Detecting the richness and dissimilarity patterns of Theaceae species in southern China

Ming-Gang Zhang [a],*, Chen-Yu Huang [a], Mi-De Rao [b], Jian-Wen Chen [a], Yu-Zhuo Wang [a], Xiang-Cheng Mi [b]

[a] Institute of Loess Plateau, Shanxi University, Taiyuan 030006, China
[b] State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

ABSTRACT

Theaceae is a plant family that mainly distributed in tropical or subtropical regions of Asia. Most species of Theaceae are endemic to China. Here, we used the herbarium collections aim to: a) detect the spatial species richness patterns; b) find out the geographical divisions and c) detect the environmental drivers of the dissimilarity patterns (β diversity). The current distributions of 193 Theaceae species were modeled using MaxEnt. Meanwhile, we calculated the pair wise dissimilarity between grid cell assemblages based on beta-sim index. An ordination was used to analyze the environmental drivers of the dissimilarity patterns. We found that the diversity hotspots for Theaceae species mainly locate in Fujian and Guangxi province. Two major divisions were suggested by the ordination analysis. The longitudinal division implies the climatic divergences between southeastern and southwestern China, while the latitudinal division implies the geographical barriers. For the species in southeastern China, the dissimilarity pattern is mainly driven by temperature annual range, annual precipitation and annual mean temperature. In the southwestern China, isothermality, precipitation seasonality and the rising altitude are the three main drivers. For effective conserving the Theaceae species, the protection of the microhabitats and the high diversity areas on both sides of these divisions will be essential.

© 2016 Ecological Society of China. Published by Elsevier B.V. All rights reserved.

1. Introduction

Theace CP-T: organic carbon_pooae is a plant family that mainly distributed in southern China and adjacent Southeast Asia [1,2]. Most species of Theaceae are shrubs or trees, evergreen or rarely deciduous. It is a group with high economic value. For example, traditional Chinese tea is made from the young leaves of *Camellia sinensis* var. *sinensis* and *Camellia sinensis* var. *assamica*. *Camellia oleifera* has the longest history of cultivation and utilization in China and it can be used to extract oil for food or medicine [3]. Many species in genus *Camellia* are of great ornamental value or have been used ornamentally for hybridization, especially *Camellia japonica*, *Camellia reticulata* and *Camellia sasanqua* [2,4]. In China, there are 274 species have been recorded, including 145 species in subfam. Theoideae (106 endemic), 129 species in subfam. Ternstroemioideae (98 endemic) (Editorial Committee of *Flora Reipublicae Popularis Sinicae* 1959–2004). However, less comprehensive study at macro-scale can be found for Theaceae species of China nowadays.

Detecting the compositional dissimilarity patterns of plants is crucial for explaining the spatial biodiversity patterns [5]. Based on the compositional dissimilarity patterns between spatial units (β diversity in community ecology), the major biogeographical divisions can be recognized [6]. For example, the whole world has been successfully divided into different biogeographical regions based on the compositional dissimilarity of animals [7]. At regional level, the compositional dissimilarity index was applied to delineate the floristic regions of Yunnan province in southwest China [8]. The study on compositional dissimilarity patterns of certain group of plants can provide an effective way to enhance the understanding of ecological, evolution and historical relevance [6]. Consequently, these aspects also play an importance role for managing the biodiversity and conservation planning [9].

Nowadays, species distribution modeling is proved to be an effective tool in macro-scale studies [10]. Species distribution models attempt to identify the complete distribution ranges of certain species by combining the environmental conditions and species presence information. Currently, most specimens that stored in Chinese herbaria have been

* Corresponding author.
  E-mail address: zhangmg@sxu.edu.cn (M.-G. Zhang).

digitalized and more and more species occurrence is available online [11]. Meanwhile, many independent applications are available online, such as: Generalized Linear Model (GLM), MaxEnt, Random Forest and so on. In this exercise, we aim to (1) identify the diversity hotspots of Theaceae species using the stacked layers; (2) find out the major longitudinal and latitudinal biogeographical divisions based on the compositional dissimilarity index of the spatial units; (3) detect the environmental drivers for the dissimilarity patterns.

2. Materials and methods

2.1. Species presence data

Initially, we completed a species list of Theaceae species according to Flora of China (FOC) and *Flora Reipublicae Popularis Sinicae* (FRPS) (<http://frps.eflora.cn/>). The presence information for Theaceae species was primarily obtained from the National Specimen Information Infrastructure (NSII) (<http://www.nsii.org.cn/>, accessed December 2015). The NSII digitalized most plant specimens that stored in 42 major herbaria of China and it is the most comprehensive species distribution dataset nowadays. Using the location descriptions as provided on the labels, we matched latitude and longitude information for each record by mapping the label locations on a high resolution map of China. The species presence information were scored in $20 \times 20'$ grid cells, avoiding duplicate species records in each grid cell. This spatial solution was used because it provided a good balance between geo-referencing accuracy and the spatial resolution of available species occurrence data. Species that were present in fewer than 5 grid cells were removed from the analysis because of the statistically limitations in habitat association analyses when building the species distribution models (SDMs). Finally, 193 Theaceae species were kept to be modeled for further analysis.

2.2. Environmental predictors

In order to model the spatial distribution patterns for the 193 Theaceae species, 35 environmental predictors were selected. These predictors include 19 bioclimatic predictors (1950–2000) plus altitude of the WORLDCLIM dataset (<www.worldclim.org>) and 15 soil variables selected from the Food and Agriculture Organization (FAO) database. All the layers were resampled into $20 \times 20'$ grid cells to match the spatial solution of the Theaceae species distribution information. To avoid the over-fitting problem in species distribution modeling [12], the spearman's rank correlation tests were applied to remove multi-colinearity of variables for both bio-climatic and soil predictors in SPSS 19.0. Only the least correlated variables (spearman's rho <0.75) were kept to build the model. From those variables with Spearman rho higher than 0.75 only the ecologically most meaningful factors were kept, i.e. the information of mean temperature of warmest quarter (Bio10) and mean temperature of coldest quarter (Bio11) can be both included in annual mean temperature (Bio1). The bio-climatic and soil predictors that were kept for further analyses can be found in Tables 1 and 2. All the layers were handled with the ArcGIS 9.3 (ESRI Inc. USA).

Table 1
Spearman's rank correlation coefficients (r) for elevation and 5 bio-climate predictors, Bio01: Annual Mean Temperature. Bio03: Isothermality (P2/P7) (* 100) (P2: Mean Diurnal Rang). Bio07: Temperature Annual Range (P5–P6). Bio12: Annual Precipitation. Bio15: Precipitation Seasonality (Coefficient of Variation).

jlcl	Elevation	Bio01	Bio03	Bio07	Bio12
Bio01	−0.638				
Bio03	0.716	−0.144			
Bio07	−0.168	−0.412	−0.504		
Bio12	−0.455	0.514	−0.147	−0.610	
Bio15	0.296	−0.568	0.179	0.370	−0.387

Table 2
Spearman's rank correlation coefficients (r) for the 9 soil predictors, BS_T: base saturation% topsoil. CE-S: CEC soil subsoil. CN-T: C:N ratio class topsoil. CP-T: organic carbon_pool top-soil. Depth: effective soil depth. Drain: soil drainage class. NN-T: nitrogen% topsoil. Prod: soil production index. Text.: Textural class top-soil. CEC: cation exchange capacity.

	BS-T	CE-S	CN-T	CP-T	Depth	Drain	NN-T	Prod
CE-S	.483							
CN-T	−0.010	.272						
CP-T	.086	.430	.431					
Depth	.160	−.130	−.041	.061				
Drain	.446	.293	.022	.302	.442			
NN-T	.288	.540	.350	.591	−.175	.179		
Prod	.519	.180	−.231	.157	.420	.611	.090	
Text.	−.107	.065	.318	.289	−.028	−0.002	.233	−.106

2.3. Species distribution model building

The modeling application MaxEnt (< www.cs.princeton.edu/ ~schapire/maxent/>; ver. 3.3.3 k) was used to build the species distribution model. When dealing with the presence-only data or few presence records are available for a species, MaxEnt was proved to be specifically useful [13]. Maxent was run with the following modeling rules: (1) linear features (5–10 collection records), (2) quadratic features (10–14 records), while (3) hinge features (≥15 records) [14]. The threshold independent and prevalence insensitive area under the curve (AUC) of the receiver operating characteristic (ROC) plot was used to measure the accuracy of SDMs, which was produced by Maxent. In this excise, the SDMs were built based on presence-only data and pseudo-absences, pseudo-absences means the absences sites were randomly selected at localities where no species presence was recorded. Here, the standard measures of accuracy (AUC > 0.75) were used [14].

2.4. Species diversity patterns and compositional dissimilarity of spatial units

Once all SDMs were built, we need to select a threshold for MaxEnt prediction values to define presence or absence for a species in the spatial units. We applied the 'sensitivity specificity equality' (species with 5–9 records) and the fixed '10 percentile presence' threshold (species with ≥10 records) in this study [15]. Once the threshold is set, 193 presence/absence layers of the Theaceae species become available. In the next step, we make a stack for all the presence/absence layers. Furtherly, a presence/absence matrix was created, the rows representing the spatial units and the columns representing the presence/absence of the Theaceae species. Species richness was then mapped in ArcGIS 9.3.

In order to measure the pair wise dissimilarity between grid cell assemblages, an appropriate metric is needed. In community ecology, the methods Sørensen/Bray–Curtis, Jaccard and Kulczynski were frequently used, but these are all strongly affected by differences in species richness [6]. This means that a change in community composition has a greater relative influence in relatively species-poor than in species-rich assemblages. Therefore, a metrics that are least affected by the variation in richness was chosen in this study [6]. Here, we used the beta-sim index based on Theaceae distribution data:

$\beta_{\mathrm{sim}} = 1 - \frac{a}{\min(b,c)+a}$ where a = the number of species shared between two grid cells and b and c represent the number of species unique to each grid cell. The β_{sim} value ranges from 0 to 1, where 0 means pairs of grid cells have identical taxa lists, and 1 means that they share no taxa.

2.5. Ordination and relative environmental turnover (RET)

By arranging the spatial units along reduced axes based on dissimilarity pattern, ordination is widely used to produce low dimensional projections of multivariate data. In this exercise, we used the 'metaMDS'