Promotor: Prof. dr. ir. Patrick Van Damme
Faculty of Bioscience Engineering

Rector: Prof. dr. Paul. Van. Cauwenberge
Dean: Prof. dr. ir. Guido. Van Huylenbroeck

DANIEL RICARDO JIMÉNEZ RODAS

SITE-SPECIFIC CROP PRODUCTION BASED ON FARMERS' PRODUCTION EXPERIENCES IN COLOMBIA. CASE STUDIES ON ANDEAN BLACKBERRY (*Rubus glaucus* Benth) AND LULO (*Solanum quitoense* Lam)

Thesis submitted in fulfilment of the requirement
for the degree of Doctor (PhD) in Applied Biological Sciences: Agricultural Science

Dutch translation of the title:

PLAATSSPECIFIEKE TEELT OP BASIS VAN ERVARINGEN VAN COLOMBIAANSE BOEREN. GEVALSTUDIES VAN BRAAMBES (*Rubus glaucus* Benth) EN LULO (*Solanum quitoense* Lam)

Photograph back cover: A lulo farmer with his crop of fruit growing in Darién, Colombia. Picture by Neil Palmer (CIAT).

Cover design: Corporate communications (CIAT)

Correct Citation:

Jiménez, D. 2013. Site-specific crop production based on farmers' production experiences in Colombia. Case studies on Andean blackberry (*Rubus glaucus* Benth) and lulo (*Solanum quitoense* Lam). PhD-thesis. Faculty of Bioscience Engineering, Ghent University, Belgium, 209 pp.

ISBN-number: 978-90-5989-606-2

Acknowledgements

The research was accomplished with the support of many people, but mentioning them all would be impossible.

I am particular grateful to James Cock. He was the one who brought the idea of site-specific agriculture in many fields of research in Colombia, and I had the privilege to be selected by him to work in a thesis focused on methodologies for analysing data in this field of research. As scientist he gave me valuable comments elements, encouragements and criticism to write this thesis and continue with my career as a researcher.

I am very grateful to Andy Jarvis for encouraging me to start this Ph.D, working with/for him has been an honour. He has trusted in my capabilities and gave me the guidelines to continue with my research; he also provides me not only with scientific suggestions to make progress in my professional career but also with friendly advices which are inappreciable.

During my years of research assistant I had the opportunity of learning from others researchers and experts on tropical species and modelling, I want to express my gratitude to Geo Coppens and Xavier Scheldeman for initiating me into the world of research, I want to thank specially Xavier for supporting me during the development of this PhD, I am very grateful to him and his family.

As the present study was based on five years of work in Switzerland, France and Colombia. In Switzerland, I am indebted to the staff and colleagues at the HEIG-VD. Special thanks go to professor Andres Perez-Uribe and Eduardo Sánchez. I would also like to thank Etienne Messerli and all the REDS crew; I thank Alex Corbas, Carlos Peña, and Sébastien Gerber.

I am particularly thankful to two colleagues and friends, with whom I have been sharing office for three years during my research in COCH. As they know, brainstorming continued after work in more relaxing environments. Hector Satizabal and Miguel Barreto know that without their excellent professional advices and releasers talks, it would have been difficult to make this thesis pleasant. Thanks for your patience my dear friends!!!

I thank to the “swiss-friends cooperation network” Guillaume, Bennoit, Lisa, Damien, Parya, Sonya. They welcomed me in Switzerland and let me get involve into the Swiss culture, I am very

grateful for all the attention and care I received from them, they made me feel at home in Switzerland. I also would like to thank to my “latin brothers” in Switzerland, Luis Panichelli, Juan David Villegas, Karin Juillerat, Alejandra Peyon, Cyrille Thomas, Manuela Fernandez. All these friends in Switzerland made this period the most enjoyable of the Ph.D.

In France, specifically in *Réunion* Island, I am particular grateful to Pierre Toddorof, his team and family at CIRAD, Pierre also provide me with excellent knowledge and advices to model agricultural systems. I thank to Suzy, Mathiew, Alexandre, Sylvaine, Ismael, Elisabeth. The period I spent in reunion was one of the most fantastic moments during the development of this thesis.

I thank all the support of my supervisor Prof. dr. ir. P. Van Damme and his team at Ghent during the development of this research.

Research without funding and technical support is difficult, the data used in this thesis was generated by a cooperation project between *Corporacion Biotec*, *CIAT*,*CENICAÑA* and the 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 KFH, and the *State Secretariat for Education and Research* (SER) in Switzerland.

I am very grateful to the team in Colombia with whom I shared great moments collecting, storing and interpreting the information. I would like to thank Pedro Zapata and Diana Alvarez for sharing their agronomic knowledge, and James Garcia, and Hugo Dorado for their statistical advice. I would also like to thank to fruit growers of Andean blackberry and Lulo in Colombia that participated, their happiness, enthusiasm and kindness, despite difficult living conditions are always a source of motivation for me.

I cannot find the right words to thank my family, my parents Gilberto and Martha, my brother Cesar, my sisters Lina and Luisa. They showed me the path to find happiness in little things in life, and this has keeping me very positive and in excellent mood during most of the time in my life. I got from my family the support I needed to accomplish my studies even under difficult circumstances but more important I have had always their love. Muchas gracias!!!

Last but definitely not least, I want to express my gratitude to my wife Pilar Acosta who has supported me emotionally, professionally during these last years, this thesis would never have finished without her support, patience and love.

ACRONYMS AND ABBREVIATIONS

AEPS	Agricultura Específica Por Sitio
AEZ	Agro-Ecological Zone
ANN	Artificial Neural Network
BMU	Best Matching Unit
BLUP	Best Linear Unbiased Prediction
CB	Corporación Biotec
CENICAÑA	Centro de Investigación de la Caña de Azúcar en Colombia
CSI-CGIAR	Consortium for spatial information – Consultative Group on International Agriculture
CIAT	Centro Internacional de Agricultura Tropical
CIRAD	Centre de Coopération Internationale en Recherche Agronomique pour le Développement
CVC	Corporación Autónoma Regional del Valle del Cauca
GIS	Geographic Information Systems
GPS	Global Positioning System
HEC	Homogeneous Environmental Conditions
HEIG-VD	Haute Ecole d'Ingénierie et de Gestion du canton de Vaud
ICTs	Information and Communication Technologies
IGAC	Instituto Geográfico Augustin Codazzi
MASL	Meters Above Sea Level
MLP	Multilayer Perceptron
MSE	Mean Square Error
OLS	Ordinary Least Squares regression
PA	Precision Agriculture
RASTA	Rapid Soil and Terrain Assessment
SOM	Self-Organizing Map
SRTM	Shuttle Radar Topography Mission
SSA	Site-Specific Agriculture
SSAFE	Site-Specific Agriculture Based on Farmers Experiences
SSAFT	Site-Specific Agriculture for Tropical Fruits
SSCP	Site-Specific Crop Production
SSE	Sum of Squared Error
SSER	State Secretariat for Education and Research
TCH	Tons of sugarcane per hectare
TRMM	Tropical Rainfall Measuring Mission

ABSTRACT

Every time a farmer plants and harvests a crop, this represents a unique experiment. The premise of the present research is that, if it were possible to characterise the production system in terms of management and environmental conditions, and if information on the harvested product were collected from a large enough number of harvesting events under varying conditions, it should be possible to develop analytical approaches that would be able to adequately describe the production system. This is the approach used in operational research, in which mathematical models of observed processes identify what parts in the approach should be modified to optimize the system.

Information on cropping events of sugarcane growers in the Valle del Cauca department in Colombia and compiled by the *Centro Colombiano de Investigación de la Caña de Azúcar* (CENICAÑA), was used to develop data-driven models that were then applied to two under-researched crops in Colombia: Andean blackberry (*Rubus glaucus* Benth) and lulo (*Solanum quitoense* Lam).

The research integrates information collected from small-scale growers, publicly-available environmental databases, and mathematical techniques to develop a site-specific approach for these crops. The work takes into account farmers' production experiences over a wide range of environmental and socio-economic circumstances, which should allow us to identify the combination of factors which contribute to high productivity. Data was collected involving farmers, consistent with the principles of operational and participatory research. The organization of crops' supply chains was strongly associated with the processes of collecting, managing and analyzing information.

The modelling tools developed using the sugarcane data, explained almost 90% of the yield variation of *Rubus glaucus* and were used to (a) identify factors explaining productivity by using a relevance metric known as sensitivity matrix; and (b) visualize the relations between the relevant variables using non-supervised clustering algorithm.

With *Solanum quitoense*, an iterative procedure was used that (a) identifies the variables that best explain most yield variation; (b) clusters similar environments using a neural network; and (c) analyses the effect of environment, cultural factors associated with a geographical area, and farm management skills in a mixed model, which explained more than 80 % of yield variation.

Best conditions for *Rubus glaucus* are: an average temperature of the first month before harvest of 16–18°C, minimal effective soil depth of about 65 cm, and low rainfall during the month before harvest where drainage is poor, or moderate to low rainfall in better-drained areas. The best

conditions for *Solanum quitoense* are: an average temperature of the harvest month of 15.8–19°C, soil depth 40–67 cm, and terrain slope of 13–24°. Proxies for crop management and socio-economic circumstances were integrated in the present study, as a location effect on yield was evidenced through the modelling, suggesting the influence of variables that were not possible to capture during the data collection. For instance, there was not enough information to identify management and social factors associated with high yields, although there were farms with yields higher than the average suggesting that they were managing the crop more efficiently.

Our results offer the possibility to identify management practices used by productive farmers, and extend them to less-productive farmers, so that they too can improve their yields and implement site-specific recommendations from their colleagues.

The study shows that analysis and interpretation of farmers' production experiences, combined with data of growing conditions, can provide useful information on where and how to grow both *Rubus glaucus* and *Solanum quitoense*. Operational and participatory research methodologies and farmers' production experiences are a promising tool to develop site-specific crop production for under-researched tropical fruit species. This is especially important in Colombia because of the lack of research on crops in general, and the extremely heterogeneous growing environments.

RESUMEN

Cada vez que un agricultor siembra y cosecha un cultivo, realiza un experimento irrepetible o evento único. La premisa de la presente investigación es que, si fuera posible caracterizar el sistema de producción agrícola de los agricultores, en términos de las prácticas de manejo implementadas y las condiciones ambientales, y que si información sobre las cosechas se obtuviera a partir de un número suficiente de eventos llevados a cabo bajo diferentes condiciones; debería ser posible desarrollar enfoques analíticos con la capacidad de describir adecuadamente el sistema de producción.

Éste es el enfoque utilizado en la investigación operativa, en el cual, modelos matemáticos son desarrollados a partir de observaciones para identificar que partes del enfoque deben ser modificadas y de ésta manera optimizar el sistema. Información sobre eventos de cosecha de caña de azúcar en el Valle del Cauca en Colombia registrada por el *Centro de Investigación de la Caña de Azúcar de Colombia* (CENICAÑA), ha sido empleada para desarrollar modelos basados en los datos que han sido aplicados a dos cultivos poco investigados en Colombia: mora de los Andes (*Rubus glaucus* Benth) y lulo (*Solanum quitoense* Lam).

La investigación integra información recopilada por agricultores de pequeña escala, bases de datos meteorológicas de libre acceso, y técnicas matemáticas, con el objetivo de desarrollar un enfoque específico por sitio para estos cultivos. El estudio toma en cuenta las experiencias productivas de los agricultores, llevadas a cabo bajo diferentes condiciones ambientales y socio-económicas, las cuales nos permitirán identificar la combinación de factores que conducen a una alta productividad. Los datos fueron colectados involucrando a los agricultores, y en el proceso se tuvieron en cuenta principios de la investigación operativa y participativa. La organización de las cadenas de abastecimiento de los cultivos parece estar fuertemente ligada a los procesos de recolección, manejo y análisis de información.

Las herramientas de modelado desarrolladas a partir de información de la caña de azúcar, explicaron casi el 90% de la variación del rendimiento de *Rubus glaucus* y fueron empleadas para (a) identificar los factores que explican la productividad mediante el uso de una métrica de relevancia conocida como matriz de sensibilidad, y (b) visualizar las relaciones entre la variables más relevantes a través de un algoritmo no supervisado.

En el caso *Solanum quitoense*, un proceso iterativo se utilizó para (a) identificar los factores que mejor explican la variación en el rendimiento, (b) agrupar condiciones ambientales similares a través de una red neuronal, y (c) analizar el efecto del ambiente, factores culturales asociados a una zona

geográfica en particular, y manejo del cultivo en finca por medio de un modelo mixto, el cual explicó más del 80% la variación en rendimiento.

Las mejores condiciones para sembrar *Rubus glaucus* son las siguientes: temperatura promedio mensual entre 16 y 18 °C, un mínimo de profundidad efectiva del suelo alrededor de 65 cm y bajas precipitaciones durante el mes antes de la cosecha, en áreas donde el drenaje es pobre; o precipitaciones moderadas a bajas en áreas mejor drenadas.

