Ecological Informatics 42 (2017) 79–89

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Assessing spatial distribution of *Coffea arabica* L. in Ethiopia's highlands using species distribution models and geospatial analysis methods

Binyam Tesfaw Hailu[a,*], Mika Siljander[b], Eduardo E. Maeda[c], Petri Pellikka[b]

[a] Addis Ababa University, School of Earth Sciences, P.O. Box 1176, Addis Ababa, Ethiopia
[b] University of Helsinki, Department of Geosciences and Geography, P.O. Box 68, FI-00014, Helsinki, Finland
[c] University of Helsinki, Fisheries and Environmental Management Group, Department of Environmental Sciences, P.O. Box 68, FI-00014, Helsinki, Finland

ABSTRACT

Though there is an increase in popularity of predictive modelling for assessing the geographical distribution of species, there is still a clear gap on explaining geospatial methods to derive the presence/absence of species in terms of geospatial extent besides the ambiguity of robust models. In this paper, we evaluate four major species distribution modelling methods: Artificial Neural Network (ANN), Support Vector Machines (SVM), Maximum Entropy (MaxEnt) and Generalized Linear Model (GLM) with pseudo absence and background absence data. To investigate the efficacy of these models, we present a case study using *Coffea arabica* L. species in Ethiopia as there was no species distribution modelling that has been done at a local scale especially in the coffee growing areas. We made predictions on 75% subsets and validation on 25% of the 112 presence of the species records that were collected from field observation and 0.5 m spatial resolution of true colour aerial photographs. Twelve biophysical explanatory variables; climatic, remote sensing based and landscape variables were employed in modelling. The results show that MaxEnt with pseudo absence data and SVM with background absence have highest area of understory coffee presence prediction with 12.2% and 23.1% area coverage of indigenous forest, respectively. The result from the model performance test using True Positive Rate (TPR) shows that GLM and SVM with pseudo absence data performed highest (TPR = 0.821). MaxEnt and SVM were the robust modelling methods (TPR = 0.964) using background absence data.

1. Introduction

Understanding the spatial distribution of any given species has been the focus of scientific research for centuries. Early studies such as those conducted by Von Humboldt and Bonpland (1805) attempted to give details on the reasons for how some species are present in some places and absent in others. Since mid-1980s, a new concept of quantitative species spatial distribution modelling (SSDM) has emerged, which uses geographic information systems (GIS) and remote sensing along with the increase of advanced ecological applications. Currently, vast numbers of applications exist at the local, regional, and global scales (Guisan and Zimmermann, 2000; Rushton et al., 2004; Elith et al., 2006). SDMs, also known as Ecological Niche Models, are now widely used to /indicate/ascertain/ the ecological requirements of species and also to predict their geographical distributions (Elith and Leathwick, 2009) SDMs are also used for the evaluation of mapping the spread of invasive species (Shatz et al., 2013). For example, one of this class of environmental (or ecological) niche model, the Maximum Entropy (MaxEnt) model, predicts the relative suitability of species habitat The MaxEnt model does this by using a correlation approach that evaluates the environmental conditions that satisfy a species ecological requirements (Warren and Seifert, 2011; Chemura et al., 2016). The assumption of niche theory is that uni-modal curves with symmetric Gaussian-shaped and close connection with the continuum concept when it is applied to plants (Austin and Smith, 1989). Recent evidence, Austin (2005), supports the incidence of unimodal response curves with diverse skewed asymmetric or symmetric shapes for mapping plant species. Comparisons of methods seldom use parameterization such as multiple linear versus curvilinear terms, or an ordinary type of data (presence/absence or counts), or predictors. Consequently, identifying the best SDM methodology for specific cases is often subjective.

Coffee is one of the most valuable traded commodities and it has been the second most traded commodity after oil in the post-World War II period (Davis et al., 2012; Ponte, 2002). In 2014, coffee accounted for an estimated US$ 13.9 billion of exports as some 5 billion kg were shipped and about 26 million people in 52 producing countries were engaged in total coffee sector employment (ICO, 2015). *Coffea arabica* (*C. arabica*) originally comes from the highlands of Ethiopia (Davis

* Corresponding author.
E-mail addresses: binyam.tesfaw@aau.edu.et (B.T. Hailu), mika.siljander@helsinki.fi (M. Siljander), eduardo.maeda@helsinki.fi (E.E. Maeda), petri.pellikka@helsinki.fi (P. Pellikka).

http://dx.doi.org/10.1016/j.ecoinf.2017.10.001
Received 6 July 2016; Received in revised form 2 October 2017; Accepted 4 October 2017
Available online 14 October 2017
1574-9541/

et al., 2012; Teketay, 1999). Specifically, highlands of south-western Ethiopia has *C. arabica* with most diverse populations (Gole et al., 2008). Coffee is the backbone for the economy of Ethiopia as it contributes 41% of total foreign exchange earnings in 2005 (IMF, 2006). *C. arabica* grows as a wild native understory shrub in the forests of Ethiopia's highlands (Gole et al., 2008; Hernandez-Martinez et al., 2009). Ethiopia is the world's third largest producer/exporter of *C. arabica*. It is home to at least 95% of the genetic resources for the species (Davis et al., 2012).

The habitat of wild coffee is natural forest and the site researched in this study lies within the Eastern Afromontane Biodiversity Hotspot that is known for plant diversity and its large numbers of endemic species internationally. This region is also known for the imminent threat of habitat destruction (Gil et al., 2004). The montane rainforest with wild coffee occupies a majority of the production area whereas plantation of coffee amount to only 6% of the total coffee production area in Ethiopia (Teketay, 1999). Local farmers either manage the wild coffee stands clearing some canopy trees and the competing undergrowth vegetation or harvest wild coffee fruits inside the natural forests. Coffee plants contribute to ecosystem processes provided by these forests that include habitat provisioning for a diverse wildlife community, soil conservation, and regulation of climate and atmospheric fluxes in carbon dioxide.

