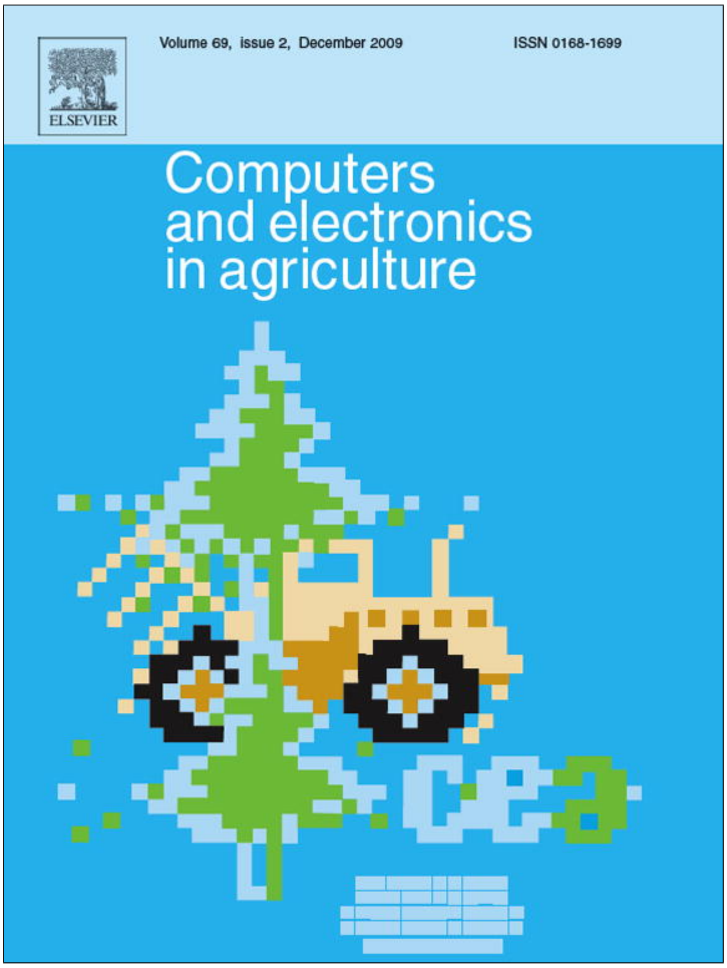

Volume 69, issue 2, December 2009 ISSN 0168-1699

Computers and Electronics in Agriculture 69 (2009) 198–208

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Analysis of Andean blackberry (*Rubus glaucus*) production models obtained by means of artificial neural networks exploiting information collected by small-scale growers in Colombia and publicly available meteorological data

Daniel Jiménez[a,c,d,*], James Cock[d,e], Héctor F. Satizábal[b,c,d], Miguel A. Barreto S[b,c,d], Andrés Pérez-Uribe[c], Andy Jarvis[e,f], Patrick Van Damme[a]

[a] *Ghent University, Faculty of BioScience Engineering: Agricultural Science, Laboratory of Tropical and Subtropical Agronomy and Ethnobotany, Coupure links 653-9000, Ghent, Belgium*
[b] *Université de Lausanne, Hautes Etudes Commerciales (HEC), Institut des Systèmes d'Information (ISI), Switzerland*
[c] *REDS Institute, University of Applied Sciences of Western Switzerland (HEIG-VD), Route de Cheseaux 1, CH 1401 Yverdon-les-bains, Switzerland*
[d] *BIOTEC, Precision Agriculture and the Construction of Field-Crop Models for Tropical Fruit Species. Recta Cali Palmira km 18, Cali, Colombia*
[e] *International Center for Tropical Agriculture (CIAT), Decision and Policy Analysis (DAPA), Recta Cali Palmira km 18, A.A. 6713, Cali, Colombia*
[f] *Bioversity International, Recta Cali Palmira km 18, A.A. 6713 Cali, Colombia*

ARTICLE INFO

Article history:
Received 19 December 2008
Received in revised form 3 June 2009
Accepted 20 August 2009

Keywords:
Andean blackberry
Small-scale growers
Artificial neural networks
Multilayer perceptron
Self-Organizing Maps
Input relevance analysis
Publicly available meteorological data

ABSTRACT

The Andean blackberry (*Rubus glaucus*) is an important source of income in hillside regions of Colombia. However, growers have little reliable information on the factors that affect the development and yield of the crop, and therefore there is a dearth of information on how to effectively manage the crop. Site specific information recorded by small-scale producers of the Andean blackberry on their production systems and soils coupled with publicly available meteorological data was used to develop models of such production systems. Multilayer perceptrons and Self-Organizing Maps were used as computational models in the identification and visualization of the most important variables for modeling the production of Andean blackberry. Artificial neural networks were trained with information from 20 sites in Colombia where the Andean blackberry is cultivated. Multilayer perceptrons predicted with a reasonable degree of accuracy the production response of the crop. The soil depth, the average temperature, external drainage, and the accumulated precipitation of the first month before harvest were critical determinants of productivity. A proxy variable of location was used to describe overall differences in management between farmers groups. The use of this proxy indicated that, even under essentially similar environmental conditions, large differences in production could be assigned to management effects. The information obtained can be used to determine sites that are suitable for Andean blackberry production, and to transfer of management practices from sites of high productivity to sites with similar environmental conditions which currently have lower levels of productivity.

1. Introduction

The Andean blackberry (*Rubus glaucus* Benth.), also known as the Andes Berry or Mora de Castilla (Bioversity International, 2005) is a fruit native to an area ranging from the northern Andes to the southern highlands of Mexico (National Research Council, 1989). It is grown as a commercial crop in Colombia, Ecuador, Guatemala, Honduras, México and Panamá (Franco and Giraldo, 2002). It is an important source of income in hillside regions of Colombia (Sora et al., 2006). Productivity varies widely between regions and also between farms. Furthermore, the crop is harvested continuously during the year and the productivity varies throughout the year. At the same time growers have little reliable information on the factors that effect the development and yield of the crop, and consequently there is a dearth of readily available information on where to grow the crop and how to effectively manage it.

Research on the Andean blackberry is limited and with the current levels of research intensity it is unlikely that technological packages can be developed for use by growers based on traditional plot based experimentation varying individual factors that affect crop production. The heterogeneous growing conditions and the continuous production throughout the year of many tropical crops mean that a large number of experiments or treatments required to draw firm conclusions concerning the optimum management of the crop under diverse conditions. The situation of a tropical

* Corresponding author at: 90 Rue de Javel, Paris 75015, France. Tel.: +33 1 457 99 038; fax: +32 9 264 62 41.
E-mail address: danieljimenez.rodas@gmail.com (D. Jiménez).

0168-1699/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2009.08.008

crop such as the Andean blackberry contrasts strongly with that of, let us say, raspberries in a temperate climate. In the case of most temperate crops, there is a relatively short and well defined harvest period and all management is geared to optimal production in that period. In tropical perennial crops that are harvested throughout the year, the number of possible combinations of management practices that need to be tested are enormous. Thus, for example Andean blackberry production during the dry season may require totally different water and pest management practices to those required for the same crop in the wet season. A direct consequence of these multiple management options is continual experimentation by producers of crops like Andean blackberries. Every time a farmer harvests his crop, there is a unique event, an unreplicated experiment (Cock, 2007). Experience with sugarcane, which is also a perennial tropical crop that may be harvested throughout the year in the low latitude tropics, has shown that by collecting information on crop production produced with the naturally occurring variation in management and the environment, the crops response can be modeled using statistical or best fit models (Isaacs et al., 2007). This approach has later been successfully applied to another perennial tropical crops, like coffee (Niederhauser et al., 2008). Given the scarce available information and the limited resources for field work research, and the high degree of heterogeneity in both growth and management, we opted for a data-driven modeling approach to provide information to growers on how to choose apposite sites for and to better manage their crops.