En relación a *Solanum quitoense*, las mejores condiciones son: una temperatura promedio mensual entre 15 y 19 °C, profundidad efectiva del suelo entre 40 y 67 cm, pendiente del terreno entre 13 y 24 °C. Variables “proxies” para manejo del cultivo y condiciones socio-económicas fueron incluidas en el estudio, ya que un efecto de localidad fue evidenciado a través de la modelación, sugiriendo la influencia de variables que no fueron posibles de capturar durante la recolección de datos. Por ejemplo, no hubo suficiente información para identificar manejo agronómico del cultivo y factores sociales asociados con altas producciones, a pesar de que hubo fincas con rendimientos superiores al promedio, indicando que fueron más eficientes en el manejo agronómico del cultivo.

Nuestros resultados abren la posibilidad de identificar las prácticas de manejo implementadas por los agricultores más productivos, y extenderlas a otros agricultores menos productivos para que ellos también puedan aumentar sus productividades y de ésta manera poner en práctica recomendaciones específicas por sitio a partir del conocimiento de las experiencias de sus colegas.

El estudio muestra que el análisis y la interpretación de las experiencias productivas de los agricultores, combinado con información de las condiciones de desarrollo del cultivo, pueden proporcionar información útil sobre dónde y cómo cultivar *Rubus glaucus* y *Solanum quitoense*. Metodologías de investigación operativa y participativa y las experiencias productivas de los agricultores surgen como una herramienta prometedora para el desarrollo de producción agrícola específica por sitio para cultivos poco estudiados. Esto es especialmente importante en Colombia, debido a la poca investigación sobre los cultivos en general, y la heterogeneidad ambiental de los sitios donde se desarrollan los cultivos.

SAMENVATTING

Planten en oogsten is telkens weer een uniek teeltexperiment. In onze studie veronderstellen we dat als het mogelijk zou zijn om de beheerspraktijken en de omgevingsomstandigheden van een teeltsysteem te karakteriseren, en om informatie over het geoogste product in voldoende grote aantallen en onder verschillende oogstomstandigheden te verzamelen, het mogelijk zou moeten zijn om analytische methodes op te stellen die het teeltsysteem adequaat beschrijven. Deze benadering wordt gebruikt in operationeel onderzoek, waarbij mathematische modellen van geobserveerde processen die delen identificeren die zouden moeten gewijzigd worden om het systeem te optimaliseren.

Informatie over verschillende teelten van suikerrtiet in het departement *Valle del Cauca* werd verzameld door het *Centro Colombiano de Investigación de la Caña de Azúcar* (CENICAÑA) en werd ingevoerd in gegevensgestuurde modellen die werden toegepast op twee weinig bestudeerde vruchten uit Colombia: de Andes braambes (*Rubus glaucus* Benth) en de lulo (*Solanum quitoense* Lam).

Het onderzoek integreert informatie bekomen bij kleinschalige telers, omgevingsinformatie uit openbare gegevensbanken en mathematische technieken om te komen tot een plaatsspecifieke benadering voor de bestudeerde teelten. Het werk neemt de ervaringen van de telers omtrent een brede waaier aan omgevings- en socio-economische omstandigheden in rekening, om ons in staat te stellen een combinatie van opbrengstverklarende factoren te identificeren. De de telers zelf werden bij de gegevensverzameling betrokken, wat eigen is aan operationeel en participatorisch onderzoek.

De aan de hand van suikerrietgegevens ontwikkelde modeleringsmiddelen verklaarden 90 % van de variatie in de opbrengst van *Rubus glaucus* en werden gebruikt om (a) opbrengstverklarende factoren te identificeren aan de hand van een sensitiviteitsmatrix; en (b) relaties tussen de relevante variabelen te visualiseren aan de hand van een niet-gesuperviseerd clusteringsalgoritme.

Bij *Solanum quitoense* werd een iteratieve procedure gebruikt dat (a) de variabelen identificeert die de opbrengstvariatie best verklaren, (b) gelijkaardige omgevingen clustert aan de hand van een neuraal netwerk; en (c) het omgevingseffect, de aan het geografische gebied eigen culturele factoren, en de managementkwaliteiten van de teler analyseert in een gemengd model dat meer dan 80 % van de opbrengstvariabiliteit verklaarde.

De meest optimale omstandigheden voor *Rubus glaucus* zijn: een gemiddelde temperatuur tijdens de eerste maand vóór de oogst van 16–18 °C, een minimale effectieve bodemdiepte van 65 cm en

geringe regenval gedurende de maand vóór de oogst indien de bodem slecht gedraineerd is, of matige tot lage regenval in gebieden met beter gedraineerde bodems. De beste omstandigheden voor *Solanum quitoense* zijn: een gemiddelde temperatuur van de maand vóór de oogst van 15,8-19 °C, een bodemdiepte van 40-67 cm en een helling van het terrein van 13-24°. Aangezien de modellering een effect van de locatie op de opbrengst aantoonde, en dus suggereerde dat factoren buiten degene die tijdens de gegevensverzameling werden bekomen, ook een invloed hebben, werden proxies voor het teeltmanagement en socio-economische omstandigheden in onze studie geïntegreerd. Zo was er bijvoorbeeld onvoldoende informatie om sociale en managementfactoren te identificeren die geassocieerd zijn met hoge opbrengsten, hoewel er landbouwbedrijven waren met hoger dan gemiddelde opbrengsten, wat er op duidt dat men daar de teelt efficiënter beheerde.

Onze resultaten stellen ons in staat om managementpraktijken die door goed boerende telers worden gebruikt, over te dragen aan telers met lagere opbrengsten, zodat ook zij, door de plaatsspecifieke aanbevelingen van hun collega-telers ter harte te nemen, hun opbrengsten kunnen verbeteren.

Het onderzoek toont aan dat analyse en interpretatie van de ervaringen van telers, in combinatie met gegevens omtrent teeltomstandigheden, voldoende informatie oplevert omtrent de meest optimale plaatsen en praktijken voor de teelt van *Rubus glaucus* en *Solanum quitoense*. Operationele en participatorische onderzoeksmethodes zijn veelbelovende technieken bij de ontwikkeling van plaatsspecifieke teeltpraktijken bij minder bekende tropische vruchten. Dit is in het bijzonder van belang in Colombia, waar onvoldoende aandacht wordt besteed naar landbouwonderzoek in het algemeen en waar de omgevingsomstandigheden extreem heterogeen zijn.

TABLE OF CONTENTS

1 GENERAL INTRODUCTION

In the linear model of research, extension officers visit farms and transfer technology to farmers using blanket recommendations, which are based on knowledge generated on research stations. They typically do not take farmers' previous knowledge into account (Pretty, 1991; Thompson and Scoones, 1994; Altieri, 2002; Hall, 2005; Van Asten et al., 2009; Lacy, 2011). In the latter model of research, researchers measure a crop plant's responses to specific variations in a small number of factors with all other factors being controlled to the extent possible. This model was useful where the research station was considered to be representative of a large, relatively homogeneous area of land (Cock, 1985; Braun et al., 1996; Gauch and Zobel, 1997). It would seem to be less useful in heterogeneous environments where farm is less likely to be similar to the research stations so that the standard recommendations based on the findings in controlled conditions at the research station may not be the most appropriate. In the tropics, farming systems are rarely homogeneous. Climate and soils can be very diverse with considerable variation over short distances, and the linear model is thus not always the most appropriate.

Participatory research recognizes farmers' knowledge gained over time by observing the combined effects of all variables that influence crop growth and which are often impossible to control (Conroy et al., 1999). Farmers use this knowledge to optimize their agricultural systems, which is essentially the same approach used by operational research in industry. Operational research analyses observations of an industrial organization's operations in order to find better ways of performing them (Operational Research Society, 2006). Under-researched fruit species are typically perennials, and take several years to come into production. Participatory research allows data to be collected about growers' production experiences. Moreover, according to Edgerton (2004), farmers think that information from their own plots is more relevant to them than data from research stations. The present research used operational research methodologies, based on modern information technology, which allows even small-scale growers to interpret their multiple production experiences.

Sugarcane growers in Colombia collected data on their production experiences together with environmental and socio-economic factors to develop and apply a site-specific approach to their fields. They doubled their production over a period of 30 years (Isaacs et al., 2007). Similar approaches could be adapted to small-scale producers through the characterization of their crop's growing environment using public databases on climate, landscape, and topography coupled with data obtained at farm level (edaphic conditions, crop management) in order to improve their

production. This information can then be exploited through modelling approaches to understand yield variability and formulate recommendations to other small-scale growers.

Rubus glaucus and *Solanum quitoense* are two under-researched tropical fruit species grown in heterogeneous environments by smallholders, who typically have poor production and very few resources at their disposal to make informed management decisions.

The basic premises of this research are:

- as the conditions under which farmers operate are highly heterogeneous and farmers are continuously trying out something new, every time a farmer plants and harvests a crop, this represents a unique experience or " cropping event"; and
- if it were possible to (a) compile information on how individual cropping events were managed; and (b) characterise the conditions under which a large number of these experiments occur, it would be possible to deduce optimum practices for specific conditions.

The objectives of this thesis are therefore to:

- demonstrate that the principles of operational research methodologies developed for sugarcane in Colombia can be applied to under-researched crops such as *R. glaucus* and *S. quitoense*, by providing growers with the basic information, which, together with their own experiences, can be analysed and interpreted to provide insights into how yield varies with variations in the environment;

- evaluate modelling methodologies developed during this research for sugarcane, to determine their suitability as tools for modelling the response of *R. glaucus* and *S. quitoense* to variation in environmental conditions and management practices, by using information of crop response collected by small-scale producers;

- use these methods to identify the conditions that are most suitable for the production of *R. glaucus* and *S. quitoense*, based on the information obtained from those farmers who managed their crops particularly well.

Having briefly presented the utility of this approach, the remainder of this chapter gives some general information about the linear model of research and technology transfer, participatory research, and operational research based on farmers' production experiences. It then goes on to site-specific agriculture, based on operational research and farmers' production experiences.

Chapter 2 outlines the methodology for collecting data on farmers' production experiences and contrasts its potential in a well-researched crop (sugarcane) with two under-researched crops (Andean blackberry and lulo) in Colombia. Chapter 3 deals with data management and analysis, and the development of predictive and explanatory models taking into account the experience acquired during the sugarcane modelling exercise using the database provided by CENICAÑA. Chapters 4 and 5 illustrate the application of the strategies developed for sugarcane to Site-Specific Crop Production (SSCP) of Andean blackberry and lulo. The application is based on the approaches of operational and participatory research, integrated with information collected by small-scale producers coupled with publicly-available environmental data. In chapter 4, the SSCP strategy seeks to identify the most important factors that explain productivity of *R. glaucus*. In chapter 5, in addition to determining the most relevant factors explaining *S. quitoense* production, the effects of geographical location and variation within and between environmental clusters are investigated. Finally, chapter 6 presents a general discussion and provides concluding remarks, as well as recommendations for further research based on the experience acquired during the development of this thesis.

1.1 LINEAR MODEL OF RESEARCH AND TECHNOLOGY TRANSFER

The linear model of research, also called "top-down research model", has been developed using many experiments conducted in controlled conditions (Chambers and Ghildyal, 1985; Chambers et al., 1989; Pretty, 1991; Thompson and Scoones, 1994; Marsh and Pannell, 2000; Russell and Ison, 2000; Altieri, 2002; Hall, 2005; Van Asten et al., 2009; Lacy, 2011). Crops such as wheat (*Triticum* spp.) or maize (*Zea mays* L.) have increased their yield through plant breeding programs that used the linear research model (Evans and Fischer, 1999). Standard recommendations generated by this model were used for large relatively uniform areas (Cock, 1985; Braun et al., 1996; Gauch and Zobel, 1997). It is difficult however, to propose standard recommendations for a specific production site, as agricultural systems are heterogeneous in terms of environment, complexity and also socio-economic conditions (Basso et al., 2001).

In the linear model of research, the basic rationale guiding defining extension messages for the technology transfer process has been to develop varieties that are supposed to be adapted to wide regions. This model has also applied to sugarcane (*Saccharum officinarum* L) (Evenson, 1981). In general, researchers have continued to develop varieties adapted to mega-environments that are supposed to be environmentally homogenous. As a consequence, the linear model of research is often used to develop recommendations for improved crop management (Cock, 1985; Braun et al., 1996; Gauch and Zobel, 1997). Extension officers have been trained to use this model with standard recommendations for large, relatively homogeneous regions (Benor and Harrison, 1977).

Standard recommendations such as the use of herbicides in certain crops were thus successfully implemented in a wide range of conditions. Nevertheless, although the use of widely adapted varieties, technologies, and standard recommendations were successful for many decades, it is now accepted that it is more effective to generate more targeted recommendations based on and derived from a large number of on-farm measurements. Moreover, the heterogeneity of growing conditions over different agricultural systems and the wide range of crops cultivated worldwide, suggest that standard recommendations are not always the best option, so optimizing agricultural systems requires the development of site-specific recommendations (Isaacs et al., 2007; Niederhauser et al., 2008; Cock et al., 2011).

