxent was

specifically developed to model species distributions with presence-only data, has shown to outperform most other modelling applications (Elith et al. 2006, Pearson et al. 2007), is least affected by location errors in occurrences (Graham et al. 2008), and best performs when few presence records are available (Wisz et al. 2008). Maxent was set to use all species presence records for model building (explained below), by setting the 'random test percentage' to zero. The modelling rules were set to use linear features, when less than 10 records were available, adding quadratic features for SDMs developed with 10-14 records, and including hinge features for species with 15, or more, records (Raes and ter Steege 2007; Chapter 3). For each of the 2273 species an SDM was developed based on its unique presence records and the 11 environmental predictors.
As measure of SDM accuracy we used the threshold independent and prevalence insensitive area under the curve (AUC) of the receiver operating characteristic (ROC) plot (Fielding and Bell 1997, McPherson et al. 2004, Raes and ter Steege 2007; Chapter 3), produced by Maxent. All measures of SDM accuracy require absences (Fielding and Bell 1997). When these are lacking, as is the case here, they are replaced by pseudo-absences or sites randomly selected at localities where no species presence was recorded (Ferrier et al. 2002, Phillips et al. 2006). However, when SDM accuracy measures are based on presence-only data and pseudo-absences, the standard measures of accuracy (e.g. the often used measure AUC>0.7) do not apply (Raes and ter Steege 2007; Chapter 3). Therefore, we used the method presented in Raes and ter Steege (2007; Chapter 3) to test the AUC value of an SDM developed with all presence records against a bias corrected null-model of AUC values expected by chance. The AUC value of an SDM developed with *n* records is tested against

the upper 95% one-sided confidence interval (C.I.) AUC value derived from the AUC values of 1000 x *n* randomly drawn and modelled points. The random points were drawn from cells where in the past collections were made, and hence were corrected for any geographical sampling bias. For Borneo this was the case for 1837 (21.4%) of the total of 8577 grid cells (Raes and ter Steege 2007; Chapter 3).

We developed null-distributions for 5-35 records (31 distributions), for 40-50 records with intervals of 5 records (3 distributions), for 60-100 records with intervals of 10 records (5 distributions), and from 150-250 with intervals of 50 records (3 distributions). For each of these distributions we assessed the upper 95% one-sided C.I. AUC value, by ranking the 1000 AUC values and selecting the 950th value. We developed three series of C.I. values dependent on the modelling rules used by Maxent; 5-9 (only linear), 10-14 (linear & quadratic), and ≥ 15 (linear, quadratic & hinge) records. We applied a curve-fit to each of the three series against which the AUC values for all 2273 SDMs were tested. For further analyses only the significant SDMs were retained. To assess whether species represented by few records were not proportionally more often rejected than species with many records, we plotted the relative species abundance values against the relative species ranks. Similar shaped curves indicate that the sample is representative. Additionally, we tested whether the 1837 collection localities were biased in environmental predictor space. For each of the 11 predictors we divided predictor space into 10 equal-interval bins based on the ranges observed for Borneo (8577 grid cells) (Loiselle et al. 2008). Then we tested whether the frequency distributions represented by the 1837 collection localities differed from all 8577 grid cells using a Chi-square test.

Botanical richness and endemicity patterns

In order to develop patterns of botanical richness and endemicity, a threshold was set to convert the continuous Maxent SDM predictions, which range from 0 to 100, to discrete presence/absence values. Although species identifications, and georeferencing of the collection localities, were carried out with the greatest possible accuracy, we found it reasonable to assume that 10% of the records were either wrongly identified, or georeferenced. Therefore, for all SDMs represented by ≥ 10 records we used the fixed '10 percentile presence' threshold. This threshold uses the Maxent value of the 10 percentile species presence record to define all areas with a lower predicted Maxent value as absent, and with a higher value as present. For those species represented by 5-9 records we used either the 'sensitivity-specificity equality' or the 'sum maximization' threshold (Liu *et al.* 2005), dependent on which of the two corresponding omission rate values was closest to 10%.

Once the thresholds were set, the botanical richness pattern was developed by superimposing all significant SDMs. To develop the endemicity pattern we used the weighted endemism index (Crisp *et al.* 2001, Kier and Barthlott 2001, Küper *et al.* 2006, Slatyer *et al.* 2007). This index weighs species richness according to the range sizes of the species present, and is calculated by summing the inverse of the range sizes of the species present in each grid cell. A species with a range of 10 grid cells has a weight of 1/10 in every grid cell where it is present. We developed the endemicity pattern by summing the weights of all significant SDMs for all grid cells.

Data analyses

We first assessed whether the modelled species richness did not under-predict the actual number of species collected in corresponding cells. This was done by plotting the predicted number of species against the collected number of species for all 8577 grid cells. To plot the data at log-log scale, 'the collected number of species' was transformed to the logarithm of 'the collected number of species + 1'.