Crop models are basically of two types which can roughly be describe as mechanistic simulation models and best fit or statistical models. The mechanistic models have the great advantage, at least in theory, that they can be extrapolated out of the range of variation for which data exists as they are based on the basic physiological functions of the plant and their response to variation in individual parameters in the environment. Furthermore, variables that affect the observed variation in crop response to changes in the environment can be identified in causal relationships. However, these mechanistic simulation models require detailed knowledge of the functional relationships between the multiple physiological and other processes involved in crop growth and development. This knowledge base simply does not exist, and would take years to develop, for a crop like the Andean blackberry that has received little attention from researchers in the past. Statistical or best fit models are generally simpler and rely upon relationships between variations in observed crop growth and development and variations in the growing conditions. The best fit models, however, have the dual disadvantage that they can neither be used to extrapolate beyond the range of variation encompassed in the initial datasets used to develop the models, and secondly they are not able to determine whether relationships are causal or merely associations. The best fit models do, however, have the advantage that they can be constructed with a limited knowledge of the myriad individual processes and their interaction with variation in the environment that determines how a crop grows, develops and finally produces a useful product. Thus, with insufficient resources to obtain the knowledge required to develop mechanistic models, and the observation that best fit models have successfully been used in other crops, this approach was selected for Andean blackberry.

Many of the best fit models used to predict crop yields are developed using existing information on both crop production and the environment. In the case of small farm crops, such as the Andean blackberry, information on crop production is not readily available and certainly cannot readily be associated with the particular environmental conditions under which a particular crop was harvested. However, as we previously observed, every harvest is effectively an unreplicated experiment. If it were possible to characterize the production system in terms of management and the environmental conditions, and if we were able to collect information on the harvested product of a large number of harvesting events under varied conditions, it should be possible to develop best fit models for the production system. Hence, first step in developing these models was the acquisition of data on Andean blackberry production and the characterization of the production systems.

Agricultural systems are difficult to model due to their complexity and their non-linear dynamic behavior. The evolution of such systems depends on a large number of ill-defined processes that vary in time, that interact with each other, and whose relationships are often highly non-linear and very often unknown (Jiménez et al., 2008). Moreover, the available information describing these systems frequently includes both qualitative and quantitative data, the former often difficult to include in traditional modeling approaches. We surmised that bio-inspired models, such as artificial neural networks, are an appropriate alternative for developing models that can be used to improve production systems.

Artificial neural networks have been successfully used to model agricultural systems (Hashimoto, 1997; Schultz and Wieland, 1997; Schultz et al., 2000). According to Jiménez et al. (2008), these techniques are appropriate as an alternative to traditional statistical models and mechanistic models, when the input data is highly variable, noisy, incomplete, imprecise, and of a qualitative nature, as is the case of our Andean blackberry dataset. Artificial neural networks do not require prior assumptions concerning the data distribution or the form of the relationships between inputs and outputs (Sargent, 2001; Paul and Munkvold, 2005; Nagendra and Khare, 2006). They are capable of "learning" non-linear models that include both qualitative and quantitative information, and in general, they provide superior pattern recognition capabilities than traditional linear approaches (Murase, 2000; Schultz et al., 2000; Noble and Tribou, 2007). They have become a powerful technique to extract salient features from complex datasets (Chon et al., 1996; Giraudel and Lek, 2001). Furthermore, when dealing with multiple variables they can be used to produce easily comprehensible low-dimensional maps that improve the visualization of the data, and facilitate data interpretation (Barreto et al., 2007). Nevertheless, there are a number of disadvantages concerning the use of artificial neural networks, some of them are: its "black box" nature, which makes it difficult to interpret relations between the inputs and outputs, the difficulty of directly including knowledge of a ecological processes, the tendency to overtrain, and the need for enough data to be properly trained (Schultz et al., 2000; Sargent, 2001; Paul and Munkvold, 2005).

An important first step in developing models that explain variation in yield is the identification of relevant variables that affect yield: identification of these variables guides the data collection required as inputs into the model.

Several studies identify the most relevant variables, and explain given responses in agriculture through the use of multilayer preceptrons. For instance, Miao et al. (2006) implemented a neural network for identifying the most important variables for corn yield and quality. Using soil and genetic data, and a sensitivity analysis for each variable, they demonstrated that the hybrid was the most important factor explaining variability of corn quality and yield. In another study, Jain (2003) reported that the best frost prediction was obtained from the relative humidity, solar activity and wind speed from 2 to 6 h before the frost event. Paul and Munkvold (2005) predicting severity of gray leaf spot of maize (*Cercospora zeae-maydis*) in corn (*Zea mays* L.), concluded that the best variables for predicting severity were hours of daily temperatures, hours of nightly relative humidity, and mean nightly temperature. More recently, Jiménez et al. (2007) modeling sugarcane yield, suggested that crop age and water balance were highly relevant for the modeling process.

Self-Organizing Maps (SOM) have also been implemented to improve the visualization of input–input and input–output depen-

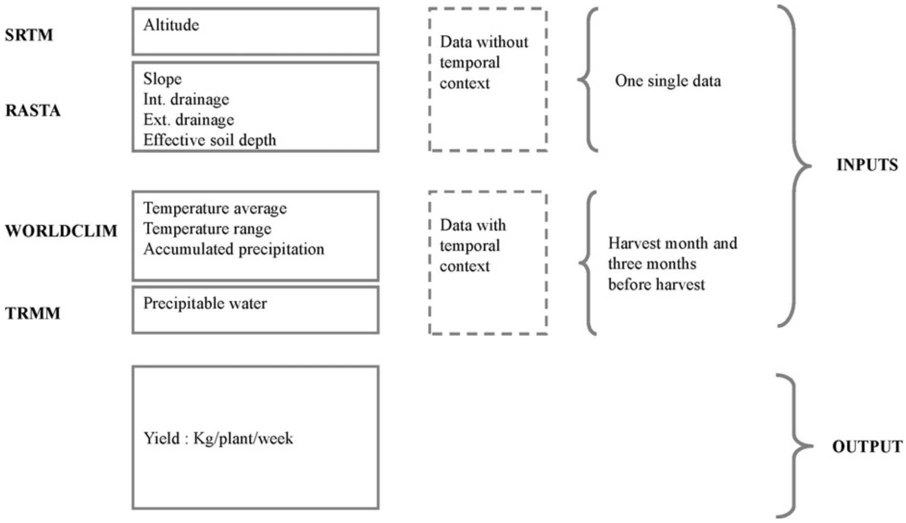

Fig. 1. Variables selected for the construction of the Andean blackberry yield model.

dencies. Thus, for example Moshou et al. (2004) found that a waveband centered at 861 nm was the variable which best discriminated healthy from diseased leaves with yellow rust (*Puccinia striiformis* f. sp. *tritici*) in wheat (*Triticum* spp. cv. Madrigal). As another example, Boishebert et al. (2006) pointed out that growing year was an important factor in the differentiation of yield of strawberry varieties.

Extension agents, expert crop advisers and growers of Andean blackberry have reached a general consensus that optimum conditions for the crop are: soils with high of organic matter content and a loamy texture, altitude between 1800 and 2400 m above sea level, average relative humidity between 70 and 80%, average temperature between 11 and 18 degree Celsius (°C), and 1500 and 2500 mm of rainfall per year (Franco and Giraldo, 2002).

The goal of this research was to demonstrate that collection of data from poor small-scale commercial producers of Andean blackberry and its analysis by means of artificial neural networks can provide growers with useful information to increase their productivity.

2. Materials and methods

2.1. Data collection and compilation

Corporación Biotec together with local Andean blackberry producers developed a simple aid based on a calendar which was used by the farmers to record information on the production of each lot planted to blackberries on their farm. The soil characteristics were determined by the soil and terrain evaluation methodology known as RASTA (Rapid Soil and Terrain Assessment) (Alvarez et al., 2004) for 20 different sites in the departments of Nariño and Caldas in Colombia. The information collected by the farmers on the calendars and with RASTA was then transferred to the database of the site-specific agriculture for tropical fruits (AEPS) project. This database includes information on location, landrace varieties, yield, harvest time and data on soil characteristics. A total of 488 records of yield from the database were included in the analysis. These records or “events” provided farmers’ estimates of the quantity (in kilograms) of fruit harvested per plant for weekly periods (Fig. 1).

Environmental information of landscape and climate was obtained for each site from a range of secondary data sources. Topography and landscape data was extracted from the Shuttle Radar Topography Mission (SRTM) (Farr and Kobrick, 2000), using the Version 3 dataset available from the CSI-CGIAR. Long-term averages for monthly temperature and precipitation were obtained from WORLDCLIM (Hijmans et al., 2005), and daily rainfall was extracted from the 3b42 product of the Tropical Rainfall Measuring Mission (TRMM) database (Bell, 1987).

