Agriculture, Ecosystems and Environment 192 (2014) 12–18

Contents lists available at ScienceDirect

Agriculture, Ecosystems and Environment

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

Environmental controls on cultivated soybean phenotypic traits across China

Qianqian Li[a,b], Yueming Hu[a,b,c,*], Feixiang Chen[a,b], Jinfeng Wang[d], Zhenhua Liu[a], Zhizhong Zhao[c]

[a] Guangdong Provincial Key Laboratory of Land Use and Consolidation, South China Agricultural University, Guangzhou 510642, China
[b] Key Laboratory of Construction Land Transformation, Ministry of Land and Resources, South China Agricultural University, Guangzhou 510642, China
[c] College of Agriculture and Animal Husbandry, Qinghai University, Xining 810016, China
[d] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Article history:
Received 29 July 2013
Received in revised form 10 March 2014
Accepted 16 March 2014
Available online 29 April 2014

Keywords:
Accession
Crude oil content
Crude protein content
Environmental variables
Seed weight
Phenotypic traits
Regression models
Soybean germplasm

ABSTRACT

The impacts of environmental variables on basic phenotypic traits of cultivated soybean varieties can be significant and vary with individual traits. However, such studies are extremely rare at the continental scale because of limited observations and potential collinearity and spatial autocorrelation among abiotic factors that can make the attribution difficult. This study was designed to explore and quantify environmental variables that could closely relate to soybean phenotypic traits across China. The data of cultivated soybean phenotypic traits (i.e., 100-seed weight, crude oil content, protein content, plant height) and environmental variables were compiled from 18,686 samples across 29 provinces of China. Different regression models were used to remove collinearity and spatial autocorrelation among selected variables. As the first attempt at a continental scale, our study shows that climatic and geographic variables contributed much more to trait variations than soils did, of which the minimum temperature was most critical, followed by longitude, and soil properties explained more variances on crude protein content than on others. Abiotic variables explained 29, 20, 17 and 38% of the observed variations ($P < 0.05$) of crude protein content, crude oil content, 100-seed weight and plant height, respectively. This result implies that, besides the effects of farming practices such as fertilization, irrigation, planting density, etc., biotic factors (e.g., genes) likely play a more important role in determining the phenotypic traits and their spatial variability. It is possible to improve soybean quality and yield by selecting suitable environments even though it is hard to develop a kind of soybean varieties with all ideal germplasm traits simultaneously.

1. Introduction

Soybean is one of the most valuable crops economically and nutritionally because it has the highest protein content and the second highest oil content of all legumes (Liu, 1997). China, the first country to domesticate soybeans and a major global soybean grower and consumer, has extensive distributions of soybean accessions. Of all soybean germplasm traits, 100-seed weight, plant height, protein content, and oil content are the most valuable phenotypic characteristics related to soybean seed quality (Borras et al., 2004; Pipolo et al., 2004). The variations in these traits are genetically controlled but could be highly influenced by environment conditions (Burton, 1989; Basra and Randhawa, 2002).

Soybean traits could vary greatly with geographical locations. For example, in the northwestern area of the United States, soybeans have higher seed oil content and lower seed protein content than those found in the southeastern states (Breene et al., 1988; Hurburgh et al., 1990). Seed weight, plant height, protein content, and oil content could vary largely with variations in temperature. In a controlled environmental greenhouse, 100-seed weight increased with increasing temperature to an optimum level (Sionit et al., 1987) and then declined (Sato and Ikeda, 1979; Baker et al., 1989). A similar temperature impact was reported on seed oil content that was positively correlated with temperature until the higher end of the optimum range (Dornbos and Mullen, 1992; Gibson and Mullen, 1996), whereas, the protein content showed a negative response (Dornbos and Mullen, 1992; Gibson and Mullen, 1996). At a plot

* Corresponding author at: Guangdong Provincial Key Laboratory of Land Use and Consolidation, South China Agricultural University, Guangzhou 510642, China.
E-mail address: ymhu163@163.com (Y. Hu).

http://dx.doi.org/10.1016/j.agee.2014.03.034

or local scale, an experimental field produces greater soybean seed weight and higher plant height in Japan than in Serbia because of higher precipitation in Japan (Miladinovic et al., 2006). In a natural environment at a regional or continental scale, the oil content is positively correlated with the length of day but negatively correlated with the growing season temperature and precipitation (Zu, 1983). The opposite relations were observed on the protein content (Hu et al., 1990). Physical, chemical, and biological properties of a soil significantly affect soybean growth and seed quality. For example, a reduction in potassium supply to the deficient level can lead to lower seed oil content and 100-seed weight but higher protein content (Sale and Campbell, 1986). Plant height was reported to positively respond to soil nitrogen status (Suyantohadi et al., 2010) and soil plowing layer thickness (Yang et al., 1996). These relationships are important hints on deciding which environmental variables should be included in models to evaluate the effects of individual factors on the selected phenotypic traits with a higher confidence.

Most of previous studies underscored the significance of correlations between soybean traits and limited environmental factors, and ignored potential collinearity and spatial autocorrelations among selected variables. Here, we explored and quantified the environmental controls on major phenotypic traits of cultivated soybean by statistically removing potential collinearity and spatial correlations among all recorded geographical, climatic, and soil variables across China.

