Contents lists available at ScienceDirect

Agricultural Systems

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

Interpretation of commercial production information: A case study of lulo (*Solanum quitoense*), an under-researched Andean fruit

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

[a] *Ghent University, Faculty of BioScience Engineering: Agricultural Science, Laboratory of Tropical and Subtropical Agronomy and Ethnobotany, Coupure Links 653-9000, Ghent, Belgium*

[b] *International Center for Tropical Agriculture (CIAT), Decision and Policy Analysis (DAPA), Recta Cali Palmira km 17, A.A. 6713, Cali, Colombia*

[c] *Université de Lausanne, Hautes Etudes Commerciales (HEC), Institute des Systèmes d'Information (ISI), CH 1015 Lausanne, Switzerland*

[d] *Previously Scientific Advisor at BIOTEC. Currently Emeritus, International Center for Tropical Agriculture (CIAT), Colombia*

[e] *REDS Institute, University of Applied Sciences of Western Switzerland (HEIG-VD), Route de Cheseaux 1, CH 1401 Yverdon-les-bains, Switzerland*

[f] *BIOTEC, Precision Agriculture and the Construction of Field-Crop Models for Tropical Fruit Species, Recta Cali Palmira km 18, Cali, Colombia*

ARTICLE INFO

Article history:
Received 3 May 2010
Received in revised form 14 September 2010
Accepted 21 October 2010
Available online xxxx

Keywords:
"Data model based"
"Algorithmic based"
Robust regression
Artificial neural networks
Mixed models
Lulo

ABSTRACT

Every time a farmer plants and harvests a crop represents a unique event or experiment. Our premise is that if it were possible to characterize the production system in terms of management and the environmental conditions, and if information on the harvested product were collected from a large number of harvesting events under varied conditions, it should be possible to develop data-driven models that describe the production system. These models can then be used to identify appropriate growing conditions and improved management practices for crops that have received little attention from researchers. The analysis and interpretation of commercial production data in the context of naturally occurring variation in environmental and management, as opposed to controlled experimental data, requires novel approaches. Information was available on both variation in commercial production of the tropical fruit, lulo (*Solanum quitoense*), and the associated environmental conditions in Colombia. This information was used to develop and evaluate procedures for the interpretation of the variation in commercial production of lulo. The most effective procedures depended on expert guidance: it was not possible to develop a simple effective one step procedure, but rather an iterative approach was required. The most effective procedure was based on the following steps. First, highly correlated independent variables were evaluated and those that were effectively duplicates were eliminated. Second, regression models identified those environmental factors most closely associated with the dependent variable of fruit yield. The environmental factors associated with variation in fruit yield were then used for more in depth analysis, and those environmental variables not associated with yield were excluded from further analysis. Linear regression and multilayer perceptron regression models explained 65–70% of the total variation in yield. Both models identified three of the same factors but the multilayer perceptron based on a neural network identified one location as an additional factor. Third, the three environmental factors common to both regression models were used to define three Homogeneous Environmental Conditions (HECs) using Self-Organizing Maps (SOM). Fourth, yield was analyzed with a mixed model with the categorical variables of HEC, location, as a proxy for cultural factors associated with a geographic region, and farm as proxy for management skills. The mixed model explained more than 80% of the total variation in yield with 61% associated with the HECs and 19% with farm. Location had minimal effects. The results of this model can be used to determine the appropriate environmental conditions for obtaining high yields for crops where only commercial data are available, and also to identify those farms that have superior management practices for given environmental conditions.

1. Introduction

Years of agronomic experimentation have lead to a wealth of knowledge on crop responses to variation in the growth environment. This knowledge has been used to develop empirically-based

* Corresponding author at: International Center for Tropical Agriculture (CIAT), Decision and Policy Analysis (DAPA), Recta Cali Palmira km 18, A.A. 6713, Cali, Colombia. Tel.: +57 2 445 0000x3729; fax: +57 2 445 0073.

E-mail address: d.jimenez@cgiar.org (D. Jiménez).

0308-521X/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.agsy.2010.10.004

crop models which quantify the crop response to variations in the growing conditions. The required level of knowledge to develop effective crop models only exists for those crops which have been the subject of intense research. For many minor and some major crops, models are not currently available. Moreover, it would take years of experimentation using traditional methodologies to build up the necessary knowledge base to develop them, particularly in perennial crops such as many tropical fruit species. We suggest that an alternative approach to years of research in controlled experiments is observing crops under varied management and different environments in the field.

Every time a farmer harvests a crop, this event represents an unreplicated experiment (Cock, 2007). We surmised that in the case of tropical fruit crops, commercial production data (CPD) could be used to evaluate crop response to variation in growing conditions caused both by inherent variation in the growth environment and also by variation in farm management practices. Our premise is that if it were possible to characterize the production system in terms of management and environmental conditions, and if information on the harvested product were collected from a large number of harvesting events under varied conditions, it should be possible to develop data-driven models for the production system (Jiménez et al., 2009). Furthermore, these data-driven models, based on producers' experiences in commercial production, are likely to provide more realistic and valuable information to growers than models based on small well manicured plots and experiments carried out in controlled environments: the results are likely to provide growers with site-specific recommendations, which they can use to better manage their crops according to the specific conditions of their farms.

This approach is essentially that of operational research, which observes an organization's operations and uses mathematical or computer models, or other analytical approaches to find better ways of doing them (Operational Research Society, 2006). This method is similar to those of total quality management which emphasizes monitoring, measurement and the systematic capture and codification of tacit knowledge (Bessant and Francis, 1999; Kannan and Tan, 2005). Similarly, in the medical profession systematic collection and analysis of information from the everyday lives of people is used to deduce factors associated with disease and hence to recommend methods of control (Framingham heart study, 2006). We suggest that modern information technology has advanced to the stage where even small-scale growers can benefit from analyzing their multiple production experiences. Existing public databases on climate, landscape, topography coupled with information collected at the farm level (edaphic conditions, management) can characterize the growth environment, and farmers can compile information on both their crop management practices and the crop response. This information can then be exploited by means of modeling approaches to understand yield variability and to provide recommendations to small-scale fruit growers.

Experience with sugarcane, coffee and Andean blackberry in Colombia, has shown that by collecting CPD generated with the naturally occurring variation in management and the environment, the crops response can be modeled (Isaacs et al., 2007; Niederhauser et al., 2008; Jiménez et al., 2009). There is however a major caveat to this approach. Due to the large number of variables that affect the crop response, the interactions and non-linearity of the responses and the inevitable errors in data collection at the farm level, a large number of data sets are required to make sense of the data. With the large data sets required to draw conclusions it is likely that novel analytical approaches will be necessary. Grimm (1999) suggests that modelers should experiment more with their models. In this paper we experiment with various analytical approaches and compare their efficacy.

Agricultural systems are difficult to model due to their complexity, non-linear dynamic behavior, and the large number of ill-defined processes that vary in time, interact with each other, and whose relationships are very often unknown (Jiménez et al., 2008). Hence, it has become necessary to develop modeling approaches able to deal with this high heterogeneity and natural variability. According to Breiman (2001) there are two approaches that can be used to predict the responses from input variables or extract information of the association of these variables to the response. They are "data model based" and "algorithmic based". Both models are based on the same original data, but they diverge in the assumptions and procedures used to estimate the model parameters. The "data model based" approach assumes that the processes generating the data have the form of a stochastic model, and the "algorithmic based model", instead of considering a stochastic model, assumes an unknown internal structure (i.e. it does not make assumptions about the underlying process or the distribution of the data) which is often complex. In the present paper a range of approaches are used to interpret variation in commercial production of lulo (*Solanum quitoense* Lam.), an Andean fruit grown in highly heterogeneous conditions by small producers with minimal access to information from traditional research programs based on controlled experiments. The efficacy of the different approaches is compared with examples from "data model based" and "algorithmic based" methods and combinations of the two approaches.