As an example, working as an extension officer in Colombia years ago, the author of this thesis made blanket recommendations following the linear model of research and technology transfer. Extension officers were instructed to make recommendations about growing and managing a specific variety of coffee (*Coffea arabica*) that was tolerant of coffee rust disease (*Hemileia vastatrix*) in the department of Caldas. The variety had been developed in very specific environmental conditions by researchers, but was supposed to be grown by farmers over a wide range of conditions. Farmers quickly realized that the variety did not perform well in conditions different from those in which it had been developed. As a result, researchers tried to adapt it to a wider range of conditions. However, given their previous experiences, farmers were unwilling to accept it (La Patria, 2011; Sandoval, 2011). Traditionally, extension officers are provided with technologies and recommendations not only on new varieties but also on pesticides, fertilizers, and other managements options to pass on to farmers. Some of them work well, but some of them do not (Lacy, 2011).

Edgerton (2004) notes that recommendations from the linear model of research are not used by all farmers. He therefore questions their possible success as farmers get most of their knowledge from their own experience and from other farmers. He concludes that most farmers usually spend a lot of time on how to adapt the recommended technologies to their own conditions. Indeed, farmers are continually doing research and are in a permanent process of technological innovation (Lyon, 1996). In a literature review of pest impact on crop yield, Rosenheim et al. (2011) showed that 88% of surveyed studies had been performed on research farms whereas only 12 % had been done on farmers' fields. Montaner (2004) remarks that researchers often only take into consideration a limited number of farmers' fields. As result they do not always know the range of management options that farmers use. Another difficulty of the linear model of research is that there is often a lack of knowledge on the part of the researchers of what really works at farm level.

Rosenheim et al. (2011) stated that in this formal research model it is common to assume that “in a well-replicated experiment, a researcher generates one or more treatments by manipulating some variable, A, while holding other conditions as nearly constant as possible; assigns those treatments randomly to experimental units; and then measures a response variable, B. If the response variable B differs significantly across treatments, then the experimenter can infer with a high degree of confidence that a change in A causes a change in B. In contrast, when a researcher observes a correlation between a natural, pre-existing variation in variables C and D, it is difficult to know whether the correlation reflects a causal influence of C on D, of D on C, or whether C and D are not causally related to each other at all, but instead are both influenced by some other variable(s) E, F, and so on, which may or may not have been measured by the experimenter.”

The conclusion is that experiments carried out on-farm by farmers themselves, under specific environmental and socio-economic conditions, generate specific knowledge for and on their production sites. Using the latter knowledge, much better targeted extension messages can be developed.

1.2 PARTICIPATORY RESEARCH

Growers know the variations that exist on their farms. They constantly learn from their experiences, and adapt and implement technologies they develop according to what they learned.
According to Conroy et al. (1999), participatory research as a formal research methodology was designed to capture all the information that can be generated by farmers.

The approach recognizes farmers’ ability to do research. The authors stated that there are four different ways to manage participatory on-farm research. First, there is the traditional linear mode in which farmers’ knowledge is less considered. Second, there is a consultative mode in which technical alternatives proposed by a researcher are tested in growers’ fields. Third, there is a collaborative mode in which ideas about the trials to be tested are generated with the participation of researchers and farmers together. Finally, there is the collegiate mode in which farmers decide on the content of the experiments to conduct. This offers a huge opportunity to use information collected by the producers themselves under specific conditions. This information can be coupled with modern information technology to characterise specific growing conditions and relate them to crop responses and thus arrive at site-specific recommendations.

The system of informally exchanging information amongst farmers formed the basis of an important revolution in agricultural production between the seventeenth to nineteenth centuries in England. Over that period, there was a threefold increase in livestock and crop production without

the involvement of research stations, government ministries of agriculture, extension institutions or pesticides. Moreover, rural transport infrastructure was also notoriously poor (Pretty, 1991; Overton, 2006). The knowledge acquired by farmers was spread through farmer-to-farmer contact, open days, workshops, rural tours, informal training, and even publications, which farmers then adapted to their own conditions (Pretty, 1991). As a consequence, one of the most important revolutions in agriculture was based on technology developed by farmers themselves (Overton, 2006).

As noted by Thompson and Scoones (1994), farmers' knowledge is often assumed as primitive, wrong, and unscientific. Additionally, researchers regard the world as easy to disaggregate into effects of independent variables. They tend to demonstrate that the factors that they have selected are well-suited for the response they want to model. Hence, they focus on statistically significant differences generated by different levels of variables that they themselves have chosen as the most relevant (Rist, 1997). But the reality is that farmers face dynamic, complex, rapidly changing, and often chaotic combinations of factors that they have to take into account in managing their farms in order to improve their agricultural systems

Rosenheim et al. (2011) pointed out that, "experiments are the ultimate intellectual playground in which researchers can attempt to implement any manipulation that they can imagine. In contrast, observational studies are restricted to conditions that actually occur in the field". The latter authors suggest that knowledge generated by researchers can be complemented by farmers' production experiences, performed under a wide range of conditions.

1.3 OPERATIONAL RESEARCH BASED ON FARMERS' PRODUCTION EXPERIENCES

Participatory research and the linear model of research both focus on doing experiments to obtain knowledge on crop performance. As Schank (2011) writes, "now, while it is difficult if not impossible to conduct controlled experiments in most aspects of our own lives, it is possible to come to understand that we are indeed conducting an experiment when we take a new job, or try a new tactic in a game we are playing, or when we pick a school to attend, or when we try and figure out how someone is feeling, or when we wonder why we ourselves feel the way we do". The author of this thesis subscribes to this view and insists that farmers are continually experimenting. However, growers do not have the tools or the methodologies to make the best use of their experimental results. On the other hand, operational research methodologies can help farmers make sense of their experiences.

As an example, operational research observes an industrial 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 used by total quality management which emphasizes monitoring, measurement and systematic capture and codification of tacit knowledge to detect trends (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 elucidate factors associated with cardiovascular disease and hence to recommend methods to control the latter (Framingham heart study, 2006).

Farmers are embedded in specific agro-ecological and socio-cultural contexts that change constantly. Hence, growers have constantly to make adjustments associated with their specific production conditions (Thompson and Scoones, 1994). These continuous modifications made by farmers in relation with ding to their particular circumstances may induce an increase in their yields. They can therefore take advantage of the analysis of their multiple experiences using the principles of operational research.

As an example of this, the Cropcheck system determines the most appropriate practices for any given condition (Lacy, 2011). It benchmarks farmer crops to identify the best practices associated with high yields. Through this approach, both productivity and profitability of wheat increased by 50% over a six-year period in Chile (Lacy, 2011). Similar approaches have been implemented in Australia in sugarcane, where the TopCrop system makes grower information available to researchers in order to support farmers to establish optimum practices for particular crops and conditions (Evans and Fischer, 1999; Schulz et al., 2001; Lawes and Lawn, 2005). Researchers have used information generated by farmers to recommend the most suitable varieties and management for cocoa (*Theobroma cacao*) in Ghana. Similarly in Canada, information from strips planted in farmers' fields has reduced the cost of producing new varieties of winter wheat (*Triticum aestivum*), and in Asia the interactions of temperature and radiation on rice (*Oryza sativa*) were elucidated from multiple trials on farmers' fields (Yan et al., 2002; Edwin and Masters, 2005; Welch et al., 2010).

1.4 SITE-SPECIFIC AGRICULTURE

At the *Centro de Investigación de la Caña de Azúcar* (CENICAÑA) in Colombia, researchers and growers combined environmental information on soils and climate with crop performance, to determine site-specific guidelines for sugarcane management, in order to do that, they combined approaches of operational and participatory research (Isaacs et al., 2007; Cock and Luna, 1996).

Both Site-Specific Agriculture (SSA) and Precision Agriculture (PA) are largely based on the principles of the observation of crop response to temporal and spatial variation. CENICAÑA defines SSA as: “the art of managing crops according to the spatial and temporal variation in conditions of the site where they are grown with a view to optimize production” (Isaacs et al., 2004). CENICAÑA has been the pioneer of SSA in Colombia. Through SSA, sugarcane yield increased from 5 t/sugar/ha/yr in 1980, to 11 t/sugar/ha/yr 2003-2005 (Figure 1.1.).

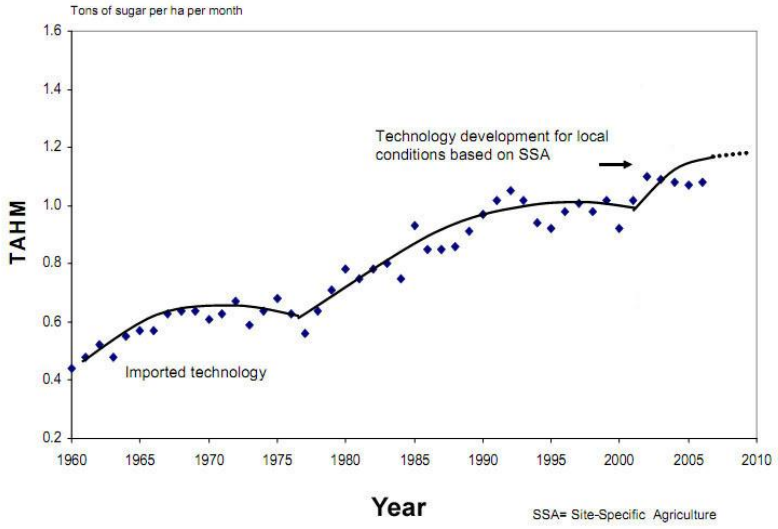


Figure 1.1. Evolution in sugarcane productivity (tons of sugar per hectare per month - TAHM) in Colombia from 1960 – 2008 (Isaacs et al., 2007)

In PA, which is also called site-specific management, crops are managed taking into account the environmental conditions under which they are grown, according to temporal and spatial variation (Cassman, 1999). Hence, PA seems to be useful when information on crop response to specific conditions is available in an attempt to optimize the agricultural system at micro-level (Figure 1.3). For instance, when a particular spot in a field is identified as a place where plants are performing poorly due to the lack of a given nutrient, then these plants can be supplied with the lacking nutrient (Bongiovanni and Lowenberg-Deboer, 2004). PA has been widely implemented in countries where environmental conditions vary less than in the tropics, and knowledge about the factors that results in high productivity is available (see Proceedings of the 14th annual symposium on precision agriculture, 2010). PA normally involves monitoring of within-site variability with wireless sensors, telemetry systems, and sophisticated analytical tools. For most crops in developing countries in the tropics, farmers do not have access to these sophisticated tools. Moreover, their first requirement often is to manage their farms well, before using PA to address within-site variation and increase yields.

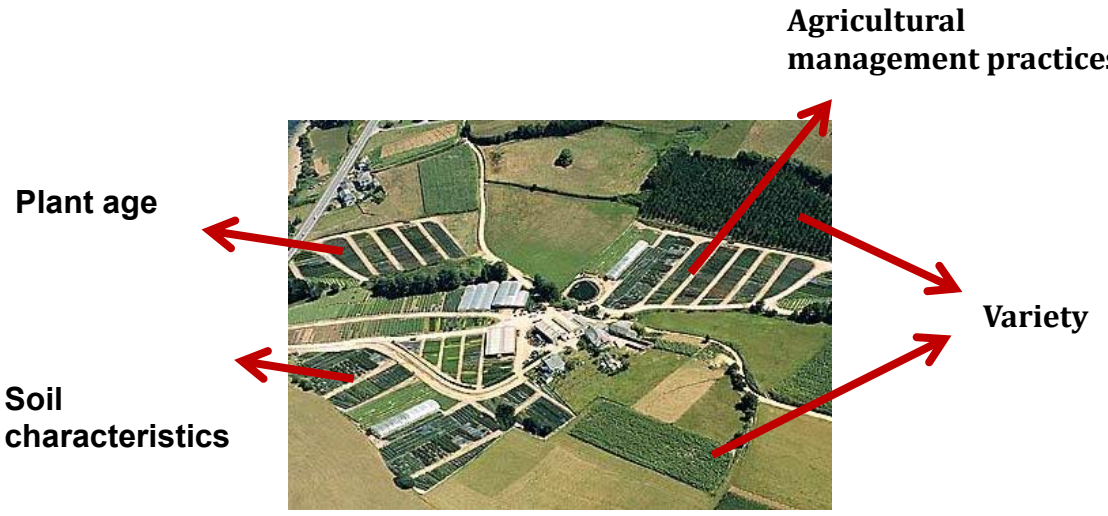


Figure 1.2. SSA looks for variation at field scale (macro-level) in which each field has features that make it relatively homogeneous in terms of environmental conditions and agricultural management practices. In SSA and PA, fields are often called Management Units (figure elaborated by the author)

PA can be implemented either in a whole field (National Research Council, 1997; Cassman, 1999; Läderach, 2011) or in specific places within a field (Basso et al., 2001; Erazo, 2011). For PA to be applied at within-field level, spatial variation is commonly associated with a single factor such as deficiency of water or specific soil nutrients, other soil features, or outbreaks of pests or diseases, which are limited to a small area. Other conditions such as management practices and climate are then assumed to remain stable over the whole plot (Cock et al., 2011).