2. Materials and methods

2.1. Data collection

The dataset of cultivated soybean germplasms was provided by Chinese Soybean Germplasm Resources Inventory (CSGRI). It contains 18,686 samples that were collected from individual farms across the China as illustrated by Fig. 1. The collection period ranged from 1950 to 2000 but about 80% of all records had been collected since 1978. Each sample came with both qualitative and quantitative measurements of phenotypic traits of soybean accessions and of their associated environmental variables from 29 soybean growing provinces of China (Fig. 1). The geographical ranges for the data set are from 18°29′ to 53°33′N and 80°17′ to 134°33′E. An average annual precipitation varies from 200 mm in the northwest to 1500 mm in southeastern coastal areas, and daylight hours gradually increase from the southeast to the northwest with the annual maximum of 3550 h in Qinghai and the annual minimum of 791 h in Sichuan. This study mainly focused on four soybean phenotypic traits: 100-seed weight (SW), crude protein content (CP), crude oil content (CO), and soybean plant height (SPH). The environmental (including geographic, climatic, and soil) variables appended to each record of these four traits and their definitions are listed in Table 1. The longitude and latitude were derived from a digitized administrative map of China, and the elevation was obtained from Shuttle Radar Topography Mission (SRTM) 1-km resolution digital elevation data. The climatic variables were derived from the database of WorldClim with a 1-km spatial resolution compiled by Hijmans et al. (2005). To more accurately represent climatic influences on soybean production, we also defined growing period (GP) temperature and precipitation for each soybean accession. The GP daylight hours were calculated according to Allen et al. (1998), and the daylight intensity was reckoned as the quotient of solar radiation and daylight hours. Soil data were extracted from the Second Chinese Soil Inventory Database.

2.2. Data analysis

The analyses were conducted with SAS version 9.1 (SAS, Inc., 2004). The all 18,686 samples were grouped with the strata variable GP. Then a training dataset consisting of 4701 samples was formed using an equal probability-based random selecting procedure from all GP-based original observation groups. In other words, the number of observations for each type of GP was used as the weighting factor to specify the stratum sample sizes, so as to prevent losses of soybean diversity during the random sampling and ensure the representativeness of selected samples. Based on our preliminary study results (not shown here), the original records of some soybean traits and environmental variables were $\log_{10}$-transformed to normalize their distributions as indicated in Table 1. Regression modeling was performed to identify the importance of each variable to each individual soybean traits. The full multiple linear regressions were conducted to evaluate the overall effect and test the collinearity between predictors. Forward stepwise MLRs were performed to identify the minimum number of significant variables so as to avoid strong multicollinearity (Cocu et al., 2005). Ridge regression models were developed to remove multicollinearity between explanatory variables (Kutner et al., 2005) based on the variance inflation factor (VIF) traces – two-dimensional plots of standardized regression coefficients (Schroeder et al., 1986), and the corresponding VIF against ridge parameter (i.e. the residual sum of squares, also called ridge trace control value or biasing constant). We selected the smallest ridge trace when all standardized coefficients were stable; if possible, VIF is better around 1 (Chatterjee and Hadi, 2006). The explanatory variables, whose a ridge trace is stable but small, or unstable when the coefficient tends to be zero, were dropped in the predictive models of the ridge regression. Variables with unstable ridge traces that do not tend toward zero are also considered to remove (Ji and Peters, 2004; Kutner et al., 2005).

Spatial autocorrelation (SAC) might exist in soybean germplasm data, which violates the assumption of the independently and identically distributed residuals (Anselin, 2002) and affects hypothesis testing and prediction (Dormann, 2007; Dormann et al., 2007; Kühn, 2007; Peres-Neto and Legendre, 2010). Therefore, SAC of the regression residuals of a ridge regression model was assessed by fitting the empirical semivariogram into the theoretical spherical semivariogram using the weighted least squares method (Jian et al., 1996; Olea, 2006). The residual maximum likelihood method was used to estimate regression coefficients (Ji and Peters, 2004).

Testing results and validating models were conducted during the model-building process. Normal probability plots of residuals (P-P plots) and the Anderson–Darling test were executed on residuals to check the normality. The residuals against the predicted value were plotted to check the heteroskedasticity. The model residuals against predictors were plotted to verify the linearity. The remaining samples were used as the validation dataset to evaluate the performance of predictive models.

3. Results

3.1. Regression analysis and modeling

The results presented in Table 2 show that soybean traits were significantly correlated with all the predictive factors either positively or negatively. Strong multicollinearity was found between environmental predictors (VIF > 10 and condition index (CI) > 100), and some of the predictors were not significant at $\alpha = 0.01$ level, such as precipitation and TDH in the crude oil model (Table S1).

Compared with the full MLR, the multicollinearity of the forward stepwise regression was greatly reduced as indicated by VIF and CI (Table S2), and all the insignificant parameter estimates were removed from the models (i.e. the number of left predictors ranged from 7 to 11 out of 19 variables potentially); whereas the ridge