2.2. Variable selection

The information compiled in the database for Andean blackberry consisted of 28 variables (Table 1). This information included categorical variables describing geographical position (large areas for departments, specific areas for particular localities within departments) and variety (thorn blackberry or normal blackberry), and environmental variables based on landscape, soil and climate (Table 1). Each yield observation was associated with the environmental variables taking into account the date of harvest (Fig. 1).

2.3. Computational models

2.3.1. Multilayer perceptron

A multilayer perceptron (Bishop, 1995) was used to model Andean blackberry yield, in such a manner that the output of the neural network, the continuous variable yield, is determined by the 28 variables we used as inputs. The Back-propagation algorithm (Bishop, 1995) was employed in order to train the neural networks. This algorithm is a gradient descent based optimizer which minimizes the difference between the desired output of the model (in the training dataset) and the actual output of the network, i.e. the mean square error (MSE).

In order to provide a mechanism for testing the model performance and to compare different models or network topologies, a training and a validation dataset were created by random sampling without replacement from the whole dataset. In this way, each training step was performed using 80% of the whole dataset, and every testing procedure to assess model performance, was performed on the remaining 20%. This method, called “split-sample” or “hold-out” validation, may assess predictive model performance, but is not recommended in its simplest form for small datasets (Goutte, 1997). However, the split sample procedure can be improved for small dataset by repeating the “split-sample” procedure several times, and then calculating the resulting performance as the average of all the tests made over the different validation subsets. Different “flavors” of this method have been used with artificial neural networks (Efron, 1983). These include “cross-validation”, “leave-one-out validation”, and “bootstrap validation”.

Table 1
Inputs used for development of Andean blackberry yield model.

Input	Variable	Type	Abbreviation	Source
1	Thorn or Normal blackberry	Cat[a]	AB_Thorn_N	AEPS
2	Nariño–Caldas (Large geographic area)	Cat[a]	Nar-Cal	AEPS
3	Nariño, la unión, chical Alto (specific geographic area)	Cat[a]	Na_un_chical	AEPS
4	Nariño, la unión,cusillo alto (specific geographic area)	Cat[a]	Na_un_cusal	AEPS
5	Nariño, la union, cusillo bajo (specific geographic area)	Cat[a]	Na_un_cusba	AEPS
6	Nariño, la unión, la jacoba (specific geographic area)	Cat[a]	Na_un_lajac	AEPS
7	Caldas Riosucio zona rural (specific geographic area)	Cat[a]	Cal_riosu_zr	AEPS
8	Altitude	Con[b]	Srtm	SRTM
9	Slope	Con[b]	Slope	SRTM
10	Internal drainage	Con[b]	IntDrain	AEPS
11	External drainage	Con[b]	ExtDrain	AEPS
12	Effective soil depth	Con[b]	EffDepth	AEPS
13	Precipitable water of the harvest month	Con[b]	Trmm_0	TRMM
14	Precipitable water of the first month before harvest	Con[b]	Trmm_1	TRMM
15	Precipitable water of the second month before harvest	Con[b]	Trmm_2	TRMM
16	Precipitable water of the third month before harvest	Con[b]	Trmm_3	TRMM
17	Average temperature of the harvest month	Con[b]	TempAvg_0	WORLDCLIM
18	Temperature range of the harvest month	Con[b]	TempRang_0	WORLDCLIM
19	Accumulated precipitation of the harvest month	Con[b]	PrecAcc_0	WORLDCLIM
20	Average temperature of the first month before harvest	Con[b]	TempAvg_1	WORLDCLIM
21	Temperature range of the first month before harvest	Con[b]	TempRang_1	WORLDCLIM
22	Accumulated precipitation of the first month before harvest	Con[b]	PrecAcc_1	WORLDCLIM
23	Average temperature of the second month before harvest	Con[b]	TempAvg_2	WORLDCLIM
24	Temperature range of the second month before harvest	Con[b]	TempRang_2	WORLDCLIM
25	Accumulated precipitation of the second month before harvest	Con[b]	PrecAcc_2	WORLDCLIM
26	Average temperature of the third month before harvest	Con[b]	TempAvg_3	WORLDCLIM
27	Temperature range of the third month before harvest	Con[b]	TempRang_3	WORLDCLIM
28	Accumulated precipitation of the third month before harvest	Con[b]	PrecAcc_3	WORLDCLIM

[a] Categorical variables.
[b] Continuous variables.

Network topology is an important issue in training a neural network model. The selection of the number of neurons in the hidden layer was made by comparing neural networks having 1,2,3,4,5,6,7,8,9 and 10 hidden units. This comparison was carried out by simple implementation of a bootstrap validation scheme (Efron, 1983). Thus, each network was tested by performing "split-sample" validations 100 times, and then the different values of the averaged MSE were compared in order to determine the network having the best performance. The topology with the lowest MSE over the validation subset had 5 units in the hidden layer neural network (Fig. 2) and was selected for further development.

An ensemble of 100 networks with the selected topology but with different initialization was built and tested in order to improve the generalization capabilities of the model (Dietterich, 2000; Brown et al., 2005). Neural networks ensembles are less affected by local minima, and have been shown to outperform their single components (Yao and Liu, 1998). In our case, the source of diversity among models was the starting point of the Back-propagation algorithm (random initialization), and the resulting model output was calculated by averaging the outputs of the 100 individual networks.

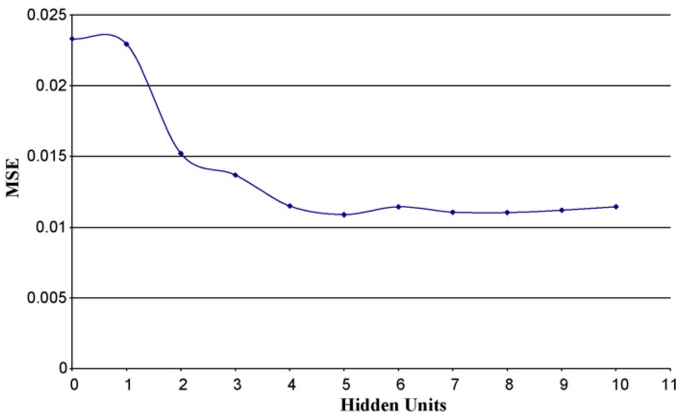

Fig. 2. MSE of artificial neural networks with different number of neurons in hidden layer.

Finally, to identify the variables which contribute most to yield; an analysis was conducted by means of the relevance metric based on sensitivity described in Satizábal and Pérez-Uribe (2007). This method assesses input relevance by calculating the partial derivative of the output of the neural network ensemble with respect to each one of the inputs. Input sensitivity should reflect input relevance because the Back-propagation algorithm finds higher connection weights to inputs having more relevant information and, in the same way, attenuates connections from noisy inputs.

2.3.2. Self-Organizing Maps

The Self-Organizing Map or SOM (Kohonen, 1995) is a non-supervised algorithm which combines clustering and visualization. The SOM maps high-dimensional datasets can be in a low-dimensional output space (generally a grid of two dimensions) with the SOM technique: observations with similar characteristics appear clustered together in the low-dimensional map produced. Such a map facilitates exploratory visual analysis of the clusters and the relationships between the variables of a complex dataset. However, a SOM does not preserve distance information. In order to address this problem the topology is disregarded, and standard clustering methods are applied to the SOM prototype vectors, and then the clusters are displayed on a lattice (Vesanto and Ahola, 1999; Barreto and Pérez-Uribe, 2007).

We chose the *K*-means algorithm to group the observations into a given number of *K* clusters. One of the limitations of this technique is the a-priori definition of the number of clusters, which is frequently unknown. To tackle this drawback, different *K* values were tested and then different groups with different number of clusters were calculated. The optimal number of *K* was then derived using the Davies–Bouldin index (Davies and Bouldin, 1979). The coordinate axes of the lattice are not clearly interpreted in terms of the original variables. Instead, variables are typically visualized by a “component plane” representation, where several lattices, one for each variable, are shown side by side. A lattice with a variable-specific coloring is called a component plane. The component plane representation is useful in finding dependencies between variables. These dependencies are perceived as similar patterns in identical areas of different component planes (Figs. 7–13). The dependency search can be eased by organizing the component planes such that similar planes are positioned near to each other (Vesanto and Ahola, 1999). In the present study, a SOM was used in order to facilitate the visualization of the relations among the productivity and the 28 environmental and geographical variables, and establish the values ranges of these variables associated with high, medium and low yield.

3. Results and discussion

3.1. Model performance

The neural network model was evaluated to ensure that its performance was acceptable for our purpose of determining relationship between the yield of the Andean blackberry and the characteristics of sites where it was grown. To evaluate the model’s performance we computed the coefficient of determination of the real Andean blackberry’s yield and the yield predicted by the model using only the data from the “hold-out” validation dataset (Fig. 3). The coefficient of determination (0.89) indicates that the model explained close to 90% of the total variation, which we considered sufficient to proceed to the next step of determining input relevance.

The fit between the real yield values and the predicted values taken from the validation data was close at the low levels of production, but was poor over the range of 69–93 (Fig. 4). At the same time, the model accurately predicted the expected yields at high yield levels. This suggests that the model can be used to determine *ex ante* the conditions and management associated with high yields and hence to provide farmers with guidelines on how to obtain high yields. In addition, the model can also determine site characteristics that are inevitably associated with poor crop performance and these can be used to indicate to farmers that a particular site and management combination is not a viable option.

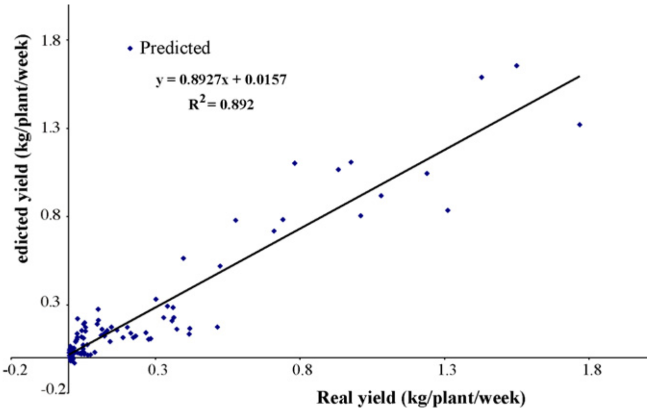

Fig. 3. Scatterplot displaying multilayer perceptron predicted yield versus real Andean blackberry yield, using only the validation dataset.

3.2. Analysis of the variables relevance

We assessed the yield response to changes in the 28 variables used in the model by obtaining the sensitivity of the model output with respect to each one of the inputs. We used the relevance metric based on sensitivity described in Satizábal and Pérez-Uribe (2007), which expresses the amount of change of the output with the variations of the inputs. The nine most important variables identified by the sensitivity metric were: soil depth, the average temperature of the first month before harvest, the specific geographical areas Nariño–la union–chical alto and Nariño–la union–cusillo bajo, the average temperature of the harvest month, the average temperature of the second month before harvest, the average temperature of the third month before harvest, external drainage and the accumulated precipitation of the first month before harvest (Fig. 5). There was a moderately sharp drop off of the sensitivity after the ninth variable (see Fig. 5). A Wilcoxon test at an alpha level of 5%

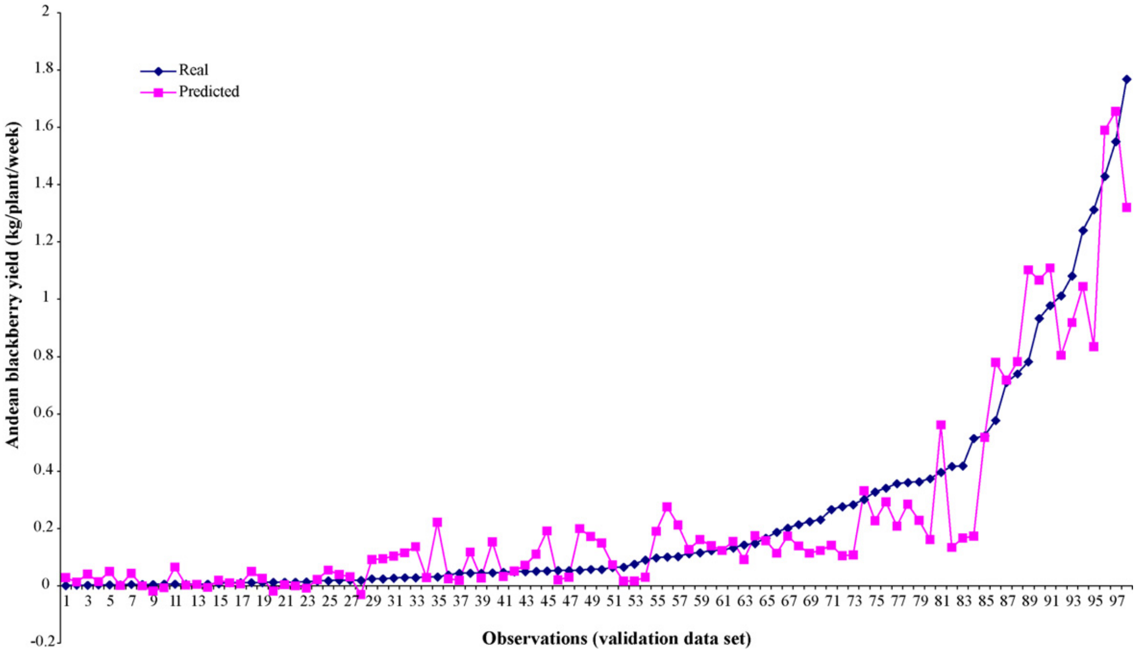

Fig. 4. Line with markers displaying the fitness of the predicted and real Andean blackberry yield through the observations from the validation dataset (yield values upwardly ordered).

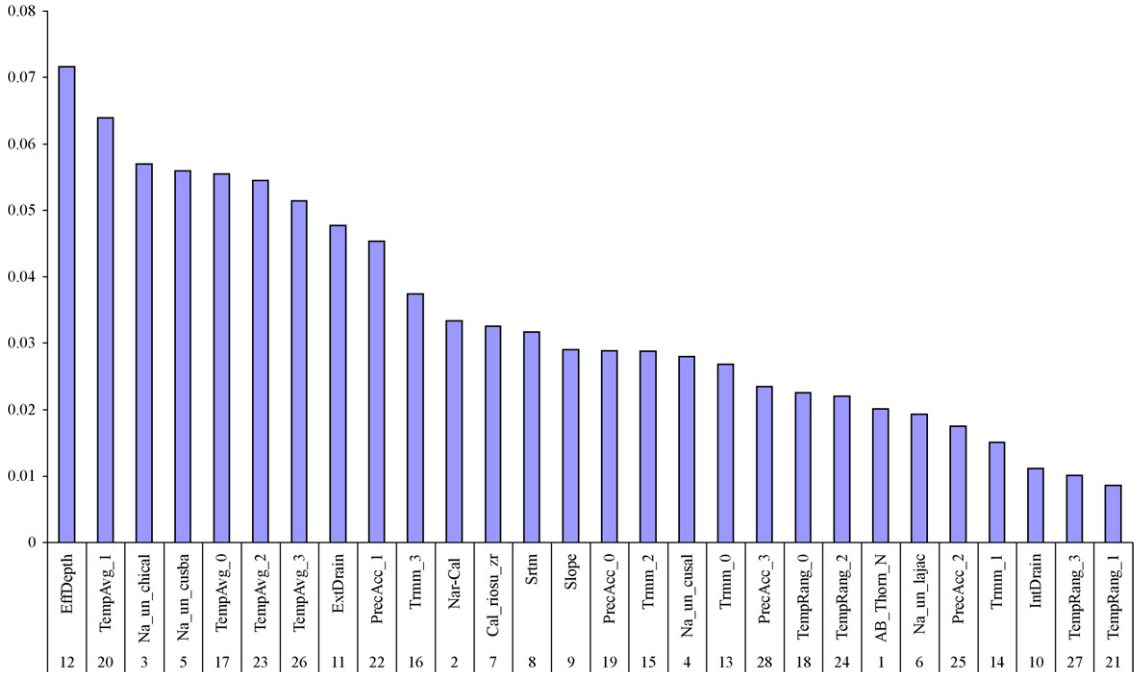

Fig. 5. Sensitivity distribution of the model with respect to the inputs.

(Table 2) indicated that the means of this group of nine variables were significantly different ($p = 0.0001$) from the rest of the variables. Hence, the nine most important variables were selected for further analysis.

3.3. *Visualization of the relations between the variables found as relevant by the sensitivity metric and clusters with similar productivity of Andean blackberry*

To further analyze the effects of the nine variables, a Kohonen map was trained with the same observations we employed to train the multilayer perceptron. The resulting bidimensional map is composed of vector prototypes which associate topological information of the original 28 variables with Andean blackberry yield (Fig. 6a). These prototypes were clustered by using the *K*-means algorithm. According to the Davies–Bouldin index, the map was divided into 6 clusters exhibiting similar features that influence Andean blackberry productivity (Fig. 6b).

3.4. *Component planes and variable dependencies*

In order to improve the visualization of the dependencies between the clusters shown in the Kohonen map (Fig. 6b) the "component planes" of Andean blackberry productivity (Fig. 7a), and the variables previously identified as the most relevant for modeling Andean blackberry yield: effective soil depth (Fig. 8), the average temperature of the harvest month, the average temperature of the first, second and third months before harvest (Fig. 9a–d), two specific geographic areas (Figs. 10 and 11), external drainage (Fig. 12), and the accumulated precipitation of the first month before harvest (Fig. 13), were separated from the Kohonen map and displayed as lattices.

3.4.1. *Productivity plane*

Yields greater than 1.16 kg/plant/week were associated with regions in cluster 2 on the Kohonen map (Fig. 7a and b). Yield values between 0.0018 and 1.16 kg/plant/week correspond to clusters 1, 3, 4, 5 and 6 in the Kohonen map. Being 3, 4, 6 the clusters with lowest yields.

3.4.2. *Effective soil depth*

Values of soil depth greater than 70 centimeters (cm) are associated with clusters 3, 4 and 6 (Fig. 8) which are all associated with low yields. In contrast, a soil depth between 40 and 70 cm appears to be related to medium to high yield clusters (1, 2, 5). The cluster with the highest yields had soil depths in the range of 60–70 cm suggesting that this level of soil depth is optimal, and that an effective soil depth greater than 70 cm is not necessary to obtain high yields. Franco and Giraldo (2002) stated that for optimal Andean blackberry development, soil depth should be deep enough to allow soil moisture retention without problems of water logging. We suggest that although soil depths above 70 cm were associated with low yields in this study this is probably due to other factors associated with the deeper soils.

3.4.3. *Average temperature of the harvest month and average temperatures of the first, second and third months before harvest*

The Kohonnen maps for temperature of the first, second and third months before harvest were similar (Fig. 9). The multilayer perceptron showed that the average temperature of the first month before harvest was more important than the others temperatures (that occurs due to small differences captured to better fit the output). However, the similarity of the components of temperature is probably due to the low monthly variation in temperature under the equatorial conditions of this study. The similarity of the temperature patterns induced us to analyze them as a group rather than

Table 2
Wilcoxon test at an alpha level of 5% comparing means of relevance between the nine most important variables identified by the sensitivity metric and means of the rest of variables.

T	*T* (expected value)	*T* (variance)	*Z* (observed value)	*Z* (critical value)	Two-tailed *p*-value
171.000	85.500	527.250	3.724	1.960	0.0001

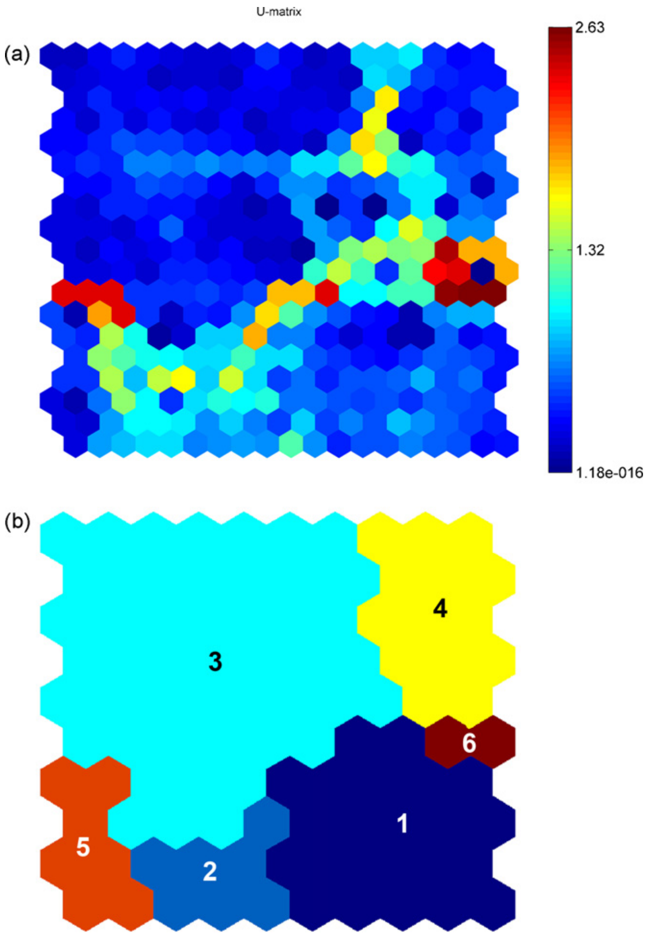

Fig. 6. Kohonen map showing the resulting clusters. (a) *U*-matrix displaying the distance among prototypes. The scale bar (right) indicates the values of distance. The upper side exhibits high distances, whilst the lower displays low distances. (b) Kohonen map displaying the 6 clusters obtained after using the *K*-means algorithm and the Davies–Bouldin index.

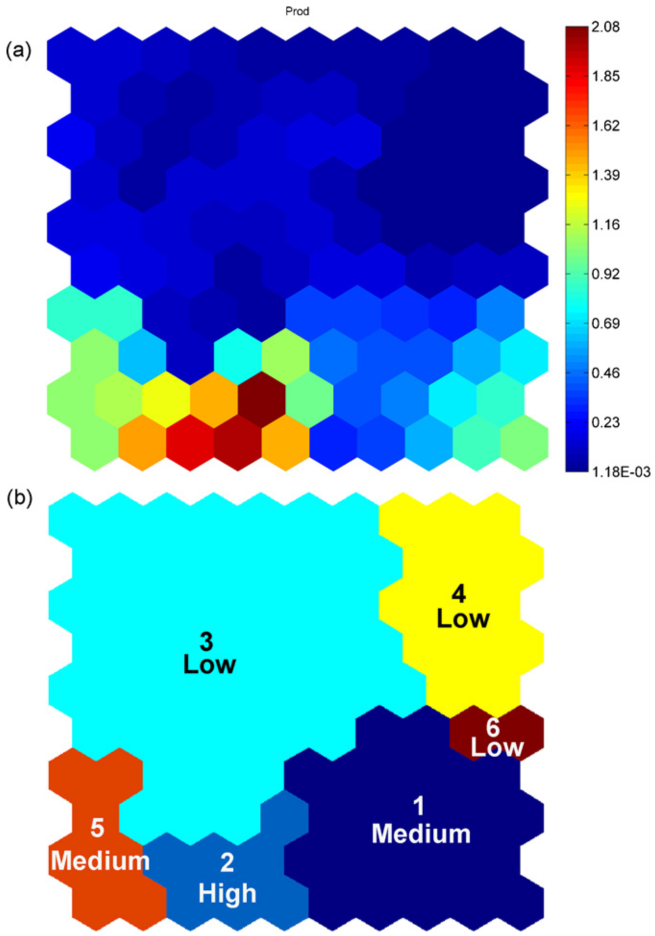

Fig. 7. (a) Component plane of Andean blackberry yield, the scale bar (right) indicates the range value of productivity in kg/plant/week .The upper side exhibits high values of yield, whereas the lower displays low values. (b) Kohonen map displaying the resultant 6 clusters and their labels according to yield values.

separately. It is immediately evident that cluster 6 with temperatures of about 24 °C is not suitable for high yields of blackberries (Fig. 9). Clusters 1, 2 and 5 with medium to high yields are related to temperatures between 16 and 18 °C (Fig. 9a–d) and low yields appear to be associated with temperatures in the range of 14–15 °C. Andean blackberry experts suggest the optimal temperature for a healthy growth of this crop is between 11 and 18 °C. We suggest a narrower temperature range with 16–18 °C associated with high yields and lower yields as the temperature moves above or below this range.

3.4.4. *Geographic areas as proxy for crop management*

Proxies can be used to estimate the effect of either immeasurable or unobservable variables on a given phenomenon (Thomas et al., 1990; Steckel, 1995; Goodman et al., 1996; Adami et al., 1999; Filmer and Pritchett, 1999; Montgomery et al., 1999). In our study, geographical areas were integrated into the model with the aim of capturing the effect of variables that were not measured. The geographical proxies were added to the analysis specifically to take into account management and social factors which were not captured by the data collection process and which are likely to be related to the geographic location of a site. For example, farmers from a given locality are likely to use similar management practices that will differ from those used by other communities living in distant localities. The localities Nariño–la union–chical alto (Fig. 10), and Nariño–la union–cusillo bajo (Fig. 11) were associated with cluster 2 which is characterized by the highest yields. Whilst the association with high yields could be a consequence of specific local environmental conditions not accounted for by the environmental variables used in the model, we suggest that is more likely that they are due to particular crop management practices related to local knowledge and socio-economic circumstances. In sugarcane certain groups of farmers consistently obtain higher yields than others even in the same edapho-climatic conditions (Isaacs et al., 2007). The difference is due to better management by certain groups which is related to socio-economic factors including access to knowledge on optimal production practices.

3.4.5. *External drainage and accumulated precipitation of the first month before harvest*

Scrutiny of the external drainage lattice (Fig. 12) gave no obvious clues as to how drainage affects the yield of blackberries. In fact medium yield in cluster 5 is associated with poor external drainage and in cluster 2 with high yields the external drainage is highly variable. However, in all clusters with medium or high yields poor external drainage is associated with low precipitation of the first month before harvest (Fig. 13): not only does this appear to be true from the Kohonnen maps, but it also makes agronomic sense. Good external drainage is evidently more important when rainfall is greater. This example clearly indicates how the visual inspection of the Kohonnen maps can assist in understanding how various factors effect the growth and development

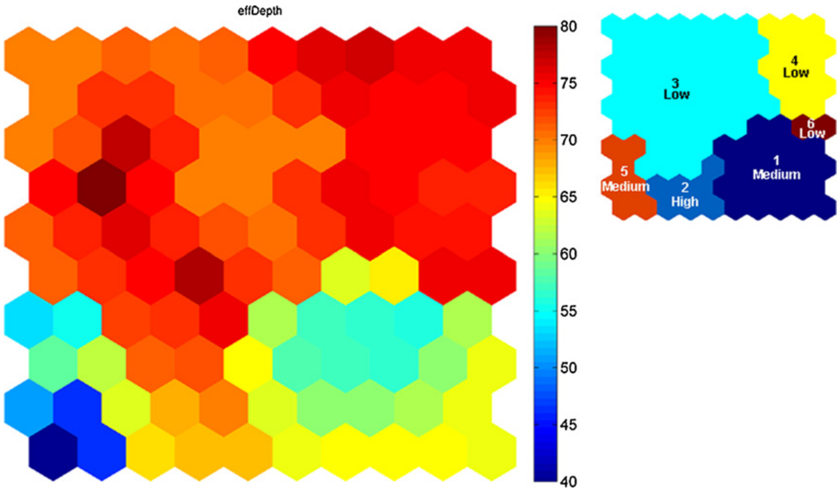

Fig. 8. Component plane of effective soil depth. The scale bar (right) indicates the range value in centimeters of soil depth, the upper side of the scale exhibits high values, whereas the lower displays low values.

of the crop and the interactions between them. Further inspection of Figs. 12 and 13 indicate that excellent external drainage is not sufficient to overcome the effects of high or moderate precipitation with moderate external drainage in cluster 3. Overall there was a tendency for low rainfall to be advantageous but there were exceptions. However, when the two variables, precipitation of the first month before harvest and external drainage are taken together it is clear that low rainfall accompanied with varied external drainage conditions can provide good yields, but that heavier precipitation of the first month before harvest with poor drainage is not conducive to high levels of productivity.

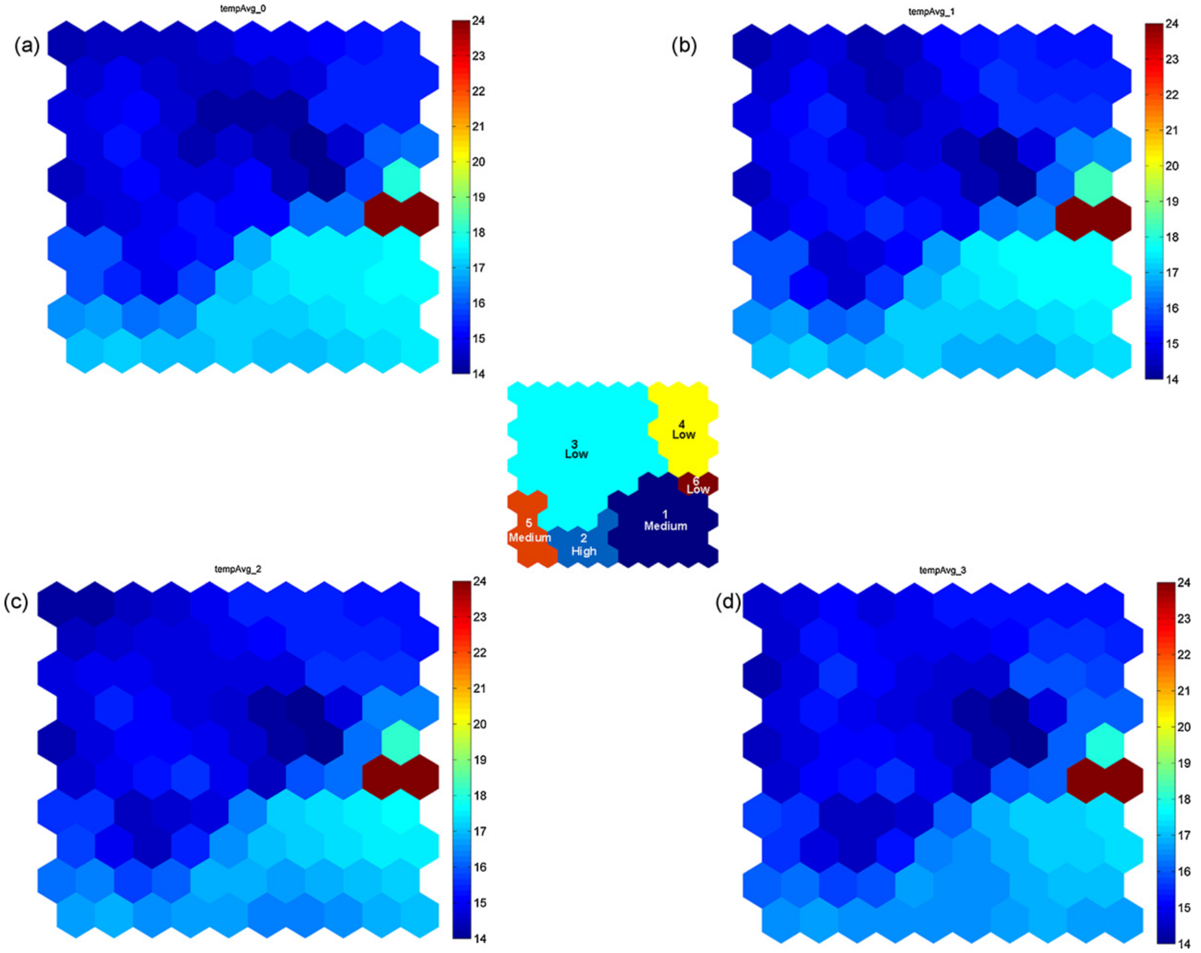

Fig. 9. Components planes of the averages temperature: (a) temperature of the harvest month, (b) average temperature of the first month before harvest, (c) average temperature of the second month before harvest, and (d) average temperature of the third month before harvest. In all figures, the scale bar (right) indicates the range value in ◦C of temperature. The upper side exhibits high values, whereas the lower displays low values.

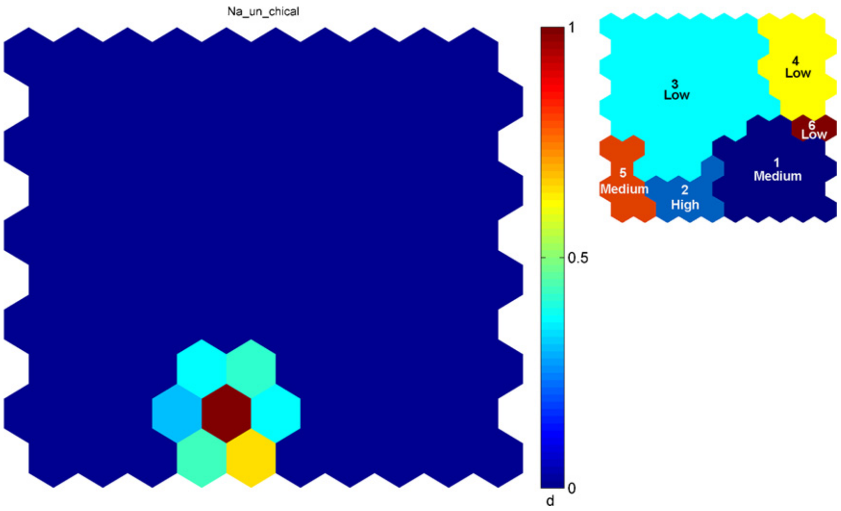

Fig. 10. Component plane of the specific geographic area Nariño–la union–chical alto. The highest values indicate presence and the lowest absence as they are categorical variables.

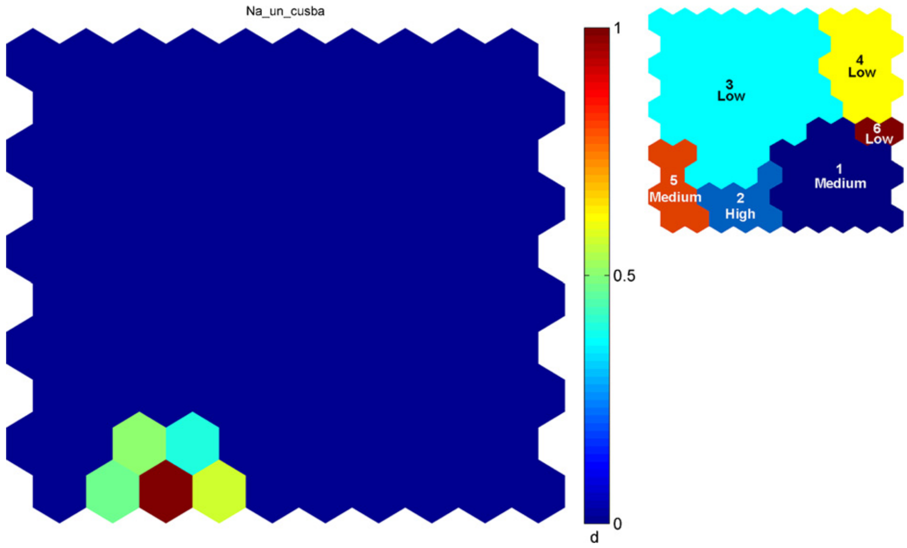

Fig. 11. Component plane of the specific geographic area Nariño–la union–cusillo bajo. The highest values indicate presence and the lowest absence as they are categorical variables.

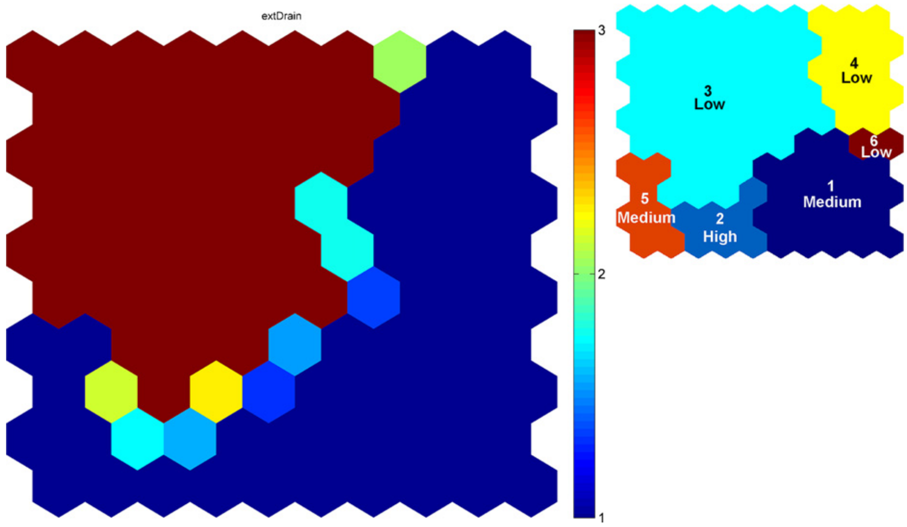

Fig. 12. Component plane of external drainage. In the scale bar (right), the highest value 3 indicates excellent or fast drainage, 2 moderate drainage, and 1 poor or slow drainage.

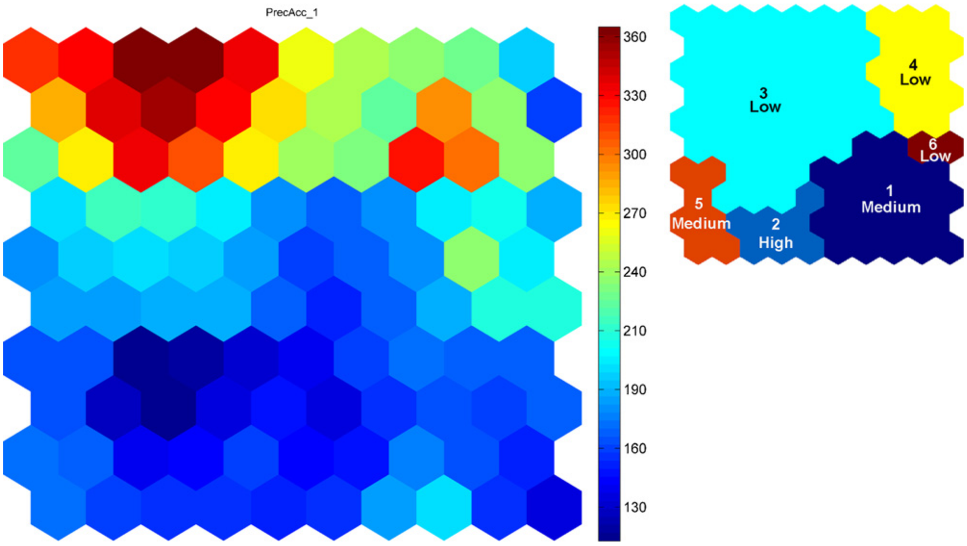

Fig. 13. Component plane of the accumulated precipitation of the first month before harvest. The scale bar (right) indicates the range value in millimeters of rainfall, the upper side of the scale exhibits high values, whereas the lower displays low values.

4. Conclusions

Data collected by small farmers in the Andes couple with information from existing data bases was successfully used to characterize specific production events and to relate production to site and time specific events. The analysis approach focuses first on identifying those variables that explain most of the yield variability by means of artificial neural networks (multilayer perceptron), and then using the Self-Organizing Maps as a tool for dimensionality reduction and visualization of input–input and input–output dependencies.

Artificial neural networks were found to be an effective tool for managing the highly variable, noisy, and qualitative nature of agricultural information collected by farmers and linked to publicly available climate databases. Multilayer perceptrons were used to develop a model based on dataset with 28 variables. This model explained close to 90% of the variation in a validation set. Sensitivity analysis was used to identify the most important variables in determining variation in yield. Self-Organizing Maps were then used to group Andean blackberry yield from different sites according to similarity of growth conditions and management. Data was not available to directly evaluate management practices, so localities were used as a proxy for management. The SOM provided a straightforward manner to visualize the distribution of the variables that affected yield. "Component planes" generated by SOM illustrated the association of these variables with yield and identified two geographic areas as highly productive. The optimal conditions for high yields are an average temperature between 16 and 18 °C, an effective soil depth between 60 and 70 cm, and low rainfall during the first month before harvest in poor external drainage locations or moderate to low rainfall in better drained areas.

The identification of geographic areas with higher yields than those that would be expected solely from the environmental conditions suggests that the farmers in those geographical areas were managing their crops particularly effectively. However, there was not sufficient information to precisely determine which management factors led to the high yields. At the same time the mere identification of areas with farmers that properly manage their crops, offers the chance for these farmers to disseminate their knowledge to other farmers with similar environmental conditions so that they too can improve yields.

Acknowledgements

This work is part of a cooperation project between BIOTEC, CIAT, and HEIG-VD (Switzerland) named "Precision agriculture and the construction of field-crop models for tropical fruits". The economical support is given by several institutions in Colombia (MADR, COLCIENCIAS, ACCI) and the State Secretariat for Education and Research (SER) in Switzerland.

References

Adami, J., Gridley, G., Nyren, O., Dosemeci, M., Linet, M., Glimelius, B., Ekbom, A., Zahm, S.H., 1999. Sunlight and non-Hodgkin's lymphoma: a population-based cohort study in Sweden. Int. J. Cancer 80, 641–645.

Alvarez, D.M., Estrada, M., Cock, J.H., 2004. RASTA (Rapid Soil and Terrain Assessment). Universidad Nacional De Colombia, Palmira, Colombia.

Barreto, M., Jiménez, D.R., Pérez-Uribe, A., 2007. Tree-structured Self-Organizing Map component planes as a visualization tool for data exploration in agro-ecological modelling. In: Proceedings of the 6th European Conference on Ecological Modelling (ECEM'07), Trieste, Italy, pp. 55–56.

Barreto, M., Pérez-Uribe, A., 2007. Improving the correlation hunting in a large quantity of SOM component planes. In: Proceedings of the International Conference on Artificial Neural Networks (ICANN 07), Porto, Portugal, pp. 379–388.

Bell, T.L., 1987. Space-time stochastic model of rainfall for satellite remote-sensing studies. J. Geophys. Res.-Atmos. 92, 9631–9643.

Bioversity International, 2005. Information Sheet on *Rubus glaucus* in New World Fruits Database. URL: http://www.bioversityinternational.org/Information_Sources/Species_Databases/New_World_Fruits_Database/. Accessed July 16, 2008.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Boishebert, d.V., Giraudel, J.L., Montury, M., 2006. Characterization of strawberry varieties by SPME–GC–MS and Kohonen self-organizing map. Chemometr. Intell. Lab. Syst. 80, 13–23.

Brown, G., Wyatt, J.L., Harris, R., Yao, X., 2005. Diversity creation methods: a survey and categorisation. Inform. Fusion 6 (1), 5–20.

Chon, T.S., Park, Y.S., Moon, K.Y., Cha, E.Y., 1996. Patternizing communities by using an artificial neural network. Ecol. Model. 90, 69–78.

Cock, J., 2007. Sharing commercial information. In: Innovation Workshop for the Agricultural Sector: Site Specific Agriculture based on Sharing Farmers Experiences, CIAT, Cali, Colombia, October, URL: http://biotec.univalle.edu.co/Memorias.htm.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE. T. Pattern. Anal. 1, 95–104.

Dietterich, T.J., 2000. Ensemble methods in machine learning. In: Multiple Classifier Systems First International Workshop (MCS 2000), Cagliari, Italy, pp. 1–15.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. 78 (382), 316–331.

Farr, T.G., Kobrick, M., 2000. Radar topography mission produces a wealth of data American geophysical. Union Eos. 81, 583–585.

Filmer, D., Pritchett, L., 1999. The effect of household wealth on educational attainment: evidence from 35 countries. Popul. Dev. Rev. 25, 85–120.

Franco, G., Giraldo, M., 2002. Condiciones ambientales del cultivo de la mora. In: Corporacion colombiana de investigacion agropecuaria, regional nueve (Eds.), El cultivo de la mora, CORPOICA, Manizales, pp. 1–3.

Giraudel, J.L., Lek, S., 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecol. Model. 146, 329–339.

Goodman, k., Correa, P., Tengana, H.J., Ramirez, H., DeLany, J.P., Pepinosa, O.G., Quiñones, M., Parra, T., 1996. *Helicobacter pylori* infection in the Colombian Andes: a population-based study of transmission pathways. Am. J. Epidemiol. 144, 290–299.

Goutte, C., 1997. Note on free lunches and cross-validation. Neural. Comput. 9 (6), 1245–1249.

Hashimoto, Y., 1997. Applications of artificial neural networks and genetic algorithms to agricultural systems. Comput. Electron. Agric. 18, 71–72 (special issue).

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Clim. 25, 1965–1978.

Isaacs, C.H., Carbonell, J.A., Amaya, A., Torres, J.S., Victoria, J.I., Quintero, R., Palma, A.E., Cock, J.H., 2007. Site specific agriculture and productivity in the Colombian sugar industry. In: Proceedings of the 26th congress International Society of Sugar Cane Technologists (ISSCT), Durban, South Africa.

Jain, A., 2003. Predicting air temperature for frost warning using artificial neural networks. Thesis. Institute for Artificial Intelligence, The University of Georgia, USA.

Jiménez, D.R., Pérez-Uribe, A., Satizabal, H.F., Barreto, M., Van Damme, P., Tomassini, M., 2008. A survey of artificial neural network-based. modeling in agroecology. In: Prasad, B. (Ed.), Softcomputing Applications in industry. Springer, Berlin, Heidelberg, pp. 247–269.

Jiménez, D.R., Satizábal, H.F., Pérez-Uribe Andrés, 2007. Modelling sugar cane yield using artificial neural networks. In: Proceedings of the 6th European Conference on Ecological Modelling (ECEM'07), Trieste, Italy, pp. 244–245.

Kohonen, T., 1995. Self-Organizing Maps. Springer, USA.

Miao, Y., Mulla, D.J., Robert, P.C., 2006. Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. Precision Agric. 7, 117–135.

Montgomery, M.R., Gragnolati, M., Burke, K.A., Paredes, E., 1999. Measuring living standards with proxy variables. Demography 37, 155–174.

Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., Ramon, H., 2004. Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks. Comput. Electron. Agric. 44, 173–188.

Murase, H., 2000. Artificial intelligence in agriculture. Comput. Electron. Agric. 29, 1–2 (special issue).

Nagendra, S.M.S., Khare, M., 2006. Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. Ecol. Model. 190, 99–115.

National Research Council, 1989. Lost Crops of the Incas: Little Known Plants of the Andes with Promise for Worldwide Cultivation. National Academy Press, Washington, DC, USA, 415 pp.

Niederhauser, N., Oberthür, T., Kattnig, S., Cock, J., 2008. Information and its management for differentiation of agricultural products: the example of specialty coffee. Comput. Electron. Agric. 61 (2), 241–253.

Noble, P.A., Tribou, E.H., 2007. Neuroet: an easy-to-use artificial neural network for ecological and biological modeling. Ecol. Model. 203, 87–98.

Paul, P.A., Munkvold, G.P., 2005. Regression and artificial neural network modeling for the prediction of gray leaf spot of maize. Phytopathology 95, 388–396.

Sargent, D.J., 2001. Comparison of artificial neural networks with other statistical approaches. Cancer Suppl. 91 (8), 1636–1642.

Satizábal, H.F., Pérez-Uribe, A., 2007. Relevance metrics to reduce input dimensions. In: Proceedings of the International Conference on Artificial Neural Networks (ICANN 07), Porto, Portugal, pp. 39–48.

Schultz, A., Wieland, R., 1997. The use of neural networks in agroecological modelling. Comput. Electron. Agric. 18, 73–90.

Schultz, A., Wieland, R., Lutze, G., 2000. Neural networks in agroecological modeling—stylish application or helpful tool? Comput. Electron. Agric. 29, 73–97.

Sora, D.S., Fischer, G., Florez, R., 2006. Refrigerated storage of mora de castilla (*Rubus glaucus*) fruits in modified atmosphere packaging. Agronomia Colombiana 24 (2), 306–316.

Steckel, R.H., 1995. Stature and standard of living. J. Econ. Lit. 33, 1903–1940.

Thomas, D., Strauss, J., Henriques, M., 1990. Child survival, height for age and household characteristics in Brazil. J. Dev. Econ. 33, 197–234.

Vesanto, J., Ahola, J., 1999. Hunting for correlations in data using the self-organizing map. In: Proceedings of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA), pp. 279–285.

Yao, X., Liu, Y., 1998. Making use of population information in evolutionary artificial neural networks. IEEE Trans. Syst. Man Cybern. B 28 (3), 417–425.