Agriculture, Ecosystems and Environment 179 (2013) 25–34

Contents lists available at ScienceDirect

Agriculture, Ecosystems and Environment

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

Environmental and social factors account for Mexican maize richness and distribution: A data mining approach

Carolina Ureta[a], Constantino González-Salazar[a,b,*], Edgar J. González[c], Elena R. Álvarez-Buylla[d,1], Enrique Martínez-Meyer[a]

[a] Departamento de Zoología, Instituto de Biología, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico
[b] C3 – Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, 04510 México, D.F., Mexico
[c] Departamento de Ecología y Recursos Naturales, Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City, Mexico
[d] Laboratorio de Genética Molecular, Desarrollo y Evolución de Plantas, Instituto de Ecología, Universidad Nacional Autónoma de México, Centro de Ciencias de la Complejidad, Cd. Universitaria, Mexico City, Mexico

A R T I C L E I N F O

Article history:
Received 7 January 2013
Received in revised form 19 June 2013
Accepted 22 June 2013
Available online 7 August 2013

Keywords:
Zea mays
Agrobiodiversity
In situ conservation
Center of crop origin and diversification
Ethnic groups
Data mining
Races distribution

A B S T R A C T

Food security is a key topic for human welfare worldwide. In this context, the agrobiodiversity in centers of origin and diversification (COD) for the world's staple crops will be critical for feeding the world under changing environmental and social conditions. Maize is one of the most widely cultivated cereal and is the staple food for African and Latin-American countries, including its COD: Mexico, harboring more than 60% of the world's diversity for this crop. In this study we implemented a data mining approach that allowed us to evaluate spatial relationships of environmental (altitude, climate, slope and soil) and social factors (education and ethnic groups) with the spatial distribution of Mexican races, as well as the areas that can potentially harbor the highest number of races (PRA). In contrast to commonly used species distribution approaches, the data mining method implemented here allowed the integration of contrasting types of variables and their spatial relationships with the focal entity. Our results indicate that altitude, which is related with climate, was the factor with highest predictive power for most races. However, different factors showed different degrees of association with the spatial distribution of particular races. In any case, the performance of the model increased when using all evaluated factors. In our example study case, the highly vulnerable race Palomero Tolqueño was mainly influenced by climate, implying that climate change might threaten its preservation by reducing the areas with favorable conditions for its cultivation. Importantly, however, for this and other eleven races analyzed in detail, is that the ethnic group was the factor with the greatest predictive power. This finding further reinforces the key importance of *in situ* conservation by supporting local indigenous communities who are in charge of preserving, adapting to changing challenges and cultivating local races.

1. Introduction

The preservation of agrobiodiversity at Centers of Crop Origin and Diversification (COD) is critical for global food security (Thrupp, 2000; Esquinas-Alcazar, 2005). Such diversity is dynamically and communally generated at COD and hence, can only be preserved *in situ* (Bardsley and Thomas, 2005; Ortega Paczka, 2007). Therefore, an integrated understating of the role of diverse environmental and social factors on the distribution of individual races and the areas that can potentially harbor the highest number of races (PRA) becomes critical. Yet, such analyses are scarce or have not been conducted for the staple crops of the world (FAO, 2012).

Maize (*Zea mays* L.L.) is one of the most widely cultivated cereals in the world (Nuss and Tanumihardjo, 2010). It is the basic food of many African and Latin-American countries (CIMMYT, 2012), including Mexico, its COD, where around 59 maize races with thousands of varieties comprise 60% of the maize genetic diversity in the world (Ruiz Corral et al., 2008). Although the race unit is not commonly used in cultivated plants, it has been quite useful to systematize maize diversity. Anderson and Cutler (1942) defined a race as "a group of related individuals with enough characteristics in common to permit their recognition as a group".

Such maize diversification has responded to environmental and biotic conditions, but also to social variables since it was

* Corresponding author at: Ciudad Universitaria, 04510, Instituto de Biología. D.F., Mexico. Tel.: +52 56229169; fax: +52 55680847.
E-mail addresses: carolina.ureta@hotmail.com (C. Ureta), cgsalazar7@gmail.com (C. González-Salazar), emm@ib.unam.mx (E.J. González), edgarjgonzalez@ciencias.unam.mx (E.R. Álvarez-Buylla), eabuylla@gmail.com (E. Martínez-Meyer).
[1] Present address: University of California, Berkeley, College of Natural Resources, Department of Plant and Microbial Biology, 431 Koshland Hall, Berkeley, CA 94720, United States. Tel.: +1 510 642 6731/510 642 5601.

0167-8809/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.agee.2013.06.017

domesticated around 5000 years ago (Kato et al., 2009). Mexican maize production is highly dependent on environmental factors, but it is also influenced by complex and dynamic socio-cultural processes (Sánchez González and Goodman, 1992; Bellon and Brush, 1993; Brush and Perales, 2007; Kato et al., 2009; Bellon and Hellin, 2011). Indeed, 75–80% of maize cultivation depends on small-hold farmers (mestizos and indigenous peasants) using traditional methods (Brush and Perales, 2007). Consequently, Mexican maize diversity conservation should rely on criteria that incorporate natural and social factors. Despite the acknowledgment, a study providing an integrated analysis of both types of factors is lacking and is critical for future policies concerning the preservation and use of Mexican maize diversity upon a plethora of challenges faced by maize production and diversity conservation. Among such challenges, the alteration of the distribution of adequate regions for maize cultivation due to climate change is outstanding (Ureta et al., 2012). Also of importance is the loss of rare and critical races to maintain production under specific environmental threats, meeting specific cultural or social needs, preserving maize genetic diversity for future breeding programs, and preserving to avoid or diminish farmers' vulnerability under changing environmental, sociopolitical and economic conditions.