In the tropics, there is a need to generate information for developing both knowledge and technology adapted to field level, before venturing into a higher resolution PA (Cassman, 1999; Spaans and Estrada, 2004). This was true for sugarcane in Colombia where both knowledge and technology data are available for the last 20 years. This allowed to develop SSA, which is now being refined to develop PA to deal with within-field variation (Erazo, 2011). The first step in implementing PA requires understanding of the variability over relatively homogeneous areas, before dealing with the variation in individual fields (Figure 1.2) (Spaans and Estrada, 2004). Furthermore, it has been demonstrated in both sugarcane and banana, that before to define the spatial and temporal variation of the factors that affect crop performance, information on the crop response to these factors is needed first (Cock and Luna, 1996; Cassman, 1999; Spaans and Estrada, 2004).

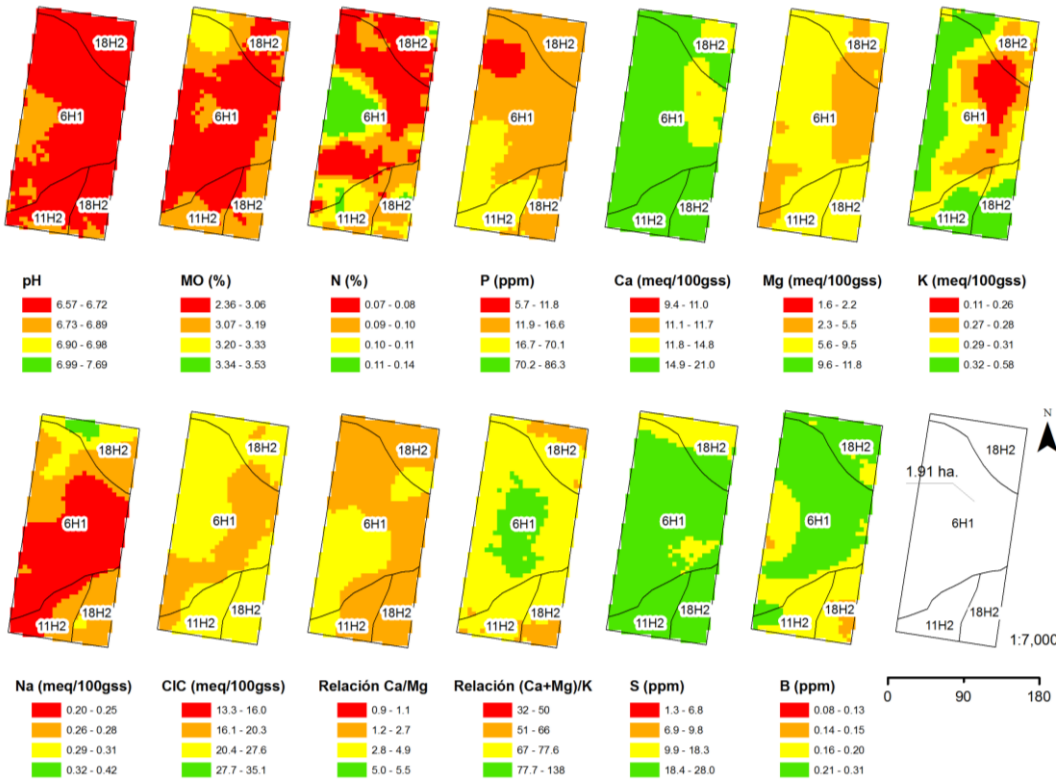


Figure 1.3. Example of PA, looking for variation at within field scale (micro-level) (spatial variability of soils effect on sugarcane yield) (Erazo, 2011)

Recently, and similarly to sugarcane, SSCP based on the approaches of operational and participatory research was applied to coffee in Colombia. As an example, brew quality is important in the coffee market, but the conditions that yield high quality are very narrow, and are largely associated with bean size. Management practices implemented by farmers in commercial fields that give high quality were used to define denomination of origin criteria for Colombian coffee (Niederhauser et al., 2008; Cock et al., 2011).

1.4.1 SSCP FOR UNDER-RESEARCHED CROPS IN COLOMBIA

For under-researched crops, the combination of (applying approaches) of operational research (coupled) with SSCP techniques, offers an alternative to experiments on research stations. As far as we know, this combination of methods has not been applied before to under-researched crops grown by small farmers in Colombia nor elsewhere. However, caution is needed when applying this method of operational research to agriculture, as the latter has been developed for industrial processes where variability is typically reduced to the extent possible. In agriculture, even though conditions cannot be controlled to the same extent as in industry, the principles of the methodology can be preserved. As a member of a multidisciplinary team, the author was involved in the development and validation of tools to collect, analyse, and interpret information on farmers`

production experiences in an attempt to apply SSCP methods to *Rubus glaucus* and *Solanum quitoense*.

The environmental conditions under which *R. glaucus* and *S. quitoense* grow are quite diverse, varying a lot in time and space. These conditions are therefore difficult to control, so that productivity varies widely between regions and even between farms. Moreover, both crops are harvested continuously during the year and productivity fluctuates throughout the year.

For temperate crops, management can be optimized, and is applied to short and well-defined cropping and harvest periods. In contrast, tropical crops present a multitude of different management options and environments. For example, production during the dry season may require totally different water and pest management techniques to those required in the wet season. A direct consequence of this need for multiple management options is the necessity for continuous experimentation by producers. Furthermore, most tropical fruit plants are perennials, and have been neglected by traditional agricultural research. In Colombia, most research on tropical fruit species has addressed fruit quality and/or biochemical composition (Estrada, 1992; Osorio et al., 2003; Flórez et al., 2008; Pulido et al., 2008; Acosta et al., 2009). There are only a few studies on these crops' responses to variations in management and/or environmental conditions, hence *R. glaucus* and *S. quitoense* are under-researched.

In summary, the lack of knowledge/data on the most suitable conditions to grow tropical fruit species, heterogeneous growing conditions, and year-round production of many tropical species would require an impossibly large number of experiments to meet the requirements of the conventional linear model of research and technology.

Collecting response data directly from farmers by compiling their production experiences on their own fields appears to be an effective way to obtain data needed to develop SSCP for under-researched tropical fruit species in Colombia. It is clear that the collection and analysis of large datasets is necessary to be able to draw conclusions from the natural occurring variation on farmers' sites. Modern information technology that can be used to capture this variability includes Global Positioning Systems (GPS), publicly-available environmental information, satellite imagery, fast computers, and analytical software packages suited to handle large, categorical datasets. These allow the large amount of information generated by describing and monitoring farmers' production experiences to be collected, processed, analysed, and interpreted. The processed information can provide insights on plant responses to individual factors and the interaction between them, but more importantly, indicates a crop's likely response to environmental and management factors on the farmers' production sites.

Crop responses to specific factors can be analysed in different ways, but for each production site in the site-specific approach used here, the so-called *event* is used as the unit of analysis (Cock, 2007). “An event occurs in a particular site within a given period of time and it is normally taken as the period between planting and harvest, or as the period between one harvest and the next, in crops which are not replanted after each harvest” (Cock et al., 2011).

1.5 THE CONTEXT OF THE PRESENT RESEARCH

Most of the research reported here was done over a three years period, covering July 2005–July 2008, as part of a cooperation project between Colombia and Switzerland, “Site-Specific Agriculture for Tropical Fruits” (SSAFT). In Colombia, the institutions involved were *Corporación BIOTEC* (CB), the *Centro Internacional de Agricultura Tropical* (CIAT) and CENICAÑA. The latter contributed the sugarcane database used for developing the modelling tools and knowledge about sugarcane. In Switzerland, *the Haute Ecole d'Ingénierie et de Gestion du canton de Vaud* (HEIG-VD) contributed through its experience in dealing with real-world data to build models by means of supervised and unsupervised artificial neural networks.

The main objective of our research was to increase the competitiveness of smallholder fruit growers in Colombia, through the development and integration of strategies based on integrating farmers’ production experiences with publicly-available environmental data using computational models to understand yield variability. *The State Secretariat for Education and Research* (SSER) in Switzerland granted three scholarships to two electronic engineers and one agronomist for the development of the models at HEIG-VD. The engineers developed the artificial neural networks models. The author of this thesis coordinated data collection, compiled the data in centralized databases, analysed the data, and interpreted the model outputs from an agronomic point of view. The strategy at HEIG-VD regarding the development of artificial neural network modelling techniques for analysing and interpreting the data is summarized in Figure 1.4.

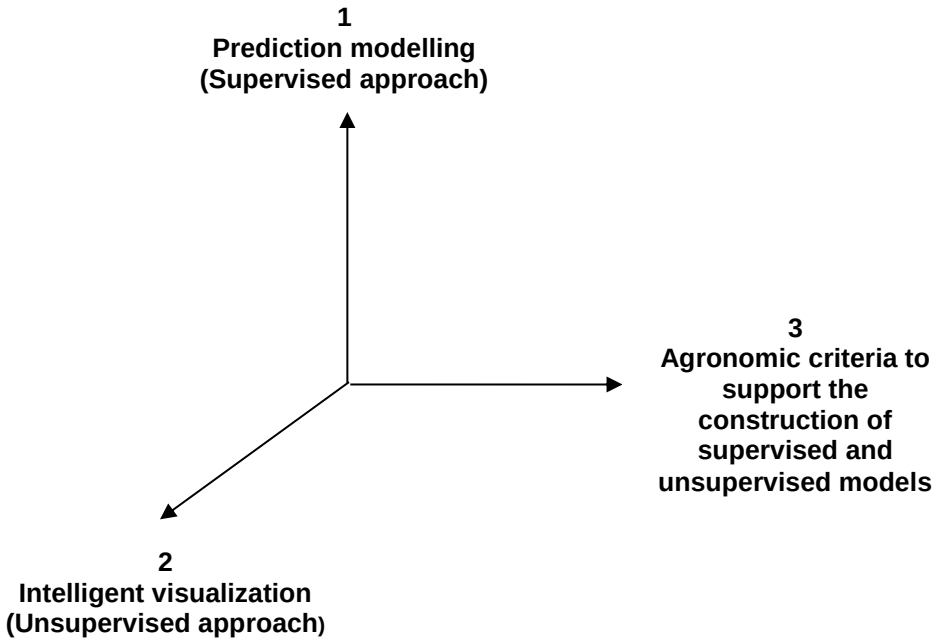


Figure 1.4. Modelling strategy at HEIG-VD: three axes of research were identified to develop the computational models. The first two axes were conducted by two electronic engineers, while the author's research focused on the third axis

The agronomic part focused on 3 elements:

(a) literature review of the applications of supervised and unsupervised techniques in agriculture;

(b) coordinate data collection in Colombia and support the electronic engineers responsible for the construction of prediction and intelligence visualization models, with agronomic knowledge, identify the most suitable variables to build the mathematical models, and facilitate data interpretation; and

(c) extend the experience achieved in the development of the sugarcane models to *Rubus glaucus* and *Solanum quitoense*.

2 METHODOLOGY FOR COLLECTING FARMERS' PRODUCTION EXPERIENCES, HARVEST EVENTS AND ENVIRONMENTAL INFORMATION

2.1 INTRODUCTION

The hypotheses that this research seeks to verify are that: (a) modern information technology can be used to combine information on farmers' production experiences with publicly-available environmental databases; and (b) the principles of operational and participatory research facilitate the task of collecting, characterizing and interpreting data on many cropping events that occur under a wide range of conditions.

The research presented in this thesis is based on methodologies that were developed in Colombia for the well-studied crop sugarcane (Quintero and Castilla, 1992; Carbonell et al., 2001; Isaacs et al., 2004 and 2007; Torres et al., 2004; Cock et al., 2011). These methodologies were adapted to crops such as *Rubus glaucus* and *Solanum quitoense*. The author of this research participated in the development and validation of these methodologies.

The analysis and modelling of the information as applied to both under-researched crops was published as two peer-reviewed papers in international journals as part of this thesis and is presented in detail in chapters 4 and 5. However, in the original article versions there is only a brief description of the methodology due to length constraints required by the journals. In this chapter we provide more detailed explanation of the procedures followed to collect the information compared to the published papers. In addition, in the present section, important background information on the sugarcane production system is briefly provided. The methodology for collecting farmers' production experiences, standardization and compilation of information in centralized databases is then illustrated, highlighting the adaptations which were required to fit within the social and technical context of under-researched fruit crops grown by small farmers.

The methodology comprises of:

data collection: compilation of farmers' production experiences based on cropping events that describe the environmental and management conditions under which sugarcane, *R. glaucus* and *S. quitoense* are cultivated, and

data management of the information captured: the strategy to store and transfer the information collected for sugarcane, *R. glaucus* and *S. quitoense*

2.2 CROPS UNDER STUDY

The operational research approach for tropical fruit species grown by many independent small farmers requires the collection of data from multiple sites continuously over time. Unlike sugarcane, there is not just one harvest event at approximately yearly intervals. This major difference between continuously harvested crops coupled with major differences in the organization of the sugarcane sector indicated that specialized approaches to data collection were needed for the Andean fruit species we studied here. The latter approaches were developed as part of this research.

2.2.1 SUGARCANE

Sugarcane is an important crop in the Valle del Cauca department in Colombia (Figure 2.1). Crop management is based on the known agro-ecological characteristics of each production unit, and soil–plant–water relations (Isaacs et al., 2007). CENICAÑA is the national research institution for sugarcane, financed and managed directly by the sugarcane sector. Unlike for most crops in Colombia, where there is a lack of organization of the supply chain and where they do not have a dedicated research institution, CENICAÑA provides sugarcane growers with advise and technology based on a close relationship between growers and researchers, who work together as a team.