Artificial neural networks (ANNs) were selected as an "algorithm based" approach and multiple linear regression as "data model based" approach. Furthermore, based on the recommendation of Schultz et al. (2000) "data model based" were combined with "algorithmic based" methods in order to benefit from the advantages of both. "Data model based" approaches such as multiple linear regression and mixed models combined with Best Linear Unbiased Prediction (BLUP), are frequently used to understand the relationships between crop yield and environmental variation (Khakural et al., 1999; Kravchenko and Bullock, 2000; Piepho, 1994; Yan et al., 2002; Piepho and Mohring, 2005). However, these approaches are often not satisfactory due to their incapacity to take into account non-linear relationships between output and inputs (Gevrey et al., 2003; Miao et al., 2006) and do not handle outliers well, although some robust linear regressions have been developed to address this problem (Rousseeuw and Leroy, 1987; Lanzante, 1996; Faraway, 2002). Multiple regressions are also poor at handling categorical data (O'Grady and Medoff, 1988). In the case of commercial data, categorical variables such as farm, agroecological zone and location are likely to be important. ANNs, as non-parametric approaches, have several attractive theoretical properties: They do not require strong assumptions on the form or structure of the data (Sargent, 2001; Paul and Munkvold, 2005; Nagendra and Khare, 2006); and they are capable of "learning" non-linear models that include both qualitative and quantitative information. ANNs have demonstrated their utility in agricultural modeling (Hashimoto, 1997; Schultz and Wieland, 1997; Paul and Munkvold, 2005; Miao et al., 2006; Jiménez et al., 2008). Nevertheless, artificial neural networks also have disadvantages: They are computationally exhaustive; it is difficult to understand relations between the inputs and outputs due to their "black box" nature; it is difficult to include knowledge of ecological processes; they can be over-trained and give false expectations of their predictive capacity; and they require large amounts of data to be properly trained (Schultz et al., 2000; Sargent, 2001; Paul and Munkvold, 2005; Ozesmi et al., 2006; Jiménez et al., 2009).

Mixed models combine both random and fixed effects. When combined with BLUP, which provides linear estimates of fixed effects, the contribution of random effects to the output can be estimated. Robinson (1991), Yan et al. (2002) and Rabe-Hesketh and Skrondal (2008) demonstrated how this methodology could be

Please cite this article in press as: Jiménez, D., et al. Interpretation of commercial production information: A case study of lulo (*Solanum quitoense*), an under-researched Andean fruit. Agr. Syst. (2010), doi:10.1016/j.agsy.2010.10.004

used to compare the performance of varieties grown under a range of conditions in commercial fields with not all varieties being grown at all sites. Experience with sugarcane, coffee (Isaacs et al., 2007; Cock et al., submitted for publication) and shrimp production (Gitterle et al., 2009), suggests that one of the most effective means of analyzing commercial information is first to establish clusters of events with similar environmental conditions, and then determine the effects of variation of management practices within and between these environmental clusters and also to determine the effects of the environmental clusters *per se*.

The effects of many continuous environmental variables on production and quality of agricultural products are likely to be non-linear. For example, there is likely to be an optimal and non-linear response to such variables as average temperature, soil water content, soil pH, air humidity and diurnal temperature range. Thus, it is likely that non-linear methods will be optimal for determining the effects of environmental variables on crop quality and productivity and identifying clusters of events with similar environmental condition. Many variables recorded for commercial crop production are likely to be categorical (i.e. weed control, land preparation practices). Furthermore, categorical variables such as farm may be used as a categorical proxy variable for farm management skills associated with that particular farm. Isaacs et al. (2007) used groups of farmers defined by social characteristics, as categorical proxy variables for management and associated the various groups with different levels of productivity. At the same time other management practices may be described by continuous variables as is the case with such variables as fertilizer levels or number of irrigations. Mixed models with BLUP, which incorporates linear regression, were selected as more suitable for handling both categorical and continuous variables in the same model than pure regression models (Cock et al., submitted for publication).

We chose the example of lulo to evaluate different data-driven approaches to develop predictive models. We selected lulo as it is a poorly understood tropical fruit tree cultivated in Colombia, Costa Rica, Ecuador, Honduras, Panama and Peru (National Research Council, 1989; Franco et al., 2002; Osorio et al., 2003; Bioversity International, 2005; Flórez et al., 2008; Pulido et al., 2008; Acosta et al., 2009). Lulo is exclusively grown in tropical environments where they normally produce during the whole year, with high variability in yield in both space and time.

2. Methodology

The commercial production data (CPD) was collected on the farms, and the environmental conditions were characterized using both data collected on farm and from publically available climate databases.

The data was compiled in the Corporación BIOTEC databases for analysis. The resulting database, which is the result of merging information from different sources, was analyzed in an iterative way with the aim of finding both, the most apposite data set and the approach to model it.

2.1. Commercial production data (CPD)

Corporación BIOTEC in collaboration with lulo producers in the department of Nariño, Colombia, developed a simple method of keeping on farm records based on a calendar that also provided them with useful information on lulo production (BIOTEC, 2007). Twenty-one lulo producers recorded information on these calendars over the period of 2 years (January 2006 to December 2007). The records of the individual farms provided a data series on production of lulo for each farm with farmers' estimates of the quantity (in grams) of fruit harvested per plant for weekly periods.

The data collected in a data base by BIOTEC included information on location, varieties (management), yield, and harvest time for a total of 254 records In addition each site was geo-referenced using hand held GPS.

2.2. Biophysical characterization of sites

Weather stations in Colombia are often not close to the fields where most of tropical fruit species are grown, furthermore the information provided by these stations rarely represent the climate of individual production sites, largely due to the large variation in altitude in the region. Therefore, the generation of the climatic, landscape, topographic, and edaphic information of each site was obtained from the coordinates (latitude and longitude). With this spatial information, it is possible to extract biophysical information from high resolution interpolated publically available databases, through the use of automated algorithms implemented in Geographical Information Systems (GIS) and to estimate the climatic conditions of any site that has been geo-referenced.

Long-term averages for monthly temperature and precipitation were obtained from WORLDCLIM database (Hijmans et al., 2005), and daily rainfall was extracted from the 3b42 product of the Tropical Rainfall Measuring Mission (TRMM) database (Bell, 1987). Landscape and topography data was extracted from the Shuttle Radar Topography Mission (SRTM) (Farr and Kobrick, 2000) using the Version 3 data set available from the CSI-CGIAR.

With regard to soil, as farmers neither have the knowledge nor the resources to evaluate their soil and terrain using traditional methodologies and there is a lack of approaches, guidelines, books and field manuals for farmers or extension workers to characterize soils and terrain *in situ*, the RASTA (Rapid Soil and Terrain Assessment) system was developed and used to characterize soil conditions. RASTA is a simple easy to learn methodology that farmers can used to characterize soils and terrain without recourse to complicated classification schemes or laboratory analysis. Farmers were provided with RASTA kits and used these to characterize their soil and terrain (Alvarez et al., 2004).

We were aware that a number of management variables, that it was not possible to measure, could have a major impact on the outputs of the models. In order to evaluate these variables we used the locality as a proxy for the socio-economic conditions of a given group of farmers, and farm as a proxy for the management skills associated with a particular farm.

2.3. Variables

In the present study, four locations with 21 different lulo producing sites were characterized. The variables were chosen on a pragmatic basis using expert knowledge to identify those variables that were considered likely to influence production and also that could be readily recorded. The information was compiled in the database for lulo with 254 records, 19 independent variables describing information of each site and the dependent variable, productivity of lulo (Table 1). The independent and dependent variables of the "data model based" approaches correspond to the supervised "algorithmic based" inputs and output respectively. This information included continuous variables depicting biophysical information based on landscape, topography, edaphic conditions, climate, and categorical variables depicting variety and location (Table 1). Each yield observation was associated with the climate variables taking into account the date of harvest.

2.4. Models

Inspection of the database showed that there was a preponderance of low productivity data with only a few cases of high yields.

Table 1
Variables recorded in the lulo database.

Input	Variable	Type	Abbreviation	Source
1	*Location*			
1[a]	Nariño, cartago, san isidro	Cat[a]	Na_ca_san*	CPD
1b	Nariño, la unión, buenos aires	Cat[a]	Na_un_ba*	CPD
1c	Nariño, la unión, la Jacoba	Cat[a]	Na_un_jac*	CPD
1d	Nariño, la union, chical alto	Cat[a]	Na_un_chical*	CPD
	Variety, landscape, topography, edaphic conditions			
2	Thorn or no thorn	Cat[a]	Nar_Thorn_N	CPD
3	Altitude	Con[b]	Srtm*	SRTM
4	Slope	Con[b]	Slope*	SRTM
5	Internal drainage	Con[b]	IntDrain	RASTA
6	External drainage	Con[b]	ExtDrain*	RASTA
7	Effective soil depth	Con[b]	EffDepth*	RASTA
	Climate			
8	Precipitable water of the harvest month	Con[b]	Trmm_0*	TRMM
9	Precipitable water of the first month before harvest	Con[b]	Trmm_1*	TRMM
10	Precipitable water of the second month before harvest	Con[b]	Trmm_2*	TRMM
11	Average temperature of the harvest month	Con[b]	TempAvg_0*	WORLDCLIM
12	Average temperature of the first month before harvest	Con[b]	TempAvg_1	WORLDCLIM
13	Average temperature of the second month before harvest	Con[b]	TempAvg_2	WORLDCLIM
14	Accumulated precipitation of the harvest month	Con[b]	PrecAcc_0	WORLDCLIM
15	Accumulated precipitation of the first month before harvest	Con[b]	PrecAcc_1	WORLDCLIM
16	Accumulated precipitation of the second month before harvest	Con[b]	PrecAcc_2	WORLDCLIM
17	Temperature range of the harvest month	Con[b]	TempRang_0*	WORLDCLIM
18	Temperature range of the first month before harvest	Con[b]	TempRang_1*	WORLDCLIM
19	Temperature range of the second month before harvest	Con[b]	TempRang_2*	WORLDCLIM
Output	Lulo yield	Con[b]	Yield	CPD

[a] Categorical variables.
[b] Continuous variables.
* Final set of inputs/dependent variables used in the development of lulo yield regression models.

Statistical analyses often take the drivers of these scarce productivities as outliers, and yet it is precisely those factors associated with high yield that are of most interest. Given these characteristics of the data we opted for three data-driven approaches: two regressions: non-linear "algorithmic based" and linear "data model based"; and an iterative approach which used both, "data model based" and a non-supervised "algorithmic based" modeling combined with expert guidance.

The approaches we used to build the "algorithmic based" models were artificial neural networks. These non-parametric models are connectionist systems inspired from the structure and behavior of nervous systems. Connectionist systems are composed of elementary units performing simple calculations, which are interconnected by following some ordered pattern. In the case of artificial neural networks, these units are called "neurons" by analogy with the cells in the brain. How these neurons are connected determines the network topology. There are layered topologies in which the neurons are organized in a sequence of regular arrays (i.e. the multilayer perceptron), lattice topologies whereby the artificial neurons are organized in a single regular grid (i.e. the Self-Organizing Maps), hierarchical, or even free topologies. In all cases, a so called "learning" algorithm is employed in order to infer the values of the parameters of the model from a set of observations coming from the process under study. How these parameters are stored in the model depends on the representation of the information used in the neural network (Van Gelder, 1999). In the case of models using local representations, the parameters of the model are the positions of the units in the multidimensional input space (i.e. in the Self-Organizing Maps). In the case of networks using a distributed representation, parameters are stored in the form of weighted connections between units (i.e. in the multilayer perceptron).

These computational models can be used to better understand the phenomenon under study, generally employing non-supervised learning approaches (Moshou et al., 2004; Boishebert et al., 2006), or to predict the behavior of the process under new conditions through non-linear regressions based on supervised learning approaches (Jain, 2003; Paul and Munkvold, 2005; Miao et al., 2006).

In this paper, these "algorithmic based" models were developed both to build a non-linear regression through a multilayer perceptron and also to establish clusters of events with Homogeneous Environmental Conditions (HECs) by means of a non-supervised approach known as Self-Organizing Maps.

The "data model based" techniques were employed in order to construct a linear robust regression and to determine the effects of location, management and the groups of Homogeneous Environmental Conditions with mixed models in an iterative approach guided by expert opinion.

2.4.1. Robust linear regression

For the multiple regression, we selected the robust linear regression, which exploits as much information as possible without removing outliers, exceptional records or events. This technique is appropriate when there are data points that have very high leverage (a measure of how far an independent variable deviates from its mean), and when there are outliers. Robust regression is essentially a compromise between dropping the case(s) that are moderate outliers (observations with large residuals) and seriously violating the assumptions of Ordinary Least Squares regression (OLS). The robust regression, a form of OLS, was applied to the 254 observations, with production as the dependent variable. The robust regression was set to determine Cook's D values, and then drop any observation with a Cook's D value greater than 1 in an iterative process (StataCorp, 2005; Castelló-Climent, 2008).

2.4.2. Multilayer perceptron (MLP) regression

For the non-linear regression, a supervised ANN capable of handling a high degree of heterogeneity in the data was used. ANNs, unlike OLS regressions are non-parametric and make no assumptions about the structure of the variance in the original data sets (Nagendra and Khare, 2006). A Multilayer Perceptron (MLP) (Bishop, 1995), was implemented to make a non-linear regression. This supervised algorithm is a network of individual units called perceptrons, which are linked by weighted connections, and where data move through several layers (typically three), the input, hidden, and output layers. The input and output layers contain nodes that correspond to independent and dependent variables, respectively (Kaul et al., 2005). These basic perceptrons perform a simple calculation, which in our case was a sigmoidal function of the weighted sum of the values fed in as inputs. The parameters of the model, which are the connections between units, were calculated by means of the Back-propagation algorithm applied over the data sets containing yield information (Bishop, 1995). The Back-propagation algorithm is a gradient descent that minimizes the difference between the desired output of the model (in the training data set) and the actual output of the network, i.e. the mean square error (MSE) (Bishop, 1995).

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 1–10 hidden units. This comparison was carried out by a bootstrap validation scheme (Efron, 1983). 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 (0.041) over the validation subset had four units in the hidden layer and was chosen as the most suitable (Fig. 1). One hundred networks with the selected topology were built and tested in order to improve the generalization capabilities of the model (Dietterich, 2000; Brown et al., 2005).

2.4.3. Iterative model approach

The iterative approach combined "data model based" and "algorithmic based" models and was based on robust linear regressions, ANNs non-linear regression, and a combination of a non-supervised ANNs known as Self-Organizing Maps (SOM) with mixed models with Best Linear Unbiased Prediction. The iterative approach first identified the most important variables associated with yield, second used this information to identify Homogeneous Environmental Conditions and third analyzed differences in productivity related to variation between HECs and due to management variation within HECs.

2.4.4. Self-Organizing Maps (SOM)

The Self-Organizing Map (SOM) or Kohonen map (Kohonen, 1995), is a non-supervised "algorithmic based" model widely used to obtain a visual exploratory analysis of high-dimensional data sets. SOM topology consists of a lattice of fully interconnected artificial neurons. These neurons differ from perceptron units in that the information is represented in a local manner, instead of a distributed manner. SOM units store a vector of position in the multidimensional input space, while keeping neighbor connections within the lattice of neurons which often has a low dimensionality (two in almost every application).

The SOMs were trained through an iterative process where the data set was presented to the artificial neurons. At each iteration the information stored in each artificial neuron and its neighborhood is adjusted to match the examples provided. At the end of the process, the artificial neurons become prototypes (also called prototype vectors) that summarize the data set used during training. The SOM were used to map high-dimensional data sets in a lattice of two dimensions. Observations with similar characteristics, in the high-dimensional space appear grouped together in the two dimensional map.

Such a map facilitates exploratory visual analysis of clusters and the relationships between the variables of a complex data set. 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 events 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; Vesanto and Alhoniemi, 2000).

The SOM were used to define Homogeneous Environmental Conditions (HEC) based on the original set of selected environmental variables and then on those identified as important by the robust regressions and the ANN non-linear regressions. The HECs take into account both temporal and spatial variability, thus a particular farm may fall into different HECs according to changes in the weather conditions.

2.4.5. Mixed models

Mixed models were selected as they include both, random and fixed effects in the analysis. Best Linear Unbiased Prediction (BLUP) is used in linear mixed models for the prediction of random effects. (The term prediction is normally used for the estimation of random effects, whereas estimation is used for fixed effects (Robinson, 1991; Rabe-Hesketh and Skrondal, 2008)). Furthermore, whereas regression techniques are not well suited to handle data sets with many categorical variables as a result of the exponential increase

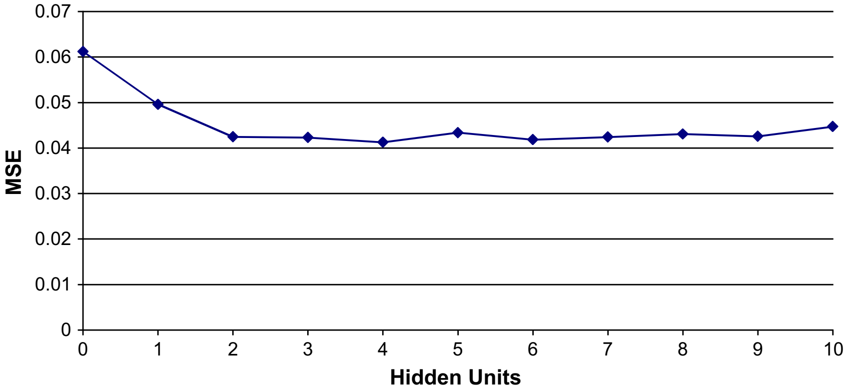

Fig. 1. MSE of artificial neural metworks with different number of neurons in hidden layer.

in volume linked to added extra dimensions to a mathematical space (Bellman, 1961; O'Grady and Medoff, 1988), mixed models are well suited to do this task. BLUPs are estimates of the realized values of the output as linear functions of the random variables; are unbiased in the sense that the average value of the estimate is equal to the average value of the quantity being estimated; best in the sense that they have minimum mean squared error within the class of linear unbiased estimators; and predictors to distinguish them from estimators of fixed effects. The mixed models assumed linear effects of the random variables with no interactions to estimate how random effects contribute to raising or lowering of the average of the output. Mixed models were selected as particularly suitable for evaluating data sets that included the categorical variables HEC, locations and farms.

2.5. Regression model testing

In order to provide a mechanism for testing the model performance and to compare different models or network topologies, both training and validation data set were created by random sampling without replacement from the whole data set for both robust regressions and MLP. In this way, each robust regression or MLP model regression was performed using 80% of the whole data set, the model performance was assessed on the remaining 20%. This method, called "split-sample" or "hold-out" validation, to assess predictive model performance, is not recommended in its simplest form for small data sets (Goutte, 1997). However, the split-sample procedure can be improved for small data set by repeating the "split-sample" procedure several times. This split sample procedure was run 100 times for both the MLP model and the robust regression model. The 100 yield estimates were then used to estimate the coefficient of determination (R^2) and the confidence limits of both the MLP and the robust regression models in order to compare the two approaches.

3. Results and discussion

3.1. Regressions

3.1.1. Selection of variables

In the iterative process of analysis, the input data sets were first pre-processed in order to eliminate variables that were highly correlated. Removal of essentially duplicated variables eliminates redundant inputs, reduces noise, and avoids the effect of several variables having the same function in the model (Faraway, 2002; Paul and Munkvold, 2005; Satizábal et al., 2007). The elimination of variables has been shown to help avoiding erroneous assignation of importance to variables when a sensitivity analysis is applied to the multilayer perceptron when analyzing data on fruit crops (Jiménez et al., 2009).

A Pearson correlation identified several variables as highly correlated: A Pearson coefficient greater than 0.8 or less than −0.8 was taken as threshold, and one of the pair of variables was eliminated from the subsequent analysis when the coefficient was beyond the threshold values (Table 2).

The decision of which variable to retain was made on the basis of expert opinion. In the case of Nariño-cartago-san isidro, thorn or no thorn (variety), the Nariño-cartago-san isidro categories for location were retained as the use of the thornless variety was considered to be just one of the several management factors that might be associated with that particular location. External drainage was chosen over internal drainage, however, in this case they are considered to be totally interchangeable. The variable average temperature for the harvest month, first month before harvest and second month before harvest were strongly correlated: the variable average temperature of the harvest month was retained. Likewise, the accumulated precipitation of the harvest month, the first and second months before harvest was strongly correlated with the temperature range throughout the different months; the variable temperature range was maintained instead of accumulated precipitation. After the elimination process twelve of the initial 19 variables were selected as drivers for the MLP non-linear regression and robust linear regressions. They were: Nariño- cartago-san isidro, Nariño- la union- buenos aires, Nariño- la union- la Jacoba, Nariño- la union- chical alto, altitude, slope, external drainage, effective soil depth, precipitable water of the harvest month, precipitable water of the first month before harvest, precipitable water of the second month before harvest, average temperature of the harvest month, temperature range of the harvest month, temperature range of the first month before harvest, and temperature range of the second month before harvest (Table 1).

3.1.2. Performance analysis and variables relevance

The mean R^2 from the 100 validations subsets was 0.69 for the MLP and 0.65 for the robust regression model (Table 3). The distribution of the R^2 provided by each approach was similar (Fig. 2) with a 95% confidence interval 0.67–0.70 for the MLP regression and 0.63–0.66 for the robust regression. Both models explained more than 60% of variability in production at $P = 0.05$. The R^2 of the MLP was significantly greater than that of the robust linear regression ($P < 0.05$ Holm–Sidak comparison at an alpha level of 5%) and thus the MLP explained significantly more of the variation (69%) than the robust regression (65%).

One of the steps followed to develop the robust regression, included the computation of forward stepwise addition procedure (Tomassone et al., 1983). This method was used to add step by step one predictor and assess the change in the MSE of the model. The change in the MSE associated with the addition of each variable illustrates the relative importance of each predictor variable (Gevrey et al., 2003). This stepwise procedure indicated that the variable slope explained 84% of the total yield variation, average temperature of the harvest month 11% and effective soil depth 4% of the total variation (Table 4).

In the case of MLP, in order to identify the variables which contribute most to yield, we used the relevance metric based on sensitivity described by Satizábal and Pérez-Uribe (2007). This assesses input relevance by calculating the partial derivative of the output of the neural network with respect to each one of the

Table 2
Pairs of variables strongly correlated.

Variable retained (abbreviation)[a]	Variable removed (abbreviation)[a]	Correlation
Na_ca_san	Nar_Thorn_N	−1
ExtDrain	IntDraina	−1
TempAvg_0	TempAvg_1	0.94
TempAvg_0	TempAvg_2	0.83
TempRang_0	PrecAcc_0	−0.83
TempRang_0	PrecAcc_1	−0.88
TempRang_1	PrecAcc_2	−0.89
TempRang_2	PrecAcc_2	−0.82

[a] List of abbreviations and their meanings are shown in Table 1.

Table 3
R^2 of predicted versus real lulo yield provided by both regressions, using 100 validation data sets.

Approach	Regression	R^2 (mean)	Confidence interval (95%)
"Data model based"	Robust (linear)	0.65	0.63–0.66
"Algorithmic based"	MLP (non-linear)	0.69	0.67–0.70

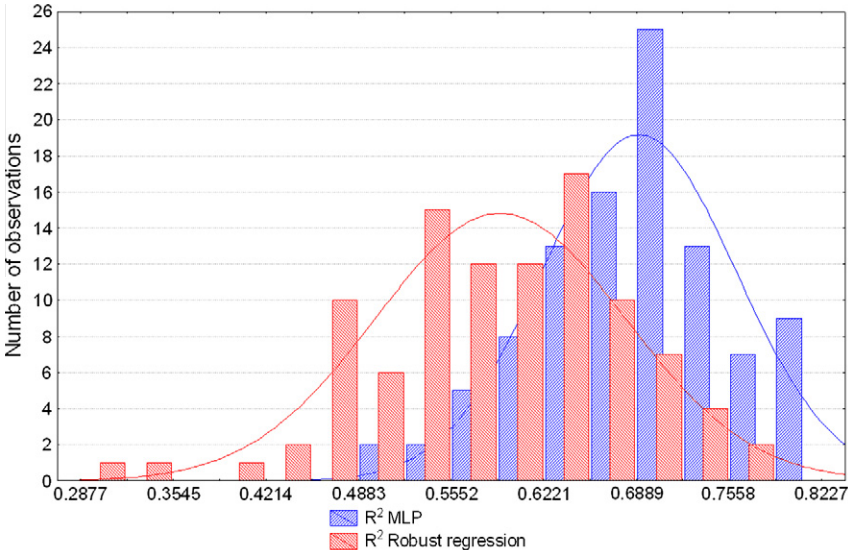

Fig. 2. Distribution of the R^2 obtained with each model.

inputs, thus the greater the partial derivate, the more relevant is the variable because the Back-propagation algorithm gives higher values to the connection weights of those inputs that are more relevant. The sensitivity metric in the MLPs identified effective soil depth, average temperature of the harvest month, slope and the locality Nariño-la union-chical alto as the most important variables associated with yield variation (Fig. 3).

The four variables selected by the sensitivity metric included the three most important variables as determined by the robust linear regression. With the exception of slope, these are the same variables that were identified as most relevant for modeling Andean blackberry yield in Colombia (Jiménez et al., 2009).

3.2. Mixed models and Self-Organizing Maps

In various crops attempts have been made to define the major environments where crops are grown and the homogeneous or mega-environments in which similar varieties or crops can be grown (see for example Cock, 1985; Braun et al., 1996; Yan et al., 2002). These relatively Homogeneous Environmental Conditions or mega-environments have been determined both by expert opinion, (see for example Cock (1985) for cassava, Braun et al. (1996) for wheat) and by analysis of the differential response or ranking of varieties in multi-locational variety trials (Yan et al., 2002.). Isaacs et al. (2007) defined various agro-ecological zones (AEZ) for sugarcane production so as to analyze the effects of management practices on cane and sugar yield within and across AEZs using commercial production data. In the case of sugarcane the AEZs were based on expert opinion and an intimate knowledge of the crop and its response to variation in climate and soil conditions. The idea behind the HECs, AEZs and mega-environments is that the crops response in any one HEC, AEZ or mega-environments is uniform or homogeneous. With lulo we were not able to define AEZs in the same manner as with sugarcane, cassava and wheat as there was neither sufficient expert knowledge of the crops response to variation in soil and climatic conditions nor were there carefully managed and controlled multilocational trials. Hence we explored an iterative approach to defining HECs for lulo.

Table 4
Variables explaining lulo yield according to a forward stepwise procedure.