Here, we developed such study using a recently published data mining approach (Stephens et al., 2009; González-Salazar et al., 2013). This method allows the incorporation of biotic and abiotic variables and has been used to model the distribution of wild species, leading to more accurate distribution maps in which it is possible to evaluate the relative importance of the different variables considered, and which should be combined to better understand and explain the distribution of any given species. Under the same principle, we integrated the role of environmental and social factors to evaluate and model the spatial distribution of individual races and of the areas that can potentially harbor the highest number of races (PRA). In comparison with traditional species' distribution approaches that have been previously used (Jackson and Robertson, 2011; Loarie, 2008; Prates-Clark et al., 2008), data mining allows variables of contrasting nature and format to be combined and considered.

Specifically, we: (a) identify environmental and social factors associated with races potential distribution (commonly represented by regions with suitable environmental conditions; in this specific study we also incorporated social factors) and PRA (b) evaluate which of these factors have the highest predictive power of the races potential distribution areas, (c) create probability distribution maps of the potential distribution of individual races and PRA. With the aim of presenting the methodological process and determining the factors influencing the races current distribution, we provide a detailed analysis for an exemplar race. For this aim we have selected the vulnerable Palomero Tolqueño (Ureta et al., 2012) and include the results for this race in the main body of this paper. However, the data and analyses for the rest of the 46 evaluated Mexican races are presented in the Appendixes.

2. Methods

2.1. Data sources

We obtained a georeferenced Mexican maize database from the Mexican Commission for Biodiversity CONABIO (2011) (Acevedo et al., 2011). This database contains 50 years of data collection containing 21,848 records, but for the purpose of our analyses, we eliminated spatially duplicated recordsat the cell level (see below) for each race because they do not provide additional useful information, thus remaining 7949 unique records. Also, from the Mexican races identified in the database, only 47 were considered in this study due to existing taxonomic synonyms (Bofo = Elotes Occidentales and Chiquito = Nal-tel de Altura) and because the status of some races is still under debate.

We considered environmental and social factors and evaluated their influence on the distribution of each Mexican race and on the distribution of the areas that can potentially harbor the highest number of races (24–35 races) (PRA). We also identified the socio-environmental drivers of races and richness.

The environmental variables included in this study are: climate, soil type, altitude, and slope. For climate we used 19 bioclimatic variables that provide extreme, seasonal and annual temperature and precipitation patterns (Téllez et al., 2011), whereas altitude and slope came from the Hydro 1 k database (USGS, 2010). Both climatic and topographic variables were originally in raster format at 1 km resolution. We also incorporated 21 soil types countrywide in vector format at a spatial scale of 1: 250,000 (CONABIO, 1995), therefore, in order to analyze our information, categorical data were used in its native format, but continuous variables (climate, altitude and slope) were each discretized into ten categories (Appendix 1).

In terms of social factors, we evaluated the average years of education and ethnic groups. Ethnic groups were established based on the ethnic language spoken (68 different languages were identified) and years of education (INEGI, 2005) were taken as a proxy of overall socioeconomic level, as previously done by the Mexican government (INEGI, 2012). This factor was input as a categorical variable with five categories, ranging from unfinished basic education to bachelors or higher degree (Appendix 1). Ethnic groups were established base on the indigenous language spoken and we worked with 62 out of 68 languages that have been recognized because six of them had just one or two unique georeferenced data points (INALI, 2011), and were insufficient to carry out the analysis.

2.2. Evaluation of factors influencing the distribution of races

We used a recently published non-parametric data mining method (Stephens et al., 2009; González-Salazar et al., 2013) to find significant geographical or spatial associations between particular races and the factors being considered. This method is based on geographical co-occurrences of factors and a target entity: presence of a particular maize race or a unit area with more than 24 races. First, we divided the map of Mexico into a 10 km × 10 km grid (resulting in 20,719 cells); then, we assigned the presence/absence of each race to map cells and the corresponding value for all the associated factors. Such spatial association allowed us to calculate the probability of finding our race under specific conditions using the following formula:

$$P(A_i|I_k) = \frac{N_{Ai\&Ik}}{N_{Ik}} \tag{1}$$

where, $N_{Ai\&Ik}$ is the number of cells in which there is a co-occurrence of the distribution of the Mexican race A_i and factor $\boldsymbol{I}_k$; and $N_{\boldsymbol{I}k}$ is the number of cells in which the factor I_k is distributed.

However, $P(A_i|I_k)$ is a simple probability that does not take into account sample size which is important to determine significance. To incorporate the sample size and avoid bias in our results, we considered the following statistical test:

$$\varepsilon(A_i|I_k) = \left(\frac{N_{Ik}(P(A_i|\boldsymbol{I}_k) - P(A_i))}{N_{\boldsymbol{I}k}P(A_i)(1-P(A_i))}\right)^{1/2} \tag{2}$$

Where $P(A_i)$ is the probability of finding our entity under study; $P(A_i|\boldsymbol{I}_k)$ the probability of finding the latter when factor I_k is present; and $N_{\boldsymbol{I}k}$ is the number of cells in which the factor I_k is distributed. This method assumes a normal distribution of epsilon values (ε) and uses as a critical value $\varepsilon(A_i|I_k) \geq 2$, which evaluates the statistical dependency between A_i and I_k relative to the null hypothesis that the distribution of A_i is independent