Figure 2.1. Location of the Valle del Cauca department within Colombia (figure elaborated by the author)

The processes of data collection and management are strongly associated with the organization of the supply chain. For instance, the weight of the sugarcane produced by each plot is measured by the sugar mills as payment to growers is based on weight and sugar content of the harvested

cane. The contracts between growers and mills indicate the area of each plot to be harvested; hence, cane tonnage can be related to specific sites with known area. Furthermore, mills collect information on planting and harvest dates, and variety planted. The mills store this information in their databases and make it available to the centralized CENICAÑA database. In the mid 1990s, the sugar industry standardized the measurement methods for the whole industry, thus all data collected is in standardized formats

The characterization of specific sites is based on multiple data source. The sugarcane growers' organization conducts soil surveys, operates automated weather stations, and produces digitized maps that facilitate data capture (Isaacs et al., 2000; Isaacs et al., 2004; Cock et al., 2011). In this research, data from the centralized CENICAÑA database was used to develop modelling tools that were then applied to *R. glaucus* and *S. quitoense.*

2.2.2 ANDEAN BLACKBERRY AND LULO

Andean blackberry (*Rubus glaucus* Benth.), also known as the Andes berry or *mora de Castilla*, (Bioversity International, 2005a), is grown commercially in Colombia, Ecuador, Guatemala, Honduras, Mexico and Panama (Franco and Giraldo, 2002). It is highly appreciated for its sweet-acid taste, dark-red colour (Figure 2.2), and pleasant aroma (Ramos et al., 2005). Lulo (*Solanum quitoense* Lam.) is a fruit native to the humid forests of the north-western Andes (Figure 2.3). It is grown commercially in Colombia (6640 ha), but production does not meet national demand (Tafur, 2006; Medina et al., 2008). It has a high-quality juice with a nice aroma, high nutritional value, and is used in the agro-industry (National Research Council, 1989; Franco et al., 2002; Franco and Giraldo, 2002; Osorio et al., 2003; Flórez et al., 2008; Pulido et al., 2008; Acosta et al., 2009; PAVUC, 2010). It is also grown in Costa Rica, Ecuador, Honduras, Panama, and Peru (Bioversity International, 2005b).

Both Andean blackberry and lulo are important sources of income for smallholders in the Colombian hillsides (Sora et al., 2006), which have limited infrastructure and produce in conditions of high environmental variability (Franco et al., 2002; Franco and Giraldo, 2002). Growers of Andean blackberry and lulo in Colombia face high incidence of pests and diseases, which are costly to control and reduce profitability (National Research Council, 1989; Estrada, 1992; Flórez et al., 2008). Furthermore, lack of effective social organization along the supply chain has led to little development of new technology and insights into improved management practices.

Well-organized and economically powerful sectors such as sugarcane, coffee and oil palm (*Elaeis guineensis*) have their own research institutes in Colombia. In contrast research on tropical fruits is

quasi non-existent because the fruit supply chains are weak, despite the initiatives of the national government to promote growers' organizations. Moreover, because of the long biological cycle of these species, it would take many years of study to define the plants' responses to environmental conditions. Most research has therefore been short-term, which means that little is known of their physiology. There is relatively little published information on the phenology of these species. Descriptions are often found in extension type literature where experienced agronomists have documented their field observations, normally for a specific ecological niche, and hence may not be extrapolated to other areas with confidence. Within these limitations the general phenological stages for Andean blackberry and Lulo are described in Figures 2.4 and 2.5.

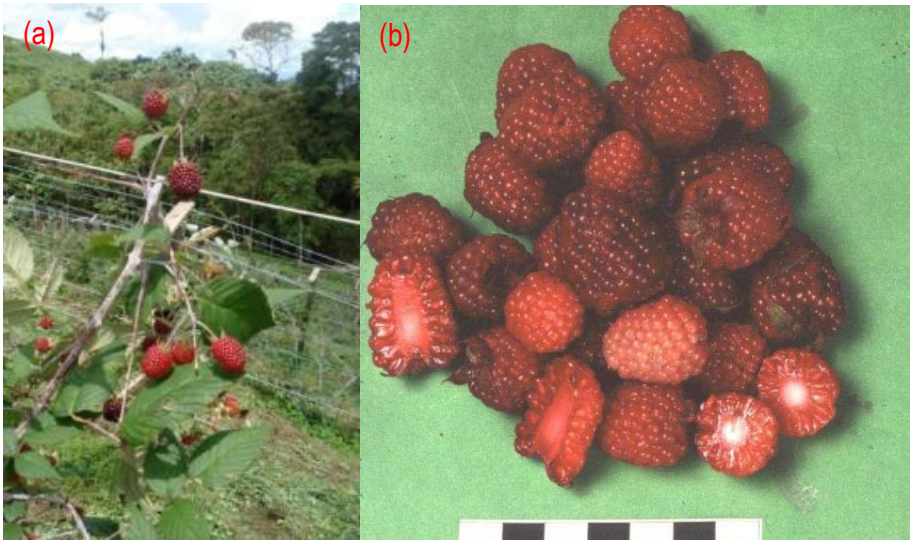


Figure 2.2. Andean blackberry: (a) branch with fruits; (b) fruit as it can be found in local markets. Taken from the New World Fruits Database (Bioversity International, 2005a)

Figure 2.3. Lulo: (a) plant exhibiting fruits; (b): fruit as it can be found in local markets. Taken from the New World Fruits Database (Bioversity International, 2005b)

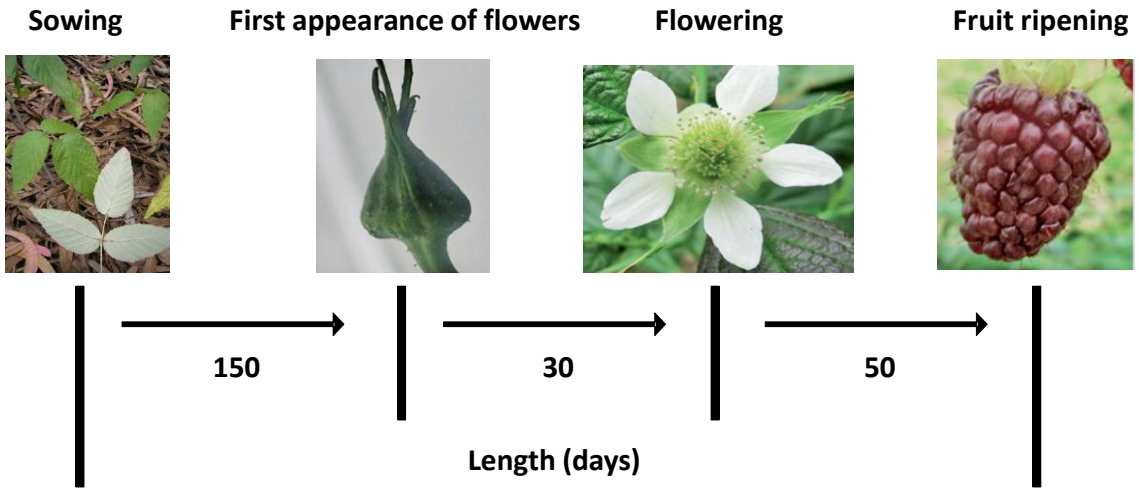


Figure 2.4. Phenological stages for Andean blackberry (Franco and Giraldo, 2002; Grijalba et al., 2010)

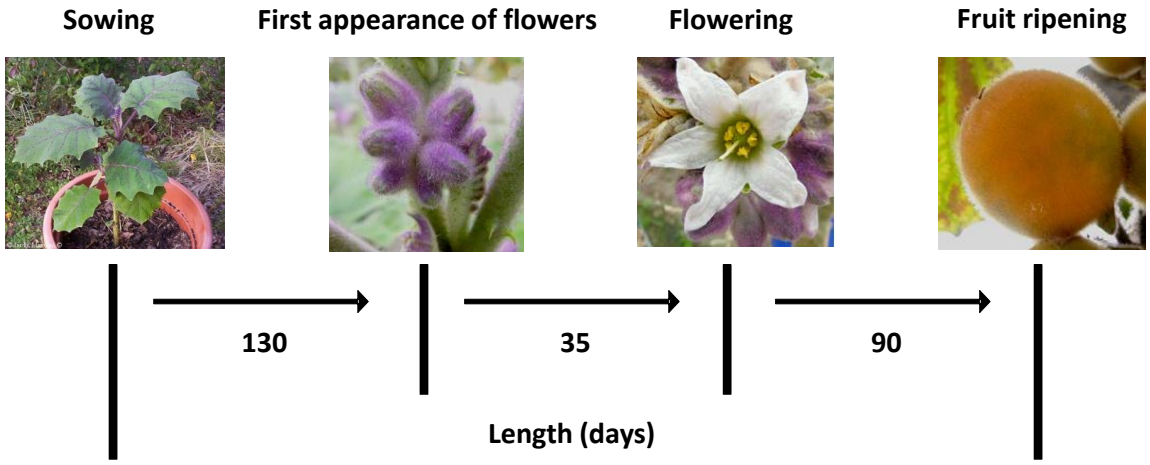


Figure 2.5. Phenological stages for Lulo (Franco et al., 2002; Garcia, 2003)

2.3 COLLECTION OF INFORMATION ON CROPPING EVENTS AND GROWING CONDITIONS

2.3.1 SUGARCANE

The *Instituto Geográfico Augustin Codazzi* (IGAC) and the *Corporación Autónoma Regional del Valle del Cauca* (CVC) have produced soil and geographic maps of the area at 1:50,000 covering 375,000 ha of the upper Cauca River valley where most sugarcane in Colombia is grown. CENICAÑA pedologists used expert opinion to provide more detailed maps. Meteorological data are available from 34 government and private weather stations. Using these, potential evapo-transpiration and water balance were calculated and mapped as humidity groups. For a more

detailed description of these studies see: Quintero and Castilla (1992); Carbonell et al. (2001); Isaacs et al. (2004) and (2007); Torres et al. (2004); and Cock et al. (2011).

The data on sugarcane is tied to what are called harvest events. A harvest event for sugarcane is the harvest of a particular plot at a specific moment in time. When harvest conditions and timing are similar, crop performance is also thought to be similar leading to the notion of environmental and temporal homogeneity. By definition, an Agro-Ecological Zone (AEZ), is a homogeneous area that shares similar environmental characteristics that influence crop responses, so that differences between crop performances within an AEZ must be due to the timing of the harvest event (Carbonell et al., 2001; Liu and Samal, 2002; Isaacs et al., 2007; Cock et al., 2011). In an attempt to obtain information on harvest events, production data was provided by the sugar mills for each plot.

2.3.2 ANDEAN BLACKBERRY AND LULO

In the case of sugarcane, the sugar industry maintained records of the environmental conditions and production of sugarcane at individual plot level. This was later on compiled into a centralized CENICAÑA database. This situation contrasts with that of the Andean fruit crops. The climatic and soil conditions of Andean blackberry and lulo individual plots were not available when this research began. Furthermore, the vast majority of Andean fruit growers simply did not maintain records on the production of their crops and how they managed them.

Farmers observe how their crop responds to the way they manage it and how it interacts with the environment. An individual farmer has only a narrow range of experience from which it is not possible to generalize. But, by applying the method of operational and participatory research, data describing a large number of harvest events within a range of management and growing conditions, can be used to develop data-driven models that will provide insights in the production system (Jiménez et al., 2009). Hence, in order to apply the operational research principles to Andean blackberry and lulo, it was necessary to establish systems to: (a) collect information on cropping events (b) characterize the growing conditions; and (c) create and compile the information collected (data management). The information collected was then used to model Andean blackberry and lulo yield.

As in sugarcane, harvest events for Andean blackberry and lulo coincide with the intervals between one harvest and the next. In contrast with sugarcane, however, there are no pre-defined homogeneous AEZs that can be used to determine which variables control crop yield. It was

necessary, therefore, to associate the productivity of each production site with its environment and the farmers' management practices using data from as many sites as feasible.

2.3.2.1 COLLECTION OF INFORMATION ON CROPPING EVENTS

In order to record information on the production of each plot planted to Andean blackberries and lulo, the author, together with researchers at *Corporación Biotec* and local Andean blackberry and lulo producers, developed a guide form based on a calendar, which was used by the farmers to record information. Data recorded on forms included a description of each plot, its location, crop species and variety or eco-type (see Appendix A2), events or harvesting experiences, and some management practices which were registered on the guide form (see Figures 2.6, 2.7 and Appendix A2).

Farmers' management information includes choice of the variety they plant. This was recorded along with other management practices, such as planting date and plant spacing. In Colombia, commercial varieties of both Andean blackberry and lulo are either thorned or thornless, with no further differentiation. There is no further genetic information available. Standard traditional practices for these crops are mainly control of pests and diseases such as *Botritis* (*Botrytis cinerea*) and *Perla de la tierra* (*Eurhizococcus colombianus*) for Andean blackberry, and (*Phytophthora infestans*) and *Picudo de la flor* (*Anthonomus* spp.) for lulo (Franco et al., 2002; Franco and Giraldo, 2002; BIOTEC, 2007).