Variable added	R^2	R^2 due to variable	% of Total
Slope	0.47	0.47	84.3
TempAvg_0	0.53	0.06	11.0
EffDepth	0.55	0.02	3.7
Total		0.55	100.0

The first step to identify production conditions that were homogeneous in terms of the environment and the weather in the period before harvest was to select the twelve variables identified by the regression models as those most closely associated with variation in productivity. The twelve variables were then used to train a

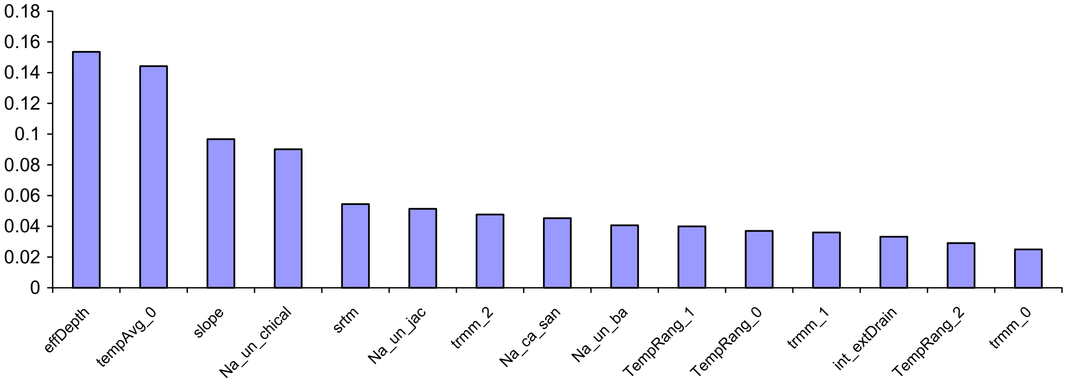

Fig. 3. Sensitivity distribution of the MLP model with respect to the inputs.

Please cite this article in press as: Jiménez, D., et al. Interpretation of commercial production information: A case study of lulo (*Solanum quitoense*), an under-researched Andean fruit. Agr. Syst. (2010), doi:10.1016/j.agsy.2010.10.004

Kohonen map and identify clusters of HEC (Fig. 4a). The Davies-Bouldin index (Davies and Bouldin, 1979) indicated that there were six major HECs (Fig. 4b).

These six HECs were then incorporated into a mixed model together with the variables farm and location (Table 5). Farm and location were both incorporated as proxy variables for crop management on the assumption that HECs covered the variation due to environmental variation and that the remaining variation must be due to management. Furthermore, we hypothesized that management in any one geographical location might be similar due to the interchange of ideas between farmers and furthermore that even in the same location there would undoubtedly be managerial differences between farms. In previous studies, the variable location or site has been incorporated into regression models to predict soybean and winter wheat yield (Yan and Rajcan, 2003; Green et al., 2007).

The mixed model with six HECs, location and farm explained more than 79% of the variation (Table 6). However, the single variable farm explained 70% of the variation, the location 8% and the HECs a negligible amount (less than 1%). Based on the MLP regressions and the robust regressions, in which environmental variables explained more than 60% of the variation with 95% confidence limits, we had expected HECs to explain a much larger proportion of the variation. This suggested that the variable, farm, was not only acting as a proxy for management effects but also for environmental conditions and that the clustering process had not identified truly homogeneous ecological conditions for crop growth and development. The most likely explanation for the HECs not being truly homogeneous in respect to crop response was that the variables used to develop the clusters were inappropriate with the variation encountered in several variables being irrelevant in terms of crop development. From the MLP and the robust regression analysis, soil depth, average temperature of the harvest month, and slope were identified as the most important environmental variables associated with variation in yield. Expert opinion concurred with the premise that soil depth and temperature were indeed likely to be important factors associated with production. However, slope came as somewhat of a surprise to the experts, although it is well known that most lulo is indeed grown on sloping ground with lulo planted on flat lands being a rarity. We therefore conducted a new cluster analysis with the three most important environmental factors identified by the regressions using the same Kohonen map procedure as that used previously: three HECs were identified (Fig. 5a and b).

A mixed model with the categorical variables of three HECs, location and farmer explained more than 80% of the variation in lulo yield (Table 6). The variable HEC explained 61% of the total variation indicating the extreme importance of environmental conditions in yield determination. The location explained less than 3% of the variation in yield suggesting that differences in local practices between locations are of little importance in determining yield. On the other hand 19% of the variation in yield, *ceteris paribus*, was attributed to the farm suggesting that the

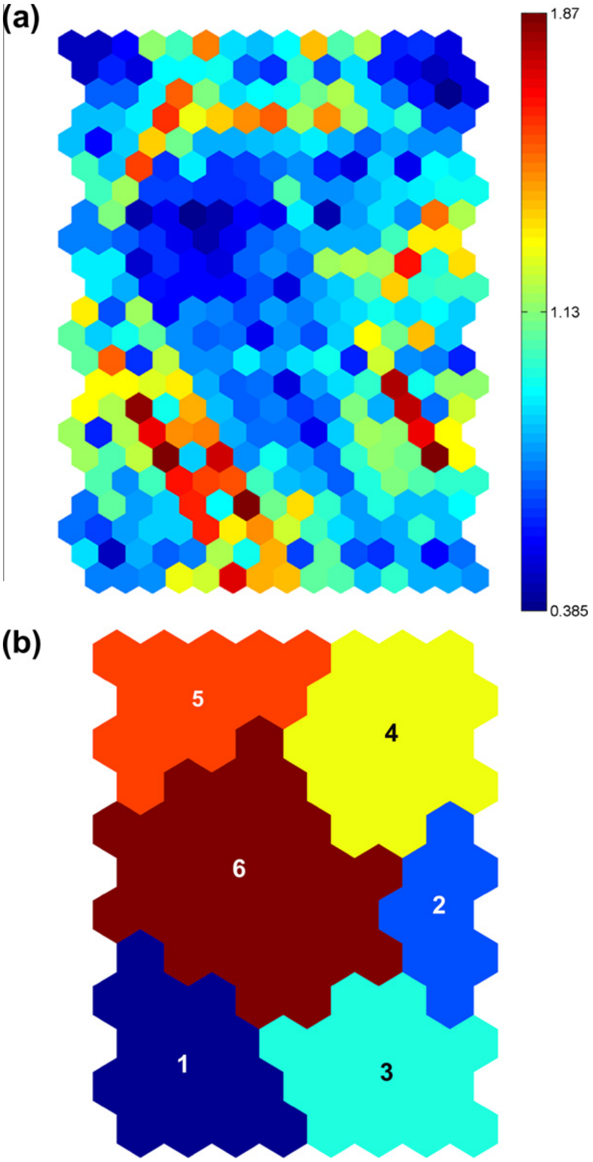

Fig. 4. (a) U-matrix displaying the distance among prototypes. The scale bar (right) indicates the values of distance. The upper side exhibits high distances, whilst the lower displays low distances. (b) Kohonen map displaying the six clusters obtained by the K-means algorithm and the Davies–Bouldin index.

Table 5
Variables integrated into the mixed model.

Variable	Abbreviation
Biophysical data used in regressions[a,b]	See Table 1
Site-farm[a]	F1, F2, F3, F4, F5, F6, F7, F8, F9, F10,... F21
Homogeneous Environmental Conditions[a]	HEC1, HEC2, HEC3...HECn
Location[a]	Na_ca_san, Na_un_ba, Na_un_jac, Na_un_chical

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

Table 6
Variance components of the mixed model estimations.

Parameters	Estimate ($g\ plant^{-1}\ wk^{-1}$)	Standard error	% of Total variance
Model including information of 6 HECs, farms, and 12 biophysical variables			
HEC	0.01	0.65	0.5
Location	0.18	0.62	8.4
Site-farm	1.50	0.34	70.4
Error	0.44	0.05	20.7
Total	2.13		100.0
Model including categorical variables of 3 HECs, location and farm			
HEC	1.85	2.01	61.2
Location	0.07	0.20	2.5
Site-farm	0.57	0.21	19.0
Error	0.52	0.04	17.3
Total	3.03		100.0

management skills of the individual farmers influenced yield. Furthermore, the high level of explanation of the total variance by the HECs suggests that the means used to define them is effective.

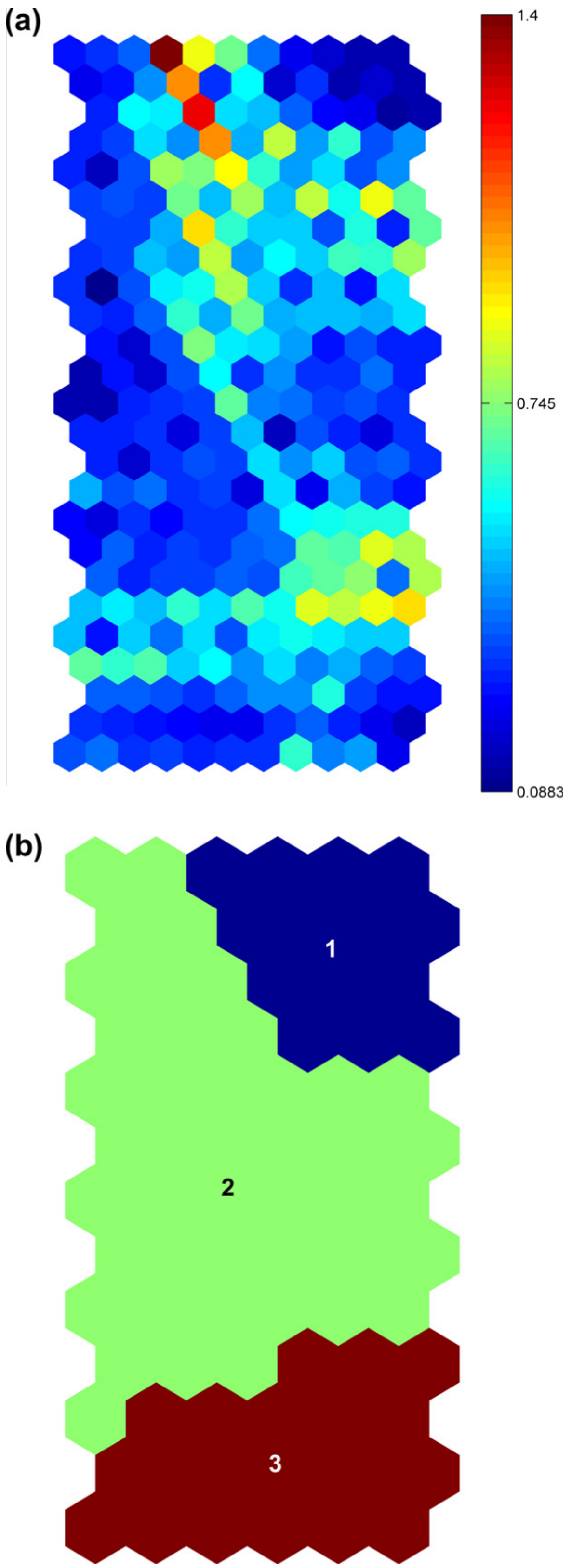

Fig. 5. (a) U-matrix displaying the distance among prototypes. The scale bar (right) indicates the values of distance. The upper side exhibits high distances, whilst the lower displays low distances. (b) Kohonen map displaying the three clusters obtained after using the *K*-means algorithm and the Davies–Bouldin index.

In the initial selection of variables, the varietal trait thorn or no thorn was eliminated as it was highly correlated with location and effectively confounded with location. Nevertheless, for farmers the effect of this trait on yield is extremely important: thornless varieties (*Solanum quitoense* var. *quitoense*) are much easier to harvest than thorny types (*Solanum quitoense* var. *septentrionale).* As location was only minimally associated with variation in yield once the effects of HEC and farm were taken into account, we decided to run the mixed model without location, but including the thorn trait as a fixed effect.

The variation explained by both HEC and farm (79%) was similar to that of the previous model (Table 8). The effect of the variable thorn or no thorn was not significant at the standard 5% level ($P = 0.168$) (Table 7). However, we suggest that caution is needed in interpreting this result as indicating that there is no difference between the yield of thorned and thornless varieties. Put in another way, there is a 7:1 probability that thorned varieties yield 28 g plant^{-1} wk^{-1} -more than thornless plants. For the farmer this is an important and commercially significant difference. This indicates that it would be advisable for producers who currently use thornless varieties to compare thorned and thornless varieties on their farms to elucidate whether the probable lack of yield is compensated for by their ease of harvest.

Inspection of Fig. 6a and b gave clues as to how HEC and farms affect productivity of lulo. HEC 3 shows a significant effect on lulo yield and consistently yielded more than HEC 2 and HEC 1 (Table 10). HEC 3 yielded 41 g plant^{-1} wk^{-1} more fruit than average, whilst HEC 2 yielded 18 g plant^{-1} wk^{-1} less than average and HEC 1 yielded 24 g plant^{-1} wk^{-1} less than the average. Comparison of the characteristics of HEC 3 with the other HECs provides an indication of environmental conditions suitable for high productivity of lulo (Table 9).

Farms 5, 6, 16, 19, and 20 in HEC 2 and farms 7 and 9 in HECs had significantly different yields to the mean. A particular farm may fall into different HECs according to changes in the environmental conditions. Thus, farms 19 and 20 had a significant effect on lulo production when they fell into HEC 2, but not when they fell into HEC 3. Nevertheless, farms 19 and 20 produced 15 and 38 g plant^{-1} wk^{-1} more than average in HEC2 and 15 and 17 g plant^{-1} wk^{-1} more than average in HEC3, suggesting that these farms effectively manage their crops and that different environmental conditions do not greatly affect good management practices required to obtain higher than average yield. Farm 7

Table 7
Variance components of the mixed model estimations, including information of variety. Fixed effect.

Lulo yield	Coefficient ($\text{g plant}^{-1}\ \text{wk}^{-1}$)	Standard error	Z	$P > Z$
Fixed effect				
Nar_Thorn_N[a]	−27.69	20.1	−1.38	0.168[n]

[a] Variable defined in Table 1.
[n] Not statistically significant difference.

Table 8
Variance components of the mixed model estimations, including information of variety (Nar_thorn_N). Random effects.

Parameters	Estimate ($\text{g plant}^{-1}\ \text{wk}^{-1}$)	Standard error	% Variation of total
Random effects			
HEC	1.41	1.55	55.4
Site-farm	0.61	0.21	23.9
Error	0.53	0.05	20.8
Total	2.56		100.0

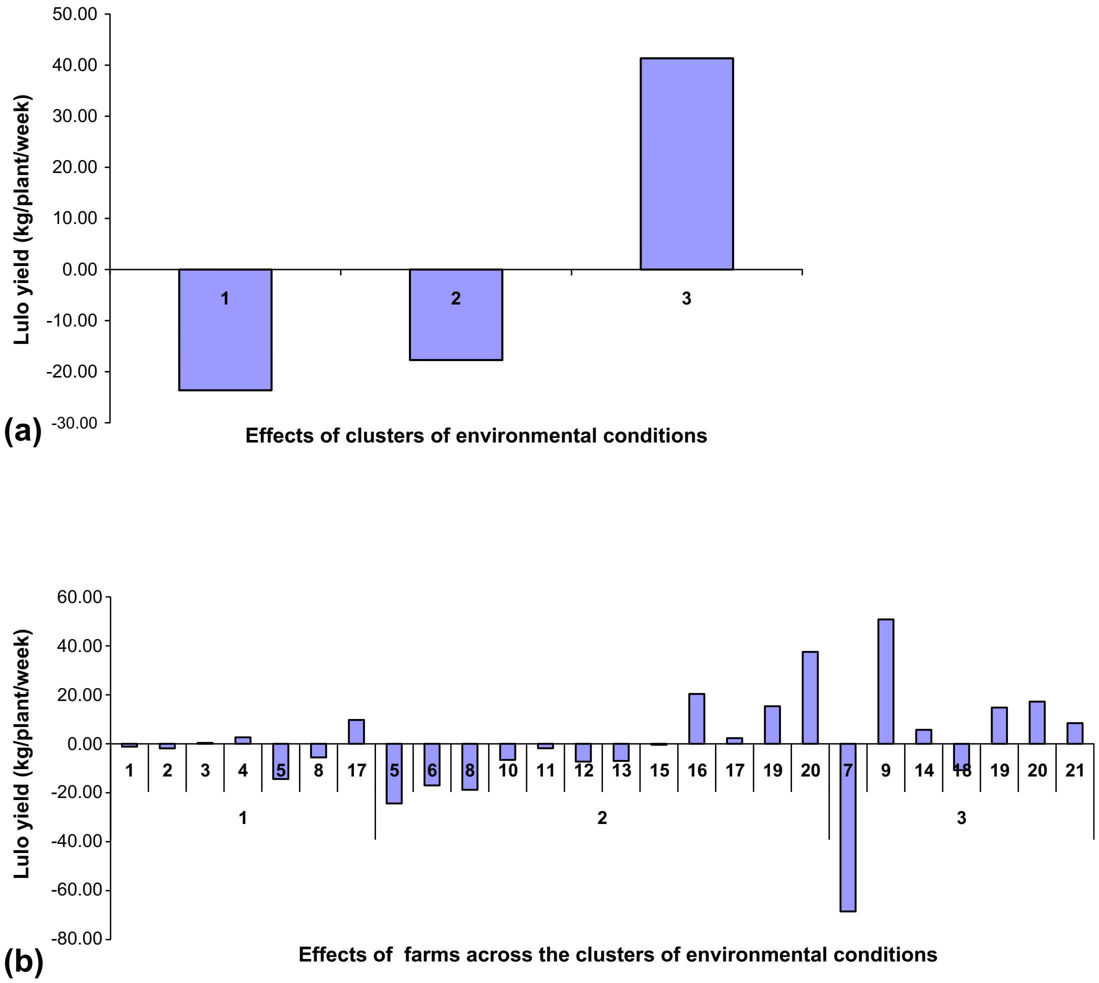

Fig. 6. Clustered columns of the effects on lulo yield estimations: (a) effect of HEC, (b) effects of farms across the HEC.

and 9 are in HEC 3 which in general produces the highest yields. However, farm 7 produced 68 g plant^{-1} wk^{-1} less than average whilst farm 9 produced 51 g plant^{-1} wk^{-1} more than average. Similarly, farm 16 even though it was in a relatively low productivity environment (HEC 2) produced significantly more, (20 g plant^{-1} wk$^{-1)}$, than average. We suggest that farm 7 probably has inappropriate management practices for obtaining high yields whilst farms 9 and 16 are effectively managed. Furthermore, by identifying well managed farms and poorly managed units in particular environmental niches and visiting them it should be possible to identify those management practices that are associated with high levels of productivity and conversely those practices which are inappropriate. We suggest that this information is extremely valuable as visits to superior farms could provide guidelines for improving yields on other farms with similar HECs.

Within HECs there is a large range of variation of yield associated with the farm, and little variation associated with location (Fig. 6b). Proxies can be used to estimate the effect of immeasurable 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). Jiménez et al. (2009) used geographic location of areas as proxies for crop management practices for Andean blackberry, suggesting that local knowledge and socio-economic circumstances would tend to be similar within geographic locations and would differ between them. In the case of Andean blackberry the variable geographic location was associated with yield variation; however, it is noteworthy that the geographic separation in the Andean blackberry study was much greater than in this study with lulo, in which location did not appear to be an important determinant of yield.

On the other hand, we suggest that farms as a variable, within homogeneous ecological conditions, provides a proxy for farmer's management skills. Although it is not possible to identify precisely what the practices or skills used by the farmers it is possible to identify "good" farmers and quantify the yield advantage that they obtain over others.

4. Conclusions

Both "Data model based" and "algorithmic based" models that used commercial production data linked to characterization of the growing conditions explained more than 60% of variability in lulo yield. The algorithmic model, using a multilayer perceptron explained more of the variation (69%) than the data model based robust regression (65%). The robust regression applied with a stepwise procedure identified slope, average temperature and soil depth as the most important environmental variables associated with variation in yield. The sensitivity analysis of the multilayer perceptron identified the same three factors as the stepwise robust regression plus one locality based variable, suggesting that both methods are appropriate to identify the most important factors associated with yield variation, but that the MLP was capable of

Please cite this article in press as: Jiménez, D., et al. Interpretation of commercial production information: A case study of lulo (*Solanum quitoense*), an under-researched Andean fruit. Agr. Syst. (2010), doi:10.1016/j.agsy.2010.10.004

discovering factors that were not identified as important by the robust regression.

Identification of HECs by taking all the measured variables and using Self-Organizing Maps (SOM), which are ''algorithmic based'', did not provide a useful clustering of HECs. However, by first identifying those factors associated with yield variation either by robust regression or by multilayer perceptron regressions, HECs associated with yield variation were successfully defined. Once the HECs were defined it was possible use a mixed model to analyze: (1) the effects of the environment using the HECs as a categorical variable; (2) cultural conditions associated with the geographic position of the individual production units using location as a categorical variable; and (3) farm management skills using the farm as a categorical variable. The mixed model has the advantage over regression models of handling multiple categorical variables of the same class, such as several farms.

The mixed model explained more than 80% of the total variation in lulo yield, with HEC and farm variables explaining most of the variation. This suggests in the case of lulo that better than average yield is primarily associated with appropriate environmental conditions (indicated by HEC) and good farm management practices (indicated by farm). Although it was not possible to identify precisely which management practices were effective, those farms with ''good'' management could readily be identified. Furthermore, observation of the range of conditions in the HEC 3, associated with higher than average yields defined that the most suitable environmental conditions for producing lulo are the combination of: an effective soil depth between 40 and 67 cm, slope between 13 and 24 degrees and an average temperature between 15.8 and 19 degree Celsius (°C). It is also noteworthy that although in this data set not all measured variables were associated with variation in lulo yield, those variables may affect yield if they are outside the range reported here. We note that an automated approach to analyzing the data using a single methodology was much less powerful than an iterative guided approach using various methodologies. Both ''data model based'' and ''algorithmic based'' models are useful tools for analyzing and interpreting the commercial production data once highly correlated independent variables had been eliminated. Regression models were particularly effective at identifying those independent variables associated with variation in the dependent variable, yield and then for defining Homogeneous Environmental Conditions (HECs) based on the previously identified independent variables. Mixed models were effective for quantifying the effects of location (local culture) and farm (farm management skills) once the HECs had been determined. The mixed model has the advantage of effectively handling multiple categorical variables. Thus we suggest that when analyzing commercial production highly correlated independent variables should first be eliminated, then either algorithmic or data based regression models should be used to identify those independent variables associated with variation in the dependent variable. Self-organizing maps can then be used to determine HECs based on the variables associated with yield. Mixed models can then be used to analyze the variation in yield attributable to both the environmental conditions (HECs) and the social and management conditions. In the particular case of lulo, proxies were used for social and management conditions. Finally, the experience with lulo suggests that useful information on where and how to grow crops successfully can be obtained from analysis and interpretation of commercial production data combined with information on site specific growing conditions. The analysis and interpretation of the data is not trivial: expert guidance is required in the process of analysis. Nevertheless, various essential principles have been established that can be used as a guide to analysis and interpretation of commercial production data, especially for crops with little formal experimental trialing.

Table 9
Environmental conditions for each HEC.

	Variable ranges		HEC
Slope (°)	EffDepth (cm)	TempAvg_0 (°C)	
5–14	21–40	15–16.5	1
8–15	32–69	15–18.9	2
13–24	40–67	15.8–19	3

Table 10
t Test for the best linear unbiased predictions (BLUPs).

Effect		Estimate (g plant^{-1} wk^{-1})	*t*	Probability > |*t*|
HEC	Farm			
1	–	−24	−0.97	0.33[n]
2	–	−18	−0.77	0.44[n]
3	–	41	1.76	**0.08**[s]
1	1	−1	−0.08	0.93[n]
1	2	−2	−0.13	0.89[n]
1	3	0	0.03	0.98[n]
1	4	3	0.19	0.85[n]
1	5	−14	−0.86	0.39[n]
1	8	−6	−0.32	0.75[n]
1	17	10	0.59	0.55[n]
2	5	−24	−2.55	**0.01**[s]
2	6	−17	−1.78	**0.08**[s]
2	8	−19	−1.44	0.15[n]
2	10	−7	−0.7	0.48[n]
2	11	−2	−0.19	0.85[n]
2	12	−7	−0.79	0.43[n]
2	13	−7	−0.8	0.42[n]
2	15	0	−0.04	0.97[n]
2	16	20	1.99	**0.05**[s]
2	17	2	0.24	0.81[n]
2	19	15	1.71	**0.09**[s]
2	20	38	4.26	**<.0001**[s]
3	7	−68	−5.12	**<.0001**[s]
3	9	51	4.56	**<.0001**[s]
3	14	6	0.48	0.63[n]
3	18	−11	−0.97	0.33[n]
3	19	15	0.84	0.40[n]
3	20	17	0.97	0.33[n]
3	21	8	0.76	0.45[n]

[n] Not statistically significant difference.
[s] Statistically significant difference.

Acknowledgements

The data used in this analysis was generated by a cooperation project between Corporacion Biotec, International Center for Tropical Agriculture (CIAT), and the ''Haute école d'ingénierie et de gestion du canton de Vaud'' (HEIG-VD), named ''Precision agriculture and the construction of field-crop models for tropical fruits''. The project was supported by funds from several institutions in Colombia (MADR, COLCIENCIAS, ACCI) and the Rectors' Conference of the Swiss Universities of Applied Sciences (KFH), and the State Secretariat for Education and Research (SER) in Switzerland.

References

Acosta, O., Perez, A., Vaillant, F., 2009. Chemical characterization, antioxidant properties, and volatile constituents of naranjilla (*Solanum quitoense* Lam.) cultivated in Costa Rica. Archivos Latinoamericanos de Nutricion 59, 88–94.

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. International Journal of Cancer 80, 641–645.

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

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. Journal of Geophysical Research-Atmospheres 92, 9631–9643.

Bellman, R.E., 1961. Adaptive Control Processes. Princeton University Press, Princeton, NJ.

Bessant, J., Francis, D., 1999. Developing strategic continuous improvement capability. International Journal of Operations & Production Management 19, 1106–1119.

BIOTEC, 2007. Agricultura específica por sitio en frutales.Calendario para toma de información en el cultivo del Lulo.

Bioversity International, 2005. Information Sheet on Solanum quitoense in New World Fruits Database. <http://www.bioversityinternational.org/informationSources/SpeciesDatabases/New/WorldFruitsDatabase/> (accessed August 2009).

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. Chemometrics and Intelligent Laboratory Systems 80, 13–23.

Braun, H.-J., Rajaram, S., Van Ginkel, M., 1996. CIMMYT's approach to breeding for wide adaptation. Euphytica 92, 175–183.

Breiman, L., 2001. Statistical modeling: the two cultures. Statistical Science 16, 199–215.

Brown, G., Wyatt, J.L., Harris, R., Yao, X., 2005. Diversity creation methods. A survey and categorisation. Information Fusion 6, 5–20.

Castelló-Climent, A., 2008. On the distribution of education and democracy. Journal of Development Economics 87, 179–190.

Cock, J., 1985. Stability of Performance of Cassava Genotypes. In: Hershey, C.H. (Eds.), Proceeding Workshop Cassava Breeding. A Multidisciplinary Review. Los Banos, Philippines, pp. 177–206.

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. <http://biotec.univalle.edu.co/Memorias.htm>.

Cock, J., Oberthür, T., Isaacs, C., Läderach, P., Palma, A., Carbonell, J., Watts, G., Amaya, A., Collet, L., Lema, G., Anderson, E., submitted for publication. Crop Management Based on Field Observations: Case Studies in Sugarcane and Coffee. Experimental Agriculture.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 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. The error rate of a prediction rule: improvement on cross-validation. Journal American Statistical Association 78, 316–331.

Faraway, J.J., 2002. Practical Regression and Anova using R. Available from the R Project. <http://cran.r-project.org/>.

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. Population and development review. Population and Development Review 25, 85–120.

Flórez, S.L., Lasprilla, D.M., Chaves, B., Fischer, G., Magnitskiy, S., 2008. Growth of lulo (*Solanum quitoense* Lam.) plants affected by salinity and substrate. Revista Brasileira de Fruticultura 30, 402–408.

Framingham Heart Study, 2006. Framingham Heart Study. A project of the National Heart, Lung, and Blood Institute and Boston University. <http://www.nhlbi.nih.gov/about/framingham> (accessed July 2006).

Franco, G., Bernal, J.E., Giraldo, M.J., Tamayo, J., Castaño, P., Tamayo, V., Gallego, J., Leomad, J., Botero M.J., Rodríguez, J., Guevara, N., Morales, J., Londoño, M., Ríos, G., Rodríguez, J., Cardona, J., Zuleta, J., Castaño, J., Ramírez, C., 2002. El cultivo del lulo: Manual técnico Corporación Colombiana de Investigación Agropecuaria (CORPOICA), Regional nueve, CORPOICA Manizales, pp. 1–103.

Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological Modelling 160, 249–264.

Gitterle, T., Martinez, W., Marimon, F., Salazar, M., Faillace, J., Suarez, A., Cock, J., 2009. Commercial Field Performance as a Measure of Genetic Improvement in the Pacific White Shrimp Penaeus (Litopenaeus) vannamei. International Symposium of Genetics in Aquaculture. Bangkok, Thailand.

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. American Journal of Epidemiology 144, 290–299.

Goutte, C., 1997. Note on free lunches and cross-validation. Neural Computation 9, 1245–1249.

Green, T.R., Salas, J.D., Martinez, A., Erskine, R.H., 2007. Relating crop yield to topographic attributes using Spatial Analysis Neural Networks and regression. Geoderma 139, 23–37.

Grimm, V., 1999. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? Ecological Modelling 115, 129–148.

Hashimoto, Y., 1997. Special issue: applications of artificial neural networks and genetic algorithms to agricultural systems. Computers and Electronics in Agriculture 18, 71–72.

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. International Journal of Climatology 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., Cock, J., Satizábal, F., Barreto, M., Pérez-Uribe, A., Jarvis, A., Van Damme, P., 2009. 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. Computers and Electronics in Agriculture 69, 198–208.

Kannan, V.R., Tan, K.C., 2005. Just in time, total quality management, and supply chain management: understanding their linkages and impact on business performance. Omega. The International Journal of Management Science 33, 153–162.

Kaul, M., Hill, R.L., Walthall, C., 2005. Artificial neural networks for corn and soybean yield prediction. Agricultural Systems 85, 1–18.

Khakural, B.R., Robert, P.C., Huggins, D.R., 1999. Variability of corn/soybean yield and soil/landscape properties across a southern Minnesota landscape. In: Robert, P.C., Rust, R.H., Larson, W.E., (Eds.), Precision Agriculture: Proceedings of the Fourth International Conference, Madison, WI, USA, pp. 573–579.

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

Kravchenko, A.N., Bullock, D.G., 2000. Correlation of corn and soybean grain yield with topography and soil properties. Agronomy Journal 92, 75–83.

Lanzante, J.R., 1996. Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples including applications to historical radiosonde station data. International Journal of Climatology 16, 1197–1226.

Miao, Y., Mulla, D.J., Robert, P.C., 2006. Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. Precision Agriculture 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. Computers and Electronics in Agriculture 44, 173–188.

Nagendra, S.M.S., Khare, M., 2006. Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. Ecological Modelling 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.

Niederhauser, N., Oberthür, T., Kattnig, S., Cock, J., 2008. Information and its management for differentiation of agricultural products: the example of specialty. Computers and Electronics in Agriculture 61, 241–253.

O'Grady, K.E., Medoff, D.R., 1988. Categorical Variables in Multiple Regression: Some Cautions. Multivariate Behavioral Research, Society of Multivariate Experimental Psychology, Fort Worth, TX, ETATS-UNIS.

Operational Research Society, 2006. <http://www.orsoc.org.uk/orshop/(0tic0pjmqgos3ajjs1zaww55)/orhomepage2.aspx> (accessed July 2006).

Osorio, C., Duque, C., Batista-Viera, F., 2003. Studies on aroma generation in lulo (*Solanum quitoense*): enzymatic hydrolysis of glycosides from leaves. Food Chemistry 81, 333–340.

Ozesmi, S.L., Tan, C.O., Ozesmi, U., 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. Ecological Modelling 195, 83–93.

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.

Piepho, H.P., 1994. Best Linear Unbiased Prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. Theoretical and Applied Genetics 89, 647–654.

Piepho, H.P., Mohring, J., 2005. Best linear unbiased prediction of cultivar effects for subdivided target regions. Crop Science 45, 1151–1159.

Pulido, S., Bojacá, C.R., Salazar, M., Chaves, B., 2008. Node appearance model for Lulo (*Solanum quitoense* Lam.) in the high altitude tropics. Biosystems Engineering 101, 383–387.

Rabe-Hesketh, S., Skrondal, A., 2008. Multilevel and Longitudinal Modeling Using Stata, second ed. Stata Press, College Station, Texas. pp. 156–160.

Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. Statistical Science 6, 15–32.

Rousseeuw, P., Leroy, A., 1987. Robust Regression and Outlier Detection.

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

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

Satizábal, H.F., Jiménez, D.R., Pérez-Uribe, A., 2007. Consequences of data uncertainty and data precision in artificial neural network sugar cane yield prediction. In: Proceedings of the 9th International Work-Conference on Artificial Neural Networks, pp. 1147–1154.

Schultz, A., Wieland, R., 1997. The use of neural networks in agroecological modelling. Computers and Electronics in Agriculture 18, 73–90.

Please cite this article in press as: Jiménez, D., et al. Interpretation of commercial production information: A case study of lulo (*Solanum quitoense*), an under-researched Andean fruit. Agr. Syst. (2010), doi:10.1016/j.agsy.2010.10.004

Schultz, A., Wieland, R., Lutze, G., 2000. Neural networks in agroecological modelling-stylish application or helpful tool? Computers and Electronics in Agriculture 29, 73–97.

StataCorp, 2005. Stata Reference Manual: Release 9. Stata Data Analysis Examples: Robust Regression, Stata Press, Texas, USA.

Steckel, R.H., 1995. Stature and standard of living. Journal of Economic Literature 33, 1903–1940.

Thomas, D., Strauss, J., Henriques, M., 1990. Child survival, height for age and household characteristics in Brazil. Journal of Development Economics 33, 197–234.

Tomassone, R., Lesquoy, E., Miller, C., 1983. La regression: nouveaux regards sur une ancienne methode statistique, Paris.

Van Gelder, T., 1999. Distributed vs. Local Representations. The MIT Encyclopedia of Cognitive. In: Wilson, R., Keil, F. (Eds.), The MIT Encyclopedia of Cognitive Sciences. MIT Press, Cambridge, pp. 236–238.

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

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11, 568–600.

Yan, W., Rajcan, I., 2003. Prediction of cultivar performance based on single- versus multiple-year tests in soybean. Crop Science 43, 549–555.

Yan, W., Hunt, L.A., Johnson, P., Stewart, G., Lu, X., 2002. On-farm strip trials vs. replicated performance trials for cultivar. Crop Science 42, 385–392.