Recent studies on *Coffea arabica* species distribution modelling used the MaxEnt method on a regional scale (Warren and Seifert, 2011; Davis et al., 2012). However, there seems to be a lack of studies about the *Coffea arabica* species on the local geographical scale and to the best of our knowledge there have been no published studies on this species that utilizes the other Species Distribution Model (SDM) modalities. The SDMs were used in this research to determine the *Coffea arabica* species distribution in Southwestern highlands of Ethiopia including MaxEnt. We evaluated the predictive capacity of SDMs for estimating the presence/absence of *Coffea arabica* species. The models were parameterized using geospatial variables i.e. the climatic variables were precipitation, minimum temperature, maximum temperature, and evapotranspiration. The remote sensing variables were: (Normalized Difference Vegetation Index (NDVI), Simple Ratio (SR), Shadow Fraction. and landscape variables: (distance to roads, distance to river, Digital Elevation Model (DEM), slope).

Mapping understorey coffee is very challenging because any cultivation that takes place under the canopies of scattered exotic trees as understory cultivations. First, we use object oriented classification of satellite image to map indigenous forests typically associated with understorey coffee shrubs. Second, we used four major predictive modelling methods: Artificial Neural Network (ANN), Support Vector Machines (SVM), Maximum Entropy (MaxEnt) and Generalized Linear Model (GLM) to associate the understorey coffee to forests and to the environmental covariates. Finally, we evaluate the effectiveness of these models.

2. Study area

The study area is located between latitude 7.95° and 8.08° North and longitude 36.3° to 36.5° East in upstream of Didessa river basin, which is located in south western Ethiopia and is a tributary of the Blue Nile River. The extent of the study area was 19,100 ha (Fig. 1). It has an altitude that varies between 1400 m above sea level (m.a.s.l.) in the Didessa River downstream to 2400 m.a.s.l. in the upstream section. The topography is rugged with slopes between 0–50°. The population density is between 0.49 and 6.87 per square kilometre area.

Mean temperatures over this area ranges between 17.5 and 20.5 at the lowest and highest altitudes, respectively. Rainfall in the area ranges from 144 mm/month in the downstream of the Didessa River to 161 mm/month in the natural forest. The study area is covered by natural forests, exotic forest plantations, agriculture areas and pastures. The false-colour SPOT 5 satellite image in Fig. 1 shows that the study area is mainly covered by forest (red colour), agricultural areas (light blue colour) and pasture (pink colour) land covers.

3. Geospatial datasets

3.1. *Coffea arabica presence data*

Coffea arabica presence data were used as a response variable in the SDM models. These data were collected from field observations in a coffee growing indigenous forest cover using hand held Global Positioning System (GPS) device with an accuracy of ± 5–7 m, and colour aerial photographs. The samples were taken randomly from indigenous forest areas, which have understory coffee.

Coffea arabica data were randomly split into calibration (75%) and evaluation (25%) data sets that follow the so called "split-sample approach" suggested by Guisan and Zimmermann (2000). This approach splits the data sets used to adjust the model (calibration data), whereas the other (evaluation data) is used to evaluate the quality of model predictions from the splitting of what was originally a single data set.

3.2. *Biophysical explanatory variables*

Twelve biophysical explanatory variables with 20 m spatial resolution that covers 19,100 ha area were used in the modelling. Four of these explanatory variables were climatic variables, four remote sensing variables, and four landscape variables.

3.2.1. *Climatic variables*

Precipitation (PT), minimum temperature (Tmin), and maximum temperature (Tmax) data that were acquired from WorldClim data (Hijmans et al., 2005) are the climatic variables incorporated in the modelling (Fig. 2). These data were clipped, projected to Adindan/UTM zone 37 N co-ordinate system and resampled to 20 m. The objective of the resampling process was to match the spatial resolution of the climate variable raster with the other input dataset for the model but not augment the spatial information in the climate datasets. Therefore, it is recognized that this method have a feature of uncertainties associated with a coarser spatial scale of the original climate datasets.

3.2.2. *Remote sensing variables*

The spectral vegetation indices derived from satellite remote sensing data include: Normalized Difference Vegetation Index (NDVI), Simple Ratio (SR), Topographic Wetness Index (TWI) and Shadow Fraction (Fig. 3). NDVI was calculated using the red band (R) and the near infrared band (NIR) of multispectral SPOT image, which was also pre-processed by adjusting for variables such as atmospheric correction and topographic correction (Hailu et al., 2014) in order to get a reliable products including SR, SR and TWI. The red band shows a high chlorophyll pigment absorptions and the NIR band shows the high reflectivity of plants ('Eq. 1').

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{R})/(\mathrm{NIR} + \mathrm{R}) \tag{1}$$

NDVI value ranges between − 1 to 1. However, in the study area, the ranges of NDVI value was from − 0.05 to 0.85 (Hailu et al., 2015). The values close to zero corresponding to pixels with dry/bare soil and values ≥ 0.5 represents to dense vegetation.

The SR, which is calculated by the ratio of the near infrared reflectance and red reflectance as shown in Eq. 2 (Xavier and Vettorazzi, 2004). The SR value in the study area ranges between 0.16 and 18.98 (Hailu et al., 2015).

$$\mathrm{SR} = \mathrm{NIR}/\mathrm{R} \tag{2}$$

The Shadow Fraction (SF) was created in ENVI software using Sequential Maximum Angle Convex Cone (SMACC) method (Gruninger et al., 2004).

Topographic Wetness Index 'Eq. (3)'was developed by Beven and