Figure 2.6. Training farmers to use the guides and calendars; left: using the guide form for Andean blackberry; centre: using the guide for lulo; right: recording information on a production site

Figure 2.7. Farmers transporting guide form to their farms

2.3.2.2 SUMMARY OF FARMER PARTICIPATION

A total of 186 small-scale farmers provided information via calendar forms. Of these, 89 recorded production data, but not all of them recorded all the information required (location of each plot, variety, management practices). Variety and number of plants was recorded by 77 growers, but only 41 of them recorded soil information to give complete datasets. Based on these 41, it was possible to characterise 742 cropping events, 488 for Andean blackberry and 254 for lulo (Table 2.2). It was not possible to collect information on specific management factors like outbreaks of pests and diseases, or application of pesticides or fertilizers. Farmers in the Nariño department shared more information than growers in other departments, probably because there are grower groups in Nariño that meet frequently to share experiences

2.3.2.3 COLLECTION OF INFORMATION ON GROWING CONDITIONS

2.3.2.3.1 SOIL INFORMATION

There are neither soil maps nor expert description of the soils where Andean blackberry and lulo are grown. Moreover, most farmers have neither the knowledge nor the resources to correctly classify their soils. Nevertheless, they do know how important soil is in determining what crops to grow and how to manage them. Due to the complexity of the mountainous terrain in which these crops are grown, there are no soil maps of the study areas at a scale large enough to be useful (O'Brien, 2004). For example, the scale of the FAO world soil map (FAO, 1974) is 1:5,000,000, which does not represent the heterogeneity of this complex terrain. Soil characteristics in the Andean regions of Colombia vary widely at very local scale, and there is usually no direct correlation between soil type and other soil features such as pH and texture (Läderach, 2009).

There is a need for a simple, easy-to-learn methodology, based on scientific and locally generated knowledge, for farmers to characterize their soils and terrain (Alvarez et al., 2004).

The *Centro Internacional de Agricultura Tropical* (CIAT) and the *Universidad Nacional Sede Palmira* developed a Rapid Soil and Terrain Assessment (RASTA) tool, which is a simple *in situ* methodology for farmers (Alvarez et al., 2004). RASTA was used in the present research to determine soil and terrain characteristics for the Andean blackberry and lulo sites. RASTA is downloadable from http://www.frutisitio.org/wp-content/uploads/2011/02/RASTA-2011.pdf (Appendix A1).

RASTA measures: (a) basic soil characteristics that can be assessed directly in the field, and (b) infers a number of other soil properties (Table 2.1). Each of the inferred traits can be obtained through standard methods (Danielson and Sutherland, 1986). RASTA mostly estimates physical soil properties such as slope, stoniness, and mottling, which change less with time compared to chemical properties that change with each fertilizer application or as nutrients are extracted by harvested crops. RASTA gives farmers a tool that they can use to assess their soils conditions without the need for expert evaluation. During the course of the research, Andean blackberry and lulo growers were trained to use RASTA (Figure 2.8), and the information they collected was processed in the subsequent analysis.

Figure 2.8. The author training farmers on the use of RASTA. Left, training farmers how to determine soil texture, right how to use the guide

Table 2.1. Soil characteristics estimated in RASTA

Basic characteristics	Inferred traits
Land form	Organic matter content+
pH (acidity or alkalinity)	Internal drainage+
Texture	External drainage+
Structure	Effective soil depth+
Hardpans	Salinity
Presence of carbonates	Sodicity
Rocky or stoniness	

+ Defined in Alvarez et al. (2004) (Appendix A1)

It was not possible to include soil pH or the presence of carbonates in the final analysis because the materials required to unequivocally measure them (indicator paper, acid) are neither easy to obtain in the area, nor easy to manipulate.

Table 2.2. Summary of the number of Andean blackberry and lulo growers who recorded information via calendars

Crop	***Departments***	***GPS***	***Cropping events***	***Production***	***Variety and number of plants***	***RASTA***	***Complete plots***
		No of farms	weekly periods	No of farms	No of farms	No of farms	No of farms
Andean blackberry	Caldas, Nariño	75	488	35	34	20	20
Lulo	Nariño, Others	111	254	54	43	21	21
Total		**186**	**742**	**89**	**77**	**41**	**41**

2.3.2.3.2 ENVIRONMENTAL INFORMATION

Climate and weather conditions in the Andean hillsides are highly variable. Often, the only data available comes from a weather station 30 km or more distant, which in these heterogeneous, mountainous landscapes, usually bears little relation to the reality of the specific farmer's site. In order to characterize individual sites, we used information from multiple sources described below.

New geospatial information on natural resources has recently become available on a global scale, including high-resolution topography (SRTM, Farr and Kobrick, 2000) and climate data (WorldClim, Hijmans et al., 2005). These databases are either available as raster maps, which represent continuous layers or grids. Maps are divided into equal-sized cells (pixels), each of which contains

a single value of the factor mapped. For raster data, the term resolution is used rather than scale (O'Brien, 2004; Läderach, 2009). Resolution is the size of one pixel and is commonly given in arc degrees. The distance represented by one degree of longitude varies with latitude, and is about 111 km at the equator. A 30 arc-second resolution is about 1 km at the equator, while 3 arc-second resolutions is approximately 90 m at the equator (Läderach, 2009).

The Shuttle Radar Topography Mission (SRTM) database contains high-resolution topographical and landscape information (Farr and Kobrick, 2000). This database is a high resolution terrain model at 90 m spatial resolution (downscaled to 30 m). The SRTM is a project between the national aeronautics and space administration and the national geospatial-intelligence agency. The database is available at http://srtm.csi.cgiar.org. SRTM uses radar interferometry to obtain digital topographic data for 80% of the Earth's land surface, at CIAT the missing data in the primary coverage were filled with secondary data (Jarvis et al., 2006; Läderach, 2009).

In order to provide researchers with useful information on climate variability, Hijmans et al. (2005) developed interpolated climate surfaces for global land areas at a spatial resolution of 30 arc-seconds (often referred to as 1-km spatial resolution) in the WorldClim database. To do so, they compiled monthly averages of climate as measured at weather stations from a large number of global, regional, national, and local sources. Most of the data cover a period of 50 years (1950–2000) and were interpolated them using “a thin-plate smoothing spline algorithm”. WorldClim contains monthly precipitation and mean, minimum and maximum temperatures. The data can be downloaded from http://www.worldclim.org

Another source of climate information is Tropical Rainfall Measuring Mission (TRMM) from which estimates of monthly average rainfall can be extracted based on the model developed by Bell (1987). The TRMM satellite was launched in November 1997 and continues in operation. The dataset corresponds to the original satellite snapshot views and is available at http:// http://mirador.gsfc.nasa.gov/cgi-bin/mirador/presentNavigation.pl?tree=project&project=TRMM. We used the TR3b42 dataset at a resolution of 10 arc-minutes (18-km). The TRMM data is an integral of atmospheric humidity, not actual rainfall, so they are called either precipitation estimates or precipitable water (Huffman et al., 1995; Kummerow et al., 1998). We call them precipitable water, which are estimates of the likely rainfall over a particular period of time. Hence using this measure gave an estimate of the actual rainfall at a given site over a given period of time. This contrasts with WorldClim which gives long term averages of rainfall at a particular time of year at a particular site. In this study, we extracted monthly data for the two-year period January 2006 to December 2007.

Production sites were geo-referenced to allow extraction of environmental data of harvesting events from the publicly-available environmental databases (see Figure 4.3 and Tables 4.1 and 5.1). The environmental data was included for the period of yield formation, which encompassed the climate when pests and diseases attack. Harvest month, first, second and third month before harvest for Andean blackberry and harvest month, first, and second month before harvest in the case of lulo. Variables that we extracted and generated from the SRTM, WorldClim and TRMM databases are summarised in Table 2.3.

Table 2.3. Environmental factors used in the present study and obtained from publicly-available environmental databases

Variable	Database Source	Units
Precipitable water - daily rainfall	TRMM	mm
Monthly total precipitation	WORLDCLIM	mm
Monthly average temperature	WORLDCLIM	mm
Monthly minimum temperature	WORLDCLIM	°C
Monthly maximum temperature	WORLDCLIM	°C
Temperature range	WORLDCLIM	°C
Altitude	SRTM	MASM
Slope	SRTM	°

2.3.2.3.3 CURRENT MODELLING SOFTWARE BASED ON PUBLICLY-AVAILABLE CLIMATIC DATA

Software packages such as Floramap (Jones and Gladkov, 2003), DIVA (Hijmans et al., 2005b), and Homologue (Jones et al., 2005) use publicly-available climatic databases to estimate the suitability of particular crops to specific climatic conditions. Floramap, DIVA, and Homologue require agronomic knowledge of the particular crop species for the user to be sure that their predictions of the suitability of a particular site are correct. As this knowledge is available for neither Andean blackberry nor lulo, the statistical approach used by these packages is inadequate.

Floramap, DIVA, and Homologue all use a pixel size of 10 arc-minutes (about 18 km at the equator), which, however, is too large to be useful in environmentally highly heterogeneous zones like the Andes where Andean blackberry and lulo grow. Therefore, modelling tools that integrate information of soils and can explain yield variability with little prior knowledge of the crop are alternatives to the statistical approach used by these packages.

2.4 DATA MANAGEMENT

CENICAÑA developed a specialized software (SEGUITEC) to link the industry's sugarcane data with the centralized CENICAÑA database (Isaacs, 1999). The software contains information on yield and sugar concentration collected over the last 20 years, and also includes weather and soil

data. This allows the compilation of a large number of cropping events over a wide range of conditions.

To apply the same methodology to Andean blackberry and lulo, a number of datasets related to harvesting events over a range of conditions is required. We therefore created a database in MS Access containing information collected by Andean blackberry and lulo farmers via calendars of production events. The database included a description of each production plot, GPS coordinates (latitude and longitude), and soil data registered via RASTA (Figure 2.9). As previously mentioned, production sites were geo-referenced in order to allow extraction of environmental data via publicly-available environmental databases (Tables 4.1 and 5.1).

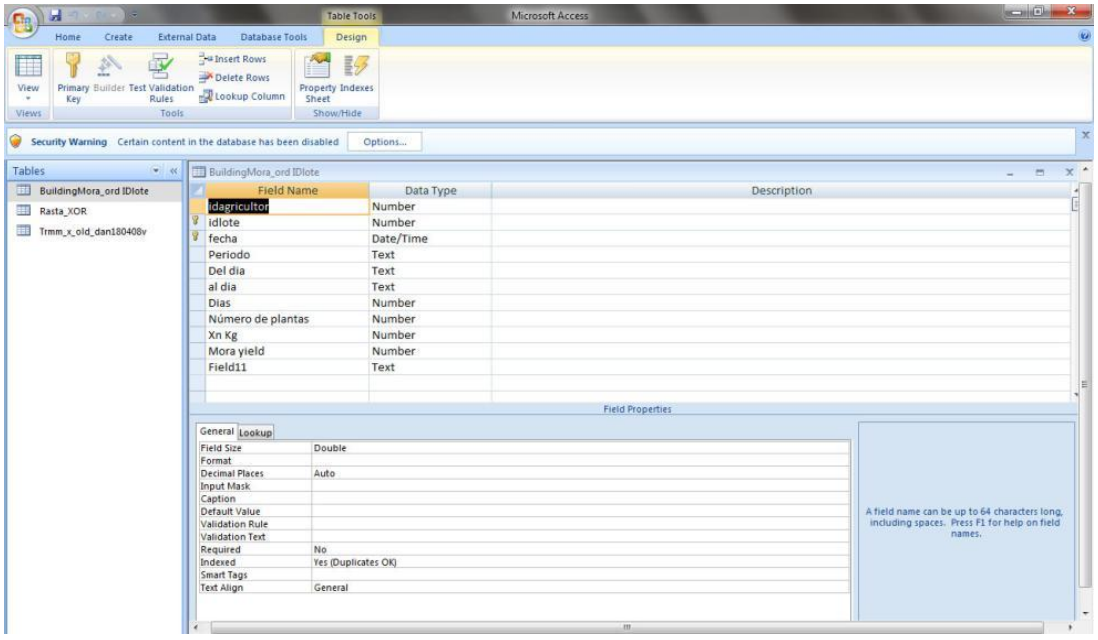


Figure 2.9. Database format used in the SSAFT project, using Microsoft Access (2003)

2.5 COMMENTS ON THE METHODOLOGIES

Sugarcane in Colombia has been researched in depth in contrast to Andean blackberry and lulo, which have received little attention. Nevertheless, the concept of harvest events, as previously defined, is common to all three, and provides an essential element for the construction of databases based on farmers' production experiences. In sugarcane, harvest events are ascribed to individual plots, which are harvested at 11–18 month intervals. In contrast, individual plants of Andean blackberry and lulo are harvested year-round. We used weekly periods as harvest events. This resulted in information on numerous cropping events that vary temporally and spatially.