The hotspots of endemic species were identified by mapping the relative residuals of the species richness – weighted endemism relationship. The relationship was assessed with a curve-fit procedure, evaluating several polynomial functions. The residuals of this relation were divided by their predicted weighted endemism values, resulting in a measure of relative residual weighted endemism.

To study which environmental predictors best explained species richness, weighted endemism, and their relative residuals we used a technique known as variation partitioning (Legendre 2008). Variation partitioning is a technique that partitions the variation of a response variable between two sets of explanatory variables, here a set of environmental predictors and a set of spatial predictors. As spatial predictors we used the nine terms of the third order polynomial trend-surface regression equation of latitude and longitude (Borcard et al. 1992, Lobo and Martin-Piera 2002, Legendre 2008). To account for possible non-linear effects between the diversity patterns and the original set of 11 environmental predictors, we added the quadratic terms of those predictors uncorrelated with the original predictors ($r<0.7$). This was the case for PCA02, PCA04 and PCA05 (Table S4.1). The variation was partitioned by performing a forward-backward stepwise multiple regression analysis for the three diversity measures and (1) the environmental predictors, (2) the spatial predictors, and (3) a combined matrix of both predictor sets. This allowed us to assess which proportion of the variation was attributed to only environmental predictors (a), to spatially structured environmental predictors (b), to spatial predictors (c), and to unexplained (residual) variation (d) (See Legendre 2008 for methodological details).

Multiple regression analyses applied to macroecological data which are driven by structured biological processes, may result in residual spatial autocorrelation (RSA). The presence of RSA is a violation of the assumption that residuals should be independent and identically distributed, and results in inflated Type I errors (Dormann et al. 2007). It was recently shown, however, that short-distance RSA, while causing inflated type I errors, does not seriously affect the interpretation of the regression coefficients estimated by ordinary least squares regressions (Diniz-Filho *et al.* 2007, Hawkins *et al.* 2007). To establish whether RSA was present in our regression residuals we assessed the Moran's I values with SAM – Spatial Analysis in Macroecology <http://www.ecoevol.ufg.br/sam> (Rangel *et al.* 2006). Since only a small proportion of the variation was explained by the spatial predictors alone (see Results), we estimated the regression coefficients for the three biodiversity measures based on the significant environmental predictors only. Model performances were tested with a 10-fold cross-validation procedure, using 90% to train the model with the significant environmental predictors selected by the full-model, and test the model with the remaining 10%, repeated 10 times. All regression analyses were performed using SPSS version 15 (SPSS Inc.).

It can be argued that this approach has a certain degree of circularity in reasoning, since the same predictors were used to develop the underlying SDMs. We argue, however, that Maxent identifies correlations with the environmental predictors independently for each species. Even if for two species the same
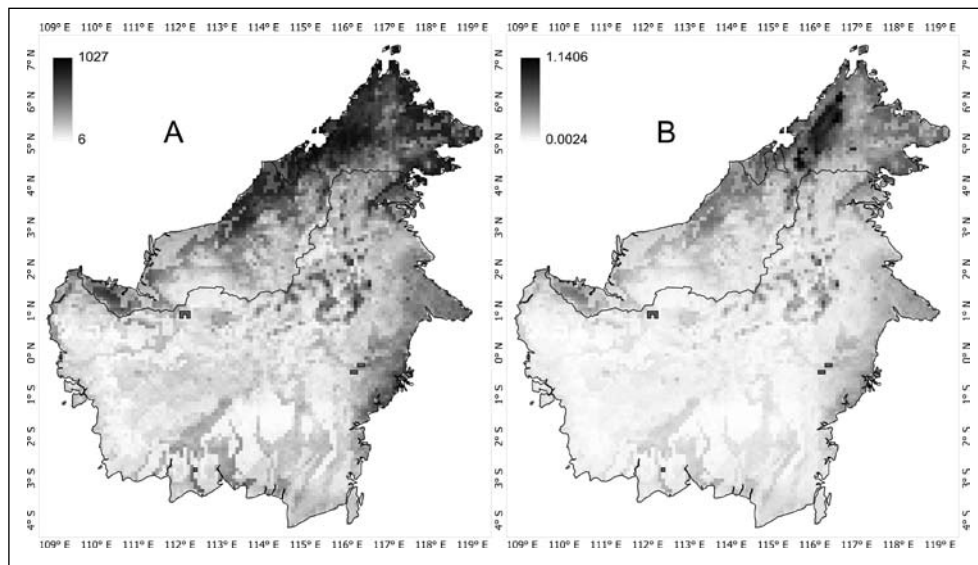
Figure 4.2. Botanical richness (A) and weighted endemism (B) pattern derived from 1439 superimposed significant species distribution models.

Figure 4.3. The predicted number of species plotted against the number of species actually collected per grid cell based on species which had a significant species distribution model. The two grid cells for which a lower number of species was predicted than actually was found (below the diagonal line) are geographically plotted on the inset. Note the log scale of both axes.

predictors are selected as being the most important to predict their distributions, these two species can have another optimum in their response. So we posit that the significant SDMs underlying the superimposed species richness, weighted endemism, and relative residual weighted endemism patterns are essentially independent.

Results

From the 2273 species which were modelled, 1439 (63.3 %) had a distribution pattern that differed significantly from a random one. The frequency distributions of the environmental conditions represented by the collection localities did not significantly differ from the distributions for all 8577 grid cells of Borneo (Fig. S4.2). The relative species rank abundance curves (Fig. S4.3) for the records of all 2273 species, and for the records of the 1439 significant species, were largely similar

in shape. This indicated that in terms of the frequency distributions of records, the 1439 species represented by a significant SDM are a representative sample of the total of 2273 modelled species.

Superimposing the significant SDMs, and derived weighted endemism values, resulted in the botanical richness and weighted endemism maps as presented in Fig. 4.2a and b, respectively. The highest richness was found in Sabah, where 1027 species of the total of 1439 species were predicted to occur in one grid cell. The lowest predicted richness was found in West Kalimantan where only 6 of the modelled species were predicted to be present. Besides Sabah and north-western Sarawak, also the high mountains in East Kalimantan had very high predicted richness values (Fig. 4.2a). The weighted endemism values (Fig. 4.2b) showed largely the same pattern as the richness pattern, albeit more concentrated around the mountains, notably the northern Crocker Range, and not extending into the lowland. The log-log plot of the predicted number of
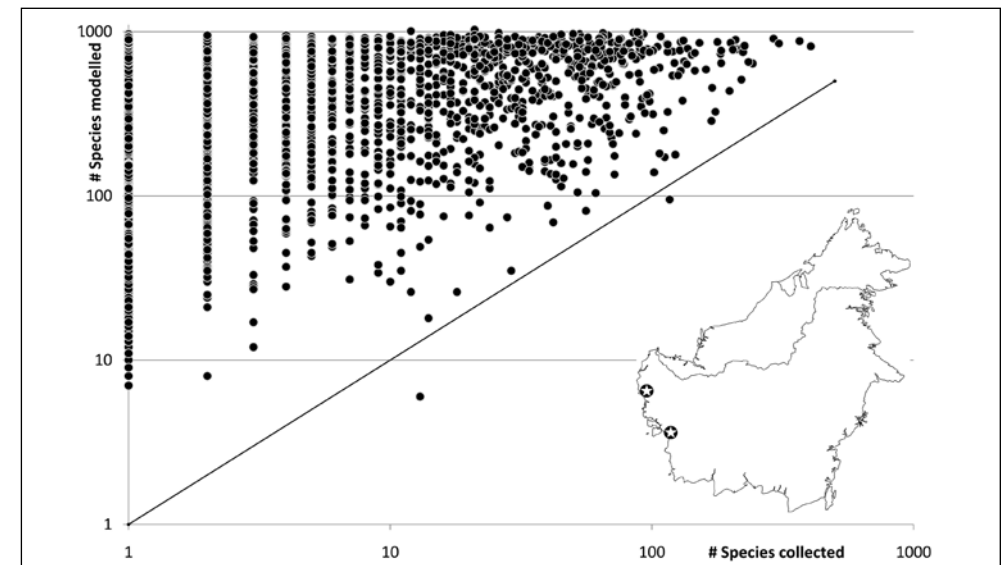
species plotted against the species counts per grid cell, showed that for cells where few species were actually collected, almost the whole range of predicted numbers of species was found (Fig. 4.3). For cells with highest numbers of actually collected species, however, the predicted number of species was never lower. Only for 2 grid cells (0.13%) a lower number of species was predicted than were actually collected (Fig. 4.3; below diagonal and inset).

The relationship between species richness and weighted endemism was best described by a quadratic function ($R^2 = 0.914$; $p<0.001$; Fig. 4.4 bottom). Mapping the relative residuals of this relationship, exposed those areas with higher weighted endemism values than can be expected on the basis of species richness (Fig. 4.4 top; black and white areas). These areas were located on Mount Kinabalu and the Crocker Range Mountains in the north, the Müller Mountains in the south of the

central mountain chain, on the east-side of the southern Meratus Mountains, and in the north of East Kalimantan's Sangkulirang peninsula. The results of the variation partitioning showed that the majority of the variation in species richness and weighted endemism can be explained by spatially structured environmental variables (Fig. 4.5; fraction b). Only very small fractions (3.4-4.5%) were attributed to spatial variables alone (Fig 4.5; fraction c). Hence, the regressions were performed on the environmental predictors only. The forward-backward stepwise regression results for both species richness and weighted endemism suggest that the most important environmental variables were Bio07 – temperature annual range, Bio04 – temperature seasonality, and ENSO (Table 4.2). Together all significant environmental variables explained 82.8, and 75.3 percent of the total variance in species richness and weighted endemism, respectively (Table 4.2; Fig. 4.5, fraction a+b). For the

Table 4.2. Results of the forward-backward stepwise multiple regression for species richness, weighted endemism, and the relative residual weighted endemism values as a function of the environmental predictors (n=8577). The order indicates the sequence in which the significant environmental predictors (Pred.; p<0.05) were included in the regression equation. $R^2_{adj.}$ indicates the cumulative adjusted coefficients of determination after the inclusion of each variable. Beta is the standardized regression coefficient, and t is the corresponding t-value for the full regression model. All models are significant at p<0.001. Legend: DEM – Digital Elevation Model; Bio04 – Temperature seasonality; Bio07 – Temperature annual range; Bio12 – Annual precipitation; Bio15 – Precipitation seasonality; PCA01-05 – PCA soil axes 1-5; ENSO – Drought impact by El Nino- events on NDVI values.

Species Richness				Weighted Endemism				Residual Weighted Endemism			
Pred.	R^2_{adj}	Beta	t	Pred.	$R^2_{adj.}$	Beta	t	Pred.	$R^2_{adj.}$	Beta	t
Bio07	0.123	-0.931	-132.034	Bio07	0.107	-0.840	-96.536	DEM	0.236	0.647	65.472
Bio04	0.546	1.026	134.268	Bio04	0.456	0.975	104.508	PCA05²	0.352	0.311	38.693
ENSO	0.640	0.258	49.341	ENSO	0.547	0.236	36.006	PCA03	0.409	-0.182	-23.512
PCA05²	0.695	-0.211	-41.642	DEM	0.599	0.345	45.455	Bio12	0.460	-0.241	-28.976
PCA03	0.745	0.223	42.251	PCA05²	0.635	-0.171	-27.745	PCA04	0.486	-0.219	-27.414
PCA05	0.774	0.225	47.701	Bio12	0.674	-0.240	-23.337	Bio15	0.518	0.195	20.319
PCA02²	0.787	-0.151	-28.427	PCA05	0.700	0.194	33.973	PCA01	0.535	0.155	19.263
Bio12	0.807	-0.157	-30.488	PCA03	0.721	0.158	24.581	ENSO	0.544	0.104	12.026
DEM	0.822	0.171	27.076	PCA02²	0.747	-0.156	-23.337	PCA05	0.551	-0.105	-13.893
PCA02	0.826	-0.044	-7.674	PCA02	0.750	-0.037	-5.420	PCA02²	0.557	0.087	9.890
PCA04²	0.827	-0.053	-8.578	PCA04	0.751	-0.022	-3.350	Bio07	0.561	0.149	13.011
PCA01	0.828	-0.036	-6.854	PCA01	0.751	-0.036	-5.696	Bio04	0.565	-0.105	-8.528
				PCA04²	0.752	-0.042	-5.163	PCA02	0.566	-0.040	-4.917
				Bio15	0.753	0.028	3.800				

Figure 4.4. The relationship between botanical richness and weighted endemism. Light grey dots represent grid cells with negative relative residual weighted endemism values; dark grey dots positive relative residual endemism values between 0-50%, white dots between 50-100%, and black dots >100%. Residual classes are mapped in the top image.

residuals, DEM (altitude), the quadratic term of PCA soil axis 5, and PCA soil axis 3 were the most important environmental variables. In total the model explained 56.6% of the variance in relative residual endemism (Table 4.2). The 10-fold cross-validation results of the models obtained through stepwise regression showed that the average $R^2_{adj.}$ values of the test-data partitions were not significantly different from the average training-data partition values (Fig. 4.6), implying that the models were not over-parameterized. The Moran´s I values of the regression residuals of species richness and weighted endemism indicated that some RSA was still present for the first three lags (Fig. 4.7).

Discussion

Botanical richness pattern

The richness pattern is based on 1439 significant SDMs, approximately 10% of the estimated number of 15,000 species expected to occur on Borneo. This is the largest dataset available today and represents all life-forms from herbs to canopy trees represented by 102 plant families treated in Flora Malesiana. The relative species rank abundance curve of the species represented by the 1439 significant SDMs lies slightly above the curve for all 2273 species that were modelled (Fig. S4.3). This indicates that a smaller proportion of species represented by few records is represented in the diversity pattern. It was shown, however,
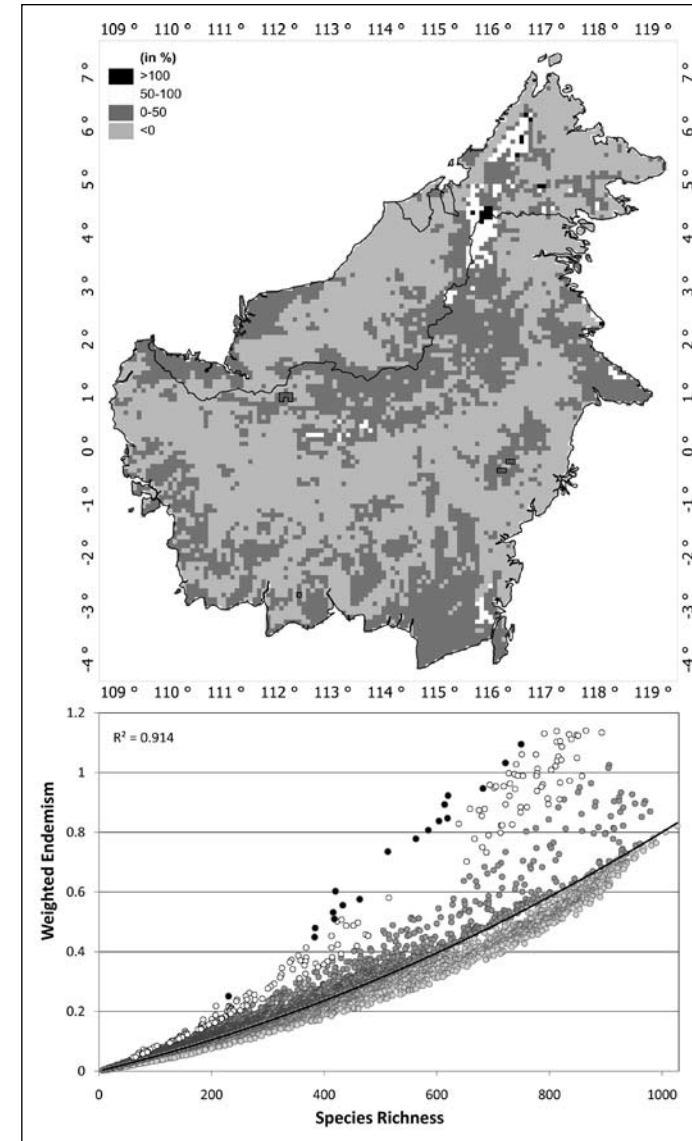
that commoner species are most responsible for richness patterns (Lennon *et al.* 2004). We argue therefore that this is not problematic for the interpretation of the botanical richness pattern. The implications for the endemicity patterns are discussed below.

The highest botanical richness is predicted

to occur in northern, and north-western Borneo (Fig. 4.2a). These results confirm the suggestion that northern Borneo is a centre of vascular plant diversity (Barthlott *et al.* 2005, Kier *et al.* 2005, Mutke and Barthlott 2005). In the lowlands the highest species richness was found in the areas west of the northern Crocker

Mountains range, while the rest of the lowlands had much lower predicted richness values. This distribution was previously suggested by Wong (1998). Furthermore, we expect high richness values for the transition area between the Sembakung-Sesajap delta and the lowland forest of northern East Kalimantan, and for the most western tip of Sarawak, around Kuching. All these areas fall within the roughly demarcated areas of high richness by Mackinnon et al. (1996) and Wong (1998) (Fig. 4.1). Additionally, our modelled richness pattern identified the eastern mountain peaks of the central mountain chain located in central East Kalimantan (Fig. 4.2a) as an area of potentially high richness, previously not recognized. Our results cannot support the south-eastern Meratus Mountains as an area of exceptional plant species richness as was indicated by MacKinnon et al. (1996).

The lowest plant species richness values were found in the southern, and south-western parts of Borneo (Fig. 4.2a). These areas largely coincide with the WWF Sundaland heath- and peat swamp forest ecoregions (Wikramanayake et al. 2002). According to Wikramanayake et al. (2002) these forest types are generally less species-rich than comparable lowland dipterocarp forest. Similarly, the peat swamp forests along the coast of Sarawak, and part of Brunei, also have lower predicted species richness values than the surrounding lowland forest. Nevertheless, we expect that the predicted richness for the lowland regions of the Kalimantan provinces still underestimates the diversity due to relatively low collection densities in this region (Raes and ter Steege 2007; Chapter 3, Fig.3.3). For a species to be modelled it should be represented by at least 5 unique records in our data set. The probability that Kalimantan endemics are represented by 5 records is much lower than for north-western Borneo endemics. On the other hand, the 10 equal-interval bins frequency distributions of

the environmental predictors for the collection localities did not differ from those of all 8577 grid cells (Fig. S4.2). A lower predicted number of species caused by underrepresentation of environmental conditions represented by the collection localities is therefore not expected. At the same time South-, East-, and Central Kalimantan represent regions most affected by the ENSO (Slik 2004) (Fig. S4.1), consistent with the lower richness values predicted (explained below).

One of the two localities that had a lower predicted number of species (95) than were actually found (117) is located in the Gunung Palung National Park (1°15´S, 110°10´E; Fig.4.3 (inset)). The 15 km² study site in this park harbours seven distinctive forest types (Cannon et al. 2007). The resolution of our analysis is 100 km², and therefore cannot correctly represent the different environmental conditions present in the park.

Weighted endemism pattern and centres of endemicity

The distribution of narrowly ranged-, or endemic species is often regarded as a metric of higher importance for conservation planning than species richness (Reid 1998, Myers et al. 2000). The fact that a smaller proportion of species represented by few collections is represented by the significant SDMs (Fig. S4.3), implicates that endemicity values are expected to be even higher than presented in this study. The northern Crocker Mountains range with Mount Kinabalu, and the high mountains of central East Kalimantan have the highest weighted endemism values (Fig. 4.2b). The latter have received little collecting effort so far, and deserve further attention since they potentially harbour many new species. Similar to our data (Fig. 4.4), a curvilinear relationship between richness and endemicity was also found for the birds of Africa (Jetz et al. 2004), indicating that more endemic species

can be expected in species rich assemblages (Witt and Maliakal-Witt 2007). Spatial mapping of the relative residuals of this relationship revealed the centres of endemicity on Borneo (Fig. 4.4): the Crocker Mountains range with Mount Kinabalu; the northern parts of the central mountain range; the high mountain peaks in east Sabah; the southern extrusions of the central mountain range (Müller Mountains); the lowland east of the southern Meratus Mountains; and the eastern Sangkulirang peninsula of East Kalimantan. It is notable that our results add the Müller Mountains and the Sangkulirang peninsula to the previously known list of biologically important sites on Borneo (Fig 4.1).

Besides the entire central mountain range having positive residual weighted endemism values, this is also the case for south-west Sarawak, the southern, and south-western areas of Borneo, and around the great lakes in southern East Kalimantan. Although the absolute richness and weighted endemism values for these areas are low, they apparently harbour species which are very characteristic for those areas, and are not found elsewhere (Fig.4.4). These areas largely coincide with the WWF peat swamp-, freshwater swamp-, and heath forest ecoregions of Borneo (Wikramanayake et al. 2002).

Botanical richness, weighted endemism, and centres of endemicity explained

The results of the variation partitioning showed that for all three diversity measures only a very small proportion of the variance is explained by spatial predictors only, and that for species richness and weighted endemism the majority is explained by spatially structured environmental variables (Fig. 4.5). Although the Moran's I values of the residuals from the partial regressions with only environmental variables were slightly higher than for the models including only spatial-, or spatial and

environmental variables combined (Fig. 4.7), they do fall well within the ranges reported by Hawkins et al. (2007). They concluded that for these ranges of RSA regression coefficients were not seriously affected. Therefore we analysed the diversity patterns with the environmental predictor dataset only (Table S4.1; Fig. S4.1). The 10-fold cross validation results suggest that the predictive models for the three diversity measure performed well and were not over-parameterized, since none of the average test $R^2_{adj.}$ values differed from the average training values (Fig. 4.6).

The most important variable, when tested alone, explaining most of the richness pattern
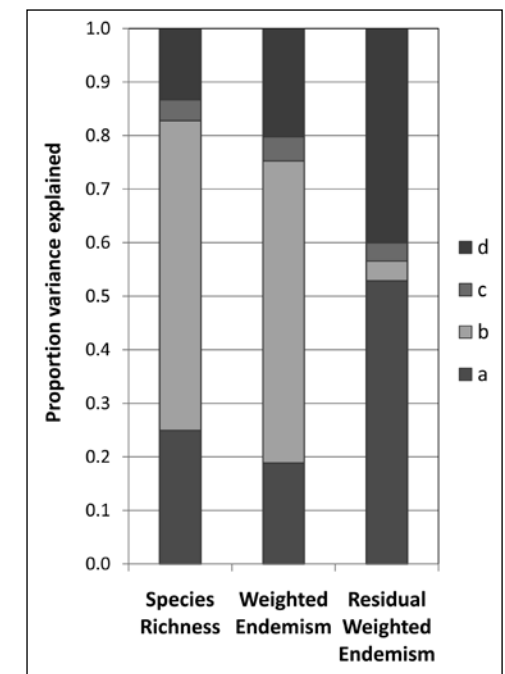

Figure 4.5. The results of variation partitioning of the forward-backward stepwise multiple regressions for species richness, weighted endemism, and relative residual weighted endemism values; a = variance explained by environmental factors, b = variance explained by spatially structured environmental factors, c = variance explained by spatial factors, and d = unexplained (residual) variance.
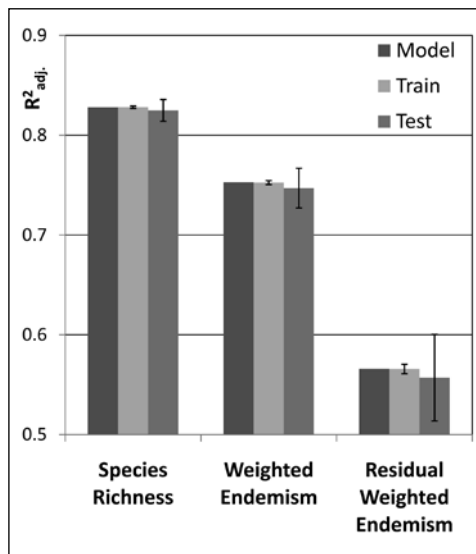
Figure 4.6. The adjusted R² values of the 10-fold cross-validation results for species richness, weighted endemism, and the relative residual weighted endemism values. Error-bars indicate ± 2SD. Average train and test R²adj. values were not significantly different (two-sample t-tests assuming unequal variances).

was the annual temperature range (Table 4.2; Bio07). The negative correlation with this variable suggests that the highest richness values were found under relative stable annual temperature conditions. The variable explaining most of the variance in species richness was temperature seasonality (Table 4.2; Bio04). This variable was positively correlated with species richness, which suggests that seasonality in temperature may be a driving factor of species richness. It should be noted however, that temperature seasonality, expressed as the standard deviation of weekly mean temperatures as a percentage of the mean annual temperatures, only ranged from 1.11-5.37%. The same two variables also accounted for almost 50% of the total explained variance of the weighted endemism pattern (Table 4.2). Stable climatic conditions maintaining high richness and endemicity values have been

found for various organisms on different continents. For the birds of Africa, low seasonality, best captured through the annual temperature range, was found to be the second most important predictor for centres of endemism (Jetz *et al.* 2004). For Amazonia, the highest botanical richness was found in areas with the shortest dry season length (ter Steege *et al.* 2003). It can be argued that habitats which face a long dry season have a larger difference in temperature between dry and wet months than habitats which remain wet throughout the year. For reptiles and amphibians in Europe, both temperature and precipitation stability were found to be important predictors of high species richness (Araújo *et al.* 2008). Araújo *et al.* (2008) even showed that it is not only contemporary climatic stability which maintains high species richness, but that stability in climate since the last glacial maximum (LGM) is an even better predictor. Similar results were found for the Australian wet tropics, predicting the highest number of species for a number of taxonomic vertebrate groups in areas which have remained climatically stable since the LGM (Graham *et al.* 2006). For Borneo there are only indirect suggestions that the areas of high richness and endemism have been stable in temperature and precipitation over longer time-scales. Geomorphic evidence suggests drier, cooler, and more seasonal climates during the LGM (Verstappen 1997), which resulted in a savanna corridor running from the southern, and south-western areas of Borneo, through the present-day Java sea and Karimata street, into south-east Asian mainland during that period (Heaney 1991, Gathorne-Hardy *et al.* 2002, Bird *et al.* 2005). There are strong indications, however, that northern Sarawak, Brunei, Sabah and East Kalimantan up to the Barito river remained forested, with everwet conditions in northern Borneo and lowland rainforest surviving around montane rainforest patches (Gathorne-Hardy *et al.* 2002). These are

the areas which coincide with the areas where the highest richness and endemicity values are predicted today.

The mechanism by which temperature seasonality (Bio04) drives high species richness and endemicity values remains speculative. There is a possible relation to phenological diversity driven by seasonal differences in abiotic conditions such as temperature and humidity (Sakai 2001). Temporal segregation of flowering minimizes interspecific overlap in flowering times, and thus ineffective pollination, or competition for pollinators. This hypothesis is known as the shared pollinator hypothesis (Sakai 2001). Whether seasonality in temperature (Bio04) within areas with a small annual temperature range (Bio07) has a clear seasonal pattern remains to be investigated, however.

Another factor of importance, explaining 9.4% and 9.1% of the variance in species richness and weighted endemism, respectively, was the ENSO drought predictor (Table 4.2; Fig. S4.1). The highest richness values were found in areas least affected by ENSO, which could indicate that severe ENSO drought impact leads to local extinction. This could also explain why the richness values for the Kalimantan lowland areas are lower than for those in north-, and north-west Borneo. These findings are supported by plot studies in East Kalimantan that found disproportionate mortality of certain species groups and tree size classes during the severe ENSO event of 1997/1998 (Slik 2004). The lower species richness values in the southern, and south-western areas of Borneo identified by the WWF as Sundaland heath- and peat swamp forest ecoregions (Wikramanayake *et al.* 2002), are explained by variables PCA052 and PCA03 (Table 4.2). Heath forests, or kerangas, are commonly found on soils known as white-sand soils, and are often covered by a layer of peat or humus. Peat swamp forests form when sediments and organic matter

builds up behind mangroves. The peat deposits can extend up to 20m (Wikramanayake *et al.* 2002). Besides along the southern coast, peat swamp habitats are also found in west Sarawak, Brunei, and around the lakes in south East Kalimantan. The variable PCA05
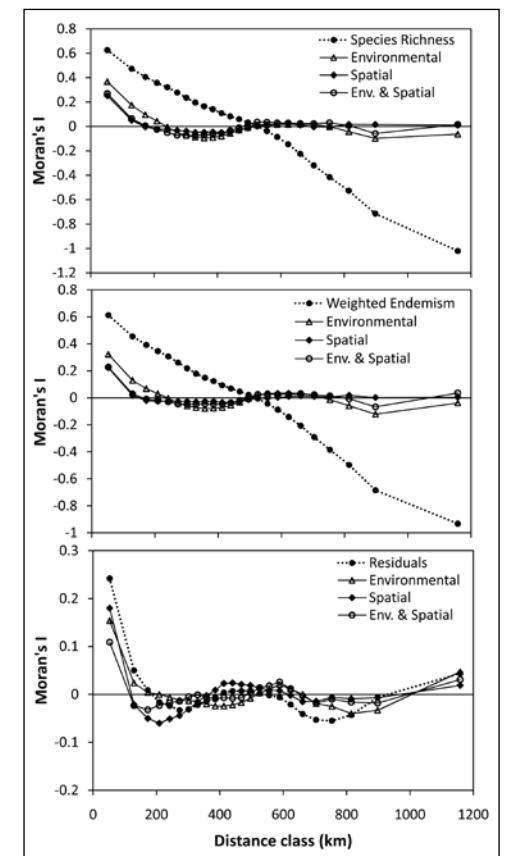

Figure 4.7. Moran's I values (equal number of pairs lags) for: species richness (A), weighted endemism (B), and the relative residuals weighted endemism values (C) (closed circles); and for the residuals of the forward-backward stepwise multiple regression results with only environmental predictors (open triangles), only spatial predictors (closed diamonds), and environmental and spatial predictors combined (open circles).

was positively correlated with the C:N-ratio of the topsoil (Table 4.1). The negative relation of PCA052 to species richness indicates that intermediate carbon content over nitrogen, characteristic for peat swamps, may have a negative effect on species richness. PCA03 was positively correlated with the textural class of the top- and subsoil (Table 4.1). The identified areas have low values for both predictors, which corresponds with coarse-textured sandy soils (FAO 2002), characteristic for kerangas and peat swamps (Whitmore 1984). Poor soil conditions, relative isolation, and the likely presence of a savanna corridor during the LGM may have resulted in low present day richness values.

The factor accounting for most of the explained variance in relative residual endemism values is the DEM (Table 4.2; Fig.4.4). Amongst others, the altitudinal range is correlated with this variable (see Methods). A large altitudinal range was also identified as the most important variable explaining the centres of African (Jetz *et al.* 2004), and South-American bird endemism (Rahbek *et al.* 2007). Jetz *et al.* (2004) argued that topographic heterogeneity might be better viewed as "a rough surrogate variable reflecting historical opportunities for allopatric speciation", which can result in centres of endemism. The mechanism suggested to drive speciation is the occurrence of narrow homothermal elevation bands serving as past and present dispersal barriers (Jetz *et al.* 2004). Other variables explaining a substantial portion of the variance in residual endemism values were PCA052, PCA03, and annual precipitation (Bio12). Where PCA052 and PCA03 explained low species richness values, the signs of the relation of these variables to relative residual endemism values were inverse, suggesting that the corresponding conditions promote speciation, resulting in positive residual weighted endemism values for the heath- and peat swamp forests (Table 4.2). Although

annual precipitation (Table 4.2; Bio12) only explained 5.1% of the variance, the annual precipitation pattern (Fig. S4.1; Bio12) showed large similarities with the pattern of the centres of endemicity (Fig. 4.4). High relative residual endemism values were found in areas with the lowest annual precipitation. All the areas are separated by wetter areas, effectively isolating the dryer areas, which might have promoted speciation.

With this study, we quantitatively analysed the Borneo wide, high-resolution botanical diversity and endemicity patterns. We showed that herbarium records can effectively be used to develop these patterns, covering areas that never have been botanically sampled. The analysis predicted an additional centre of high diversity and endemicity for the mountains in East Kalimantan, and two additional centres of endemicity; the southern extrusions of the central mountain chain, the Müller Mountains, and another on the Sangkulirang peninsula. Furthermore, our results quantitatively confirmed many of the previously recognized areas of high botanical richness and endemicity, which were based on informal expert opinions. The variables explaining most of the variance of the three biodiversity measures were comparable to other macroecological diversity studies, an indication for the reliability of our results. Additionally, our results suggested that the centres of endemicity were best explained by ecological isolation. The variables involved were altitude, soil types, and annual precipitation. Although we are confident that the estimated patterns reflect the true richness and endemicity patterns, we also stress that areas with lower values for the three diversity measures are not necessary less important for conservation. These areas may harbour species not found elsewhere, or have a forest composition, which is different from the ones

found in the 'hotspots'. We hope that our results will guide conservation efforts for the severely threatened forests of Borneo.