The organization of the supply chain (see section 2.2) determines how data can be collected and managed. The supply chain for sugarcane is highly organized and autonomous and has its own

research centre. Detailed soil maps and climate information from the industry's own weather stations, coupled with expert opinion, allows the characterization of the variation in space and time of each harvest event.

Andean blackberry and lulo do not have strong grower associations, as a consequence data collection and management requires a different approach. First, we had to develop methodologies that farmers could use to record information on their crops and their soils. In order to guarantee that they would be user-friendly, they were developed in collaboration with farmers. The validity of this approach was confirmed when farmers themselves recorded 742 cropping events on 41 plots and used RASTA to characterize their soils (Table 2.2). In the case of environmental data, GPS coordinates were successfully used to extract data from Shuttle Radar Topography Mission (SRTM), Tropical Rainfall Measuring Mission (TRMM), and WorldClim (section 2.3.2.3.2), and thus associate climatic conditions with each cropping event. The major differences between the systems developed for sugarcane and under-researched crops are summarized in Table 2.4.

Table 2.4. Comparison of the process of collecting harvest events in sugarcane, Andean blackberry and lulo

Process of collecting cropping events	Well-researched crop (sugarcane)	Under-researched crops (Andean blackberry and lulo)
Supply chain organization	Collaboration between growers, mills and CENICAÑA research institute	Absent
Capture of soil data	1:50,000 soil maps	RASTA (captured by farmers)
Capture of climate data	Industry's own weather stations	Publicly-available environmental databases
Capture of cropping events	Data management systems and GIS database	Guide forms (filled by farmers)
Definition of homogeneous AEZs	Defined through detailed soil and climate data and expert opinion	Not defined
Data management	CENICAÑA software (SEGUITEC)	MS ACCESS database

3 DATA MANAGEMENT AND ANALYSIS - LEARNING FROM THE WELL-STUDIED DATABASE OF SUGARCANE IN COLOMBIA

3.1 INTRODUCTION

The well-studied sugarcane crop in Colombia contrasts greatly with Andean blackberry and lulo (see previous chapter). Nevertheless, cropping events are common to all three crops. In the case of sugarcane, harvest information is collected by the sugar mills for each production site. In Andean blackberry and lulo, this information was generated through the use of calendar forms with information recorded by farmers themselves. As previously mentioned, one of the objectives of the present research is to demonstrate whether**,** despite the limited knowledge on both species in Colombia, it is possible to use a SSCP approach for these crops in order to provide insights into how yields vary with variations in the environment.

The first component of our research is to use farmers' observations of crop response under commercial production conditions. In Colombia, this approach has been successfully applied in well-studied crops such as sugarcane and coffee (Isaacs et al., 2007; Niederhauser et al., 2008; Cock et al., 2011). The second component is to use information obtained from farmers' observations on cropping events, combined with tools presented in the previous chapter to construct models that can be used to explain yield variability in Andean blackberry and lulo. The experience in sugarcane and coffee suggested that the most effective way to model farmers' production experiences is to create clusters of events which occur in similar environmental conditions. In the coffee and sugarcane studies, these clusters are defined with the aid of expert opinion. Once relatively homogeneous clusters are established, the effects of variation within and between clusters can be analysed (Isaacs et al., 2007; Cock et al., 2011).

In the case of sugarcane, there is a wealth of knowledge about the crop. This information exists amongst others as a result of the high degree of organization of the supply chain, expertise of agronomists and producers, and also many years of research (see chapter 2). This knowledge provided (a) the wherewithal to define relatively homogeneous clusters of agro-ecological zones, and (b) insights about the functional relationship between sugarcane production and many of the factors considered likely to influence sugarcane production.

In the case of Andean blackberry and lulo, there is neither sufficient knowledge to define agro-ecological zones in which similar environmental characteristics influence crop responses, nor is there detailed knowledge of the functional relationships between processes that can be linked to crop performance. Therefore, and in order to define clusters with similar environmental conditions,

a different approach was needed. In this research, the strategy used by the electronic engineers, who formed part of the multidisciplinary research team, outlined in chapter 1, developed methodologies to identify “climatic agro-ecozones” (Barreto, 2012).

A non-supervised artificial neural network approach showed that zones, even though geographically distant, could be clustered using this technique (Barreto, 2012). Meteorologists and sugarcane experts at CENICAÑA corroborated these results and pointed out that the climate clusters defined in that study, as far as sugarcane growth and development are concerned, are climatically similar. Hence, it was concluded that the technique can be effectively applied to cluster environmental conditions (Barreto and Pérez-Uribe, 2007; Barreto, 2012; Satizábal et al., 2012).

The neural network approach is particularly powerful when little is known about the relationships between factors (Schultz et al., 2000; Sargent, 2001; Paul and Munkvold, 2005). In the case of sugarcane it was known, for instance, that yield is not a linear function of the age of the crop since planting or the last ratoon. Therefore, yield prediction models for sugarcane included both linear and quadratic terms for yield of sugarcane as a function of the age of the crop (Cock et al., 2011). In the particular case of sugarcane yield versus crop age, there was sufficient agronomic knowledge to indicate the likely non-linear function that could approximate reality. However, in some cases (for well-researched crops and sugarcane), and in most cases (with under-researched crops for which no formal production models exist), there is little information on the exact nature of the non-linear response to variables. Many responses of crops to both management and environment are strongly non-linear: for example, the response of almost all crops to temperature is positive up to a certain point and negative beyond that optimum (Pollock, 1990).

This lack of knowledge about the true form of the response indicates the necessity of exploring approaches that make no assumptions about the functional respones in crops that have neither been studied in depth nor systematically backed up by knowledge of experts. Hence the need to explore novel methods to address SSCP development in information-poor systems, such as the under-researched crops studied in this thesis.

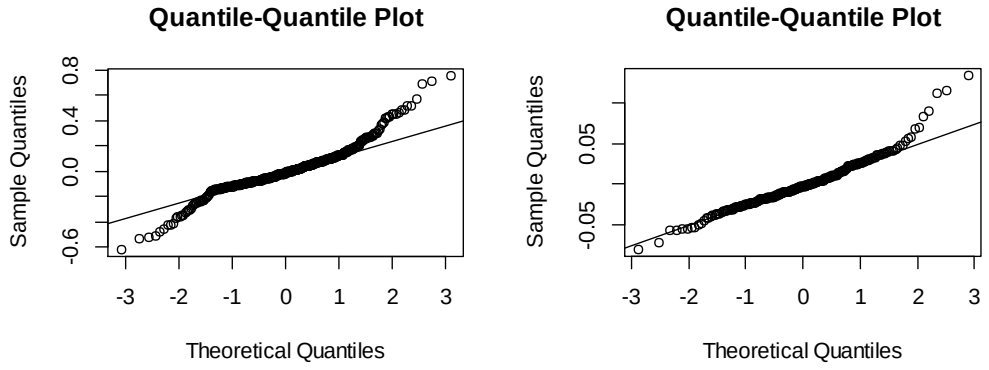
In the particular case of Andean blackberry, we used data collected by farmers on calendar forms, through RASTA and from publicly-available environmental databases to explore the relationships between crop yield and the recorded parameters. The geographical location of the site where Andean blackberry was produced had an important effect on yield (see chapter 4). We hypothesised that a geographical location effect could either be due to (a) specific environmental conditions in different locations; (b) cultural differences of the producers; or (c) a combination of both.

Based on this hypothesis and the experience in sugarcane and coffee, we decided for lulo, to use an iterative procedure based mostly on non-parametric approaches. The procedures first identify the most relevant factors linked to lulo yield; second, uses this information to define clusters having similar environmental conditions, and third, analyses the effect on yield of the environment, locations, and farms. The most relevant factors associated with lulo yield were identified through a combination of robust regression and non-linear artificial neural network. The relatively homogeneous clusters of environmental conditions were defined with a non-supervised neural network and finally a mixed model with best linear unbiased prediction was used to provide a quantitative estimate of the effects on yield of the environmental clusters, locations and farms which were treated as categorical variables (see chapter 5). The location and farms were used as proxies for management effects at both the regional level (locality) and within the regions (farm).

The decision to use non-parametric analytical methods was supported by a first analysis of the data structures, which indicated lack of normality (figures 4.2 and 5.2). For Andean blackberry and lulo, Shapiro-Wilk and Anderson-Darling were applied as procedures to test normality (Razali et al., 2011). The resulting *p* values for both tests were lower than 0.05 (Table 3.1) which indicates that there is not sufficient evidence to confirm a normal distribution of residuals. Both datasets were highly skewed and heavy tailed (Figures 3.1a and 3.1b). Parametric approaches are not generally considered appropriate for datasets that lack normality, and are highly skewed and heavy tailed.

Table 3.1. Normality tests applied to Andean blackberry and lulo datasets

Test	***Andean blackberry***	***Lulo***
Shapiro-Wilk	*p*-value = 1.48 e-13	*p*-value = 1.60e-07
Anderson-Darling	*p*-value = 2.2 e-16	*p*-value = 2.04e-05



(a) (b)

Figure 3.1. Normal quantile-quantile plot of the residuals for (a) Andean blackberry, and (b) lulo. Both datasets display heavy tailed data

In the study conducted for lulo (see chapter 5), in an attempt to identify the factors associated with yield that would then be used to determine the clusters, we used both non-parametric and parametric methods. In spite of the reservations on the applicability of parametric models to this dataset, the most relevant factors identified by the two methods were in general agreement, and the results were used in the non-parametric unsupervised neural network to determine the environmental clusters. The information of the clusters provided by the unsupervised neural network, together with the categorical variables location and farm were incorporated into a mixed model. Thus, in order to use the latter model, as it was suggested by a statistician, taking into account that it was evident that each farm was independent of the others, this new dataset should also show homogeneity of variance. Figure 3.2 shows that dispersion of residuals is around zero, and therefore there was no evidence of heteroscedasticity. Hence following the statement made by Foody (1999) about data distribution in agricultural systems, mixed models with best linear unbiased prediction were used to analyse the data within a hierarchical framework of clusters of environmental conditions, locations and farms. This mixed model with best linear unbiased estimated effects within and between environmental clusters (see chapter 5).

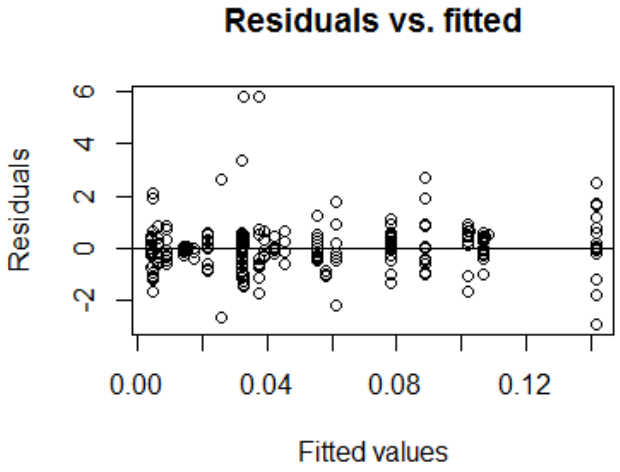


Figure 3.2. Plot of residuals against the fitted values. Residuals exhibit homogeneity of variance

Within the available non-parametric models, we chose non-linear approaches based on neural networks. We did this, not only because they are non-parametric, but also because they have proven capabilities for being able to manage noisy, incomplete, and heterogeneous, data. Furthermore, they can handle datasets where there is neither prior knowledge about data distributions nor information on possible mechanisms or functional responses to variation (Pérez-Uribe, 1998; Peña-Reyes, 2002; Barreto, 2012; Satizábal et al., 2012).

In order to evaluate these approaches, and as part of the HEIG-VD research team outlined in the introduction; the candidate provided agronomic expertise to build “non-linear supervised” and “non-linear unsupervised” models using CENICAÑA’s sugarcane database. This experience with the sugarcane databases provided the basis for applying these advanced modelling techniques to analyse productivity in Andean blackberry and lulo in highly heterogeneous conditions of both environment and management.

This chapter provides the rationale behind the use of ANNs. As part of the learning process with sugarcane, we surveyed literature on supervised and unsupervised modelling approaches based on artificial neural networks (ANNs), the results of which were published as a book chapter (Jiménez et al., 2008). Then, the sugarcane database was analysed to identify the most suitable variables to use in models and facilitate data interpretation (Barreto et al., 2007; Jiménez et al., 2007). The latter models were adapted and applied to the data collected through the methods described in chapter 2 for the two under-researched crops, e.g. Andean blackberry and lulo, (chapters 4 and 5) and were published in international journals (Jiménez et al., 2009; Jiménez et al., 2011).

3.2 A SURVEY OF ARTIFICIAL NEURAL NETWORK-BASED MODELLING IN AGRICULTURE

Adapted from: **Jiménez, D.,** Pérez-Uribe, A., Satizábal, H.F., Barreto, M., Van Damme, P. and Tomassini, M. (2008). A survey of artificial neural network-based modelling in agroecology. In: B. Prasad. (Ed.), Soft computing applications in industry. Springer-Verlag Berlin Heidelberg, pp. 247-269.

Agricultural systems are difficult to model because of their high complexity and non-linear dynamic behaviour (Basso et al., 2001). The evolution of such systems depends on a large number of ill-understood processes that vary in time, and whose relationships are often highly non-linear and very often unknown. According to Schultz et al. (2000) there are two major problems when dealing with modelling agro-ecological systems. On the one hand, there is an absence of equipment able to capture information in an accurate way, whereas on the other hand there is a lack of knowledge about such systems. Researchers are thus required to build models in both well-studied and under-studied situations, by integrating different data sources, even if this data is noisy, incomplete, and imprecise, as is the case of the crops studied in this work.

When modelling an agricultural system, we can proceed by considering the modelling problem as either a regression or a classification problem. For instance, we deal with a regression problem

when modelling natural processes such as crop yield, climate and physiological variables, vegetation dynamics, greenhouse conditions, severity of a given pest and/or disease, etc., given that the dependent variables are continuous (Philip and Joseph, 2003; Kaul et al., 2005; Chung Lu et al., 2006). On the other hand, when dealing with a classification problem, we want to model phenomena such as environmental variability, yield quality and quantity, genetic variation, soil properties, land cover, etc., given that the dependent variables of the system are categories, and that the main research idea consists of assigning the same class to individuals with similar features (e.g., by forming groups) (Levine et al., 1996).

ANNs have been shown to represent a successful tool for modelling agricultural systems by considering the latter either as regression or classification problems (Hashimoto, 1997; Schultz and Wieland, 1997; Schultz et al., 2000). Thus, ANNs can be regarded as an alternative to traditional statistics, in particular when dealing with the highly variable, noisy, incomplete, imprecise and qualitative nature of agricultural information. Such techniques have been shown to be capable of "learning" non-linear situations using qualitative and quantitative information. In general, they have shown better pattern recognition capabilities than traditional linear approaches (Murase, 2000; Schultz et al., 2000; Noble and Tribou, 2007). During the last twenty years, researchers have acquired a lot of experience using artificial neural network-based models.

This section presents a survey of artificial neural network modelling applications in agriculture, in an attempt to provide researchers in agriculture with insights about neural networks modelling in a language easily understandable to them.

3.2.1 ARTIFICIAL NEURAL NETWORKS (ANNS)

An ANN (Bishop, 1995) is a computational structure where many simple computational elements, called artificial neurons, perform a non-linear function of their inputs. Such computational units are massively interconnected and are able to model a system by means of a so-called training algorithm. This algorithm attempts to minimize an error measure that is computed in different ways depending on the specific technique used to adjust the connections (e.g. the learning algorithm).

In order to build such a model, the user collects a series of records or "examples", and presents them to the network with the aim of computing output values that might be real values when dealing with a regression problem or discrete outputs when dealing with a classification task. In an attempt to obtain the desired outputs, a training algorithm is used to adapt itself to a set of parameters, called synaptic weights (because they are supposed to model the synaptic connections between biological neurons). These parameters weigh the relative relevance of the

input variables at the level of each artificial neuron. Once a model is trained by this technique, the model is able to generalize its response to previously unknown input information (Hsieh, 2009).

There are two major approaches to train an ANN (e.g., to adapt its parameters): supervised and unsupervised learning.

In the supervised learning approach, specific examples of a target concept are given. The goal is to learn how to recognize members of a class or to build a regression model using the description attributes. In this case, the synaptic weights among neurons are adjusted in order to minimize the error between the known desired outputs and the actual output given by the neural network during the learning process. Objective functions in supervised learning algorithms have the form of error functions, which calculate the discrepancy between the actual output of the model, and the desired output taken from the dataset. The signals produced by these error functions are used to guide the adaptation of parameters of the model. The most widely used errors function is the Sum of Squared Error (SSE), (equation 3.1; Peña-Reyes, 2002; Satizábal, 2010);

$$SSE = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}\left(o_{k,n} - t_{k,n}\right)^2 \qquad (3.1).$$

In this function, *o* represents the actual output of the network, *t* represents the desired output, *n* runs for every observation in the dataset (*N* is the number of observations), and *k* runs for every output unit in the network (*c* is the number of outputs). The SSE function, being a squared subtraction, is minimized when every output of the network matches every desired output contained in the training set.

As far as the classification problem is concerned, the output of the model is a discrete value. In the case of regression problems, the objective becomes the approximation of a continuous target by using the input attributes (Peña-Reyes, 2002).

In the unsupervised learning approach, the set of examples is provided without any prior classification. The goal is to discover underlying regularities and patterns, most often by identifying clusters or subsets of similar examples. Training in this case consists of looking for a compressed representation of the examples collected (original data); the error is the difference between this representation of the original data and the original data (Bishop, 1995). In other words, there is no prior information about the categorization or labels of these examples. During the learning process, the system reproduces the distribution of the observations in the original dataset in a space with a

reduced dimensionality. Said in another way, the model generates groups over the input patterns based on correlations or similarities between examples (Hilera and Martínez, 1995; Peña-Reyes, 2002; Satizábal, 2010). Modelling an input distribution with artificial neural networks is performed by placing, e.g. artificial neurons, in specific places of the input space in such a way that the positions of the artificial neurons represent some properties of the observation distribution used in the process of training the model. Then, to measure how well the neurons reproduce the original dataset, a topology error is measured (Satizábal, 2010).

3.2.2 INFORMATION PROCESSED AND DATA PREPARATION

There are several types of information that can be obtained from agricultural or farming systems. On the one hand, information on temperature, precipitation, soil water content and yield are examples of variables with continuous representation. On the other hand, information on soil structure, texture, colour, quality, and many variables recorded on commercial crop production are likely to be categorical (e.g. weed control, land preparation practices, soils structure, and texture). All these types of information should be fully exploited in order to build a reliable model.

Because every neuron in a network represents real values, neural networks allow the direct use of these values. Conversely, neurons cannot generate categorical data directly. Thus, the latter type of data must be converted into real values in order to feed the models. This data format conversion is often called “binarization” and uses different coding schemes. A widely used coding scheme is the so-called “local encoding”, where a new binary variable is created for each category. Thus, each observation in the dataset is represented by the binary input of the original category converted to one.

Table 3.2 shows an example of how depending on the soil order to which the original data observation belongs, the observation is set to 1, whereas the rest of soil orders for this observation remain as 0.

Table 3.2. Conversion of categorical information on 5 soil orders into binary values by using “local encoding” (table elaborated by the author)

Original soil orders (category) in the dataset	Representation of the observations in the dataset used to feed the models according to their original soil order				
Observations	**Soil order**				
	Mollisol	**Vertisol**	**Entisol**	**Ultisol**	**Inceptisol**
Mollisol	1	0	0	0	0
Vertisol	0	1	0	0	0
Inceptisol	0	0	0	0	1
Mollisol	1	0	0	0	0
Inceptisol	0	0	0	0	1
Vertisol	0	1	0	0	0
Ultisol	0	0	0	1	0
Entisol	0	0	1	0	0

Yield prediction in terms of weight/area illustrates how a model could be fed with different types of information after converting categorical variables into numerical representations. In this case, the model can use climate information about, for example, temperature and precipitation, which are real values, whereas it can also use binarized data obtained from categorical values like soil taxonomy classes, structure, texture, or colour.

The more accurate data are, the more reliable the resulting model will be. Out-of-range values are considered as a source of error, therefore they should be removed from the dataset. The detection of these outliers can be guided by researchers, or by using statistical techniques. As an example, if it is known that the temperature of a certain area should be confined to values within a specific range, the researcher can detect out-of-range values by comparing the collected data with these limits and subsequently remove these outliers by hand.

Another phenomenon diminishing data accuracy and model reliability is data incompleteness (Satizábal et al., 2007). Several situations during data collection can result in missing values. For instance, meteorological stations could fail or have compatibility problems with more recently developed measurement equipment. Additionally, detection and deletion of out-of-range values reduces the amount of available data, increasing data incompleteness. In order to cope with this drawback, different approaches based on averaging or interpolations are widely used to estimate the values of the missing data. The use of clustering techniques based on artificial neural networks (Self-Organizing Maps) extends the possibilities of pre-processing data. Furthermore, researchers do not have prior information about the processes to be modelled. As a consequence, they do not

necessarily know the individual relative relevance of each variable. In practice, the data is standardized during the modelling process in order to facilitate a number of calculations and improve interpretability. Hence, different variables, having different ranges due to the diversity of sources of agricultural information, are transformed into a range that varies for example between -1 and 1.

3.2.3 NEURAL NETWORK APPROACHES

ANNs development arose from an attempt to simulate living/animal nervous systems (Figure 3.3) by combining many, simple computing elements (neurons) into a highly interconnected system and hoping that complex phenomena such as “intelligence” would emerge as the result of self-organization or learning (Sarle, 1994).

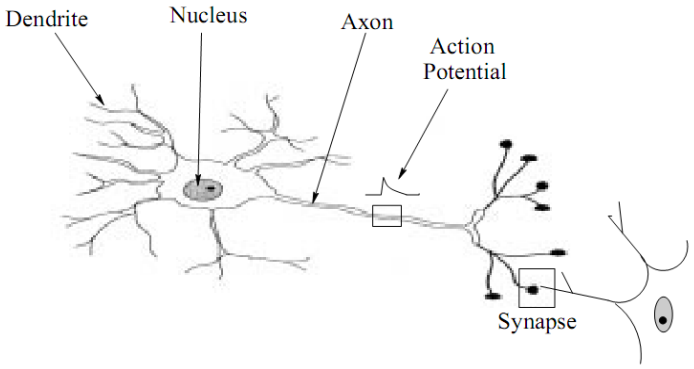


Figure 3.3. Biological neuron anatomy. Neurons receive signals via highly branched extensions, called dendrites and send information along unbranched extensions, called axons (Pérez-Uribe, 1998)

Neurons in artificial networks are linked together by a weighing value representing the synaptic connections displayed by the real neurons. Moreover, the processing units are organized in such a way that they form different layers of neurons. The way neurons are connected among layers determines the network topology (Figure 3.4). Hence, feed-forward networks are networks where information goes unidirectionally from inputs to outputs. In other words, the units of any layer are connected only to the units of a subsequent layer (Figure 3.4). In general, we use the term *recurrent networks* when there are also connections driving information to previous layers.

Among layered topologies, artificial neurons are organised in regular arrays called layers. The different neurons of the network perform different functionalities depending on the location of the layer they are placed in. Neurons belonging to the input layer transmit the input information into the network and thus feed the next layer of neurons. These following layers are called hidden layers. They are responsible for processing the information coming from the inputs by using non-linear

functions. Finally, neurons belonging to the output layer result in the model response by using the processed information coming from the preceding layers (Figure 3.4).

3.2.4 MULTILAYER PERCEPTRON

A Multilayer Perceptron (MLP) is an ANN with feed-forward topology and several hidden layers (Figure 3.4). In practice, however, only one, two or three hidden layers are used. The training process is carried out in a supervised manner whereby adjustment of model parameters can be done by means of numerous algorithms. In this respect, gradient descent strategies are the most widely used (Bishop, 1995).

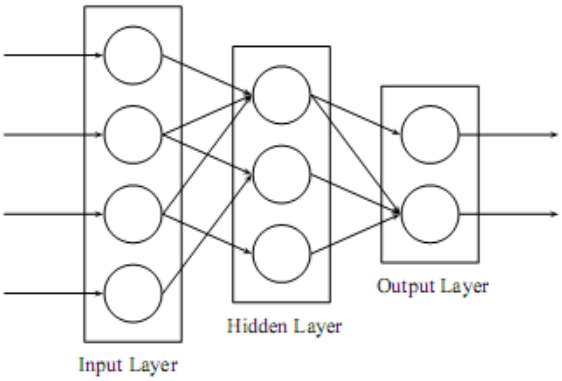


Figure 3.4. Schematic illustration of a three-layered feed-forward neural network, with one input, one hidden and one output layer. Circles (nodes) represent neurons, whilst lines stand for connections. Information flows from left to right (Satizábal, 2010)

During the training process, there is a pattern which indicates to the ANN the desired output as a function of input variables (training pattern). Training is carried out in order to minimize the error between the desired response and the predicted response by means of an optimization algorithm (Hsieh, 2009).

To compute the output of the feed-forward neural network, a node receives data from the previous layer and calculates a weighed sum of all its inputs. The equation is shown below

$$t_i = \sum_{j=1}^{n} WijXj \qquad (3.2).$$

where *n* is the number of inputs, *W* is the weight of the connection between nodes *i* and *j*, and *X* is the input from node *j*. A transfer function is then applied to the weighed value, *t*, to calculate the node output, *oi* (equation 3.3.):

$$o_i = f(t_i) \qquad (3.3).$$

The most commonly used transfer function is a sigmoidal function for the hidden and output layers, whereas a linear function is commonly used for the input and output layers (Kaul et al., 2005).

3.2.5 GRADIENT DESCENT ALGORITHMS – BACK-PROPAGATION

Gradient descent algorithms are optimization techniques employed to minimize a continuous function using information on the gradient. The gradient is a vector pointing in the direction where the evaluated function shows highest increase rate. Hence, it is possible to find a local minimum of the function by following the opposite of the direction of the gradient. The most commonly used output quantity is an error function which typically corresponds to the SSE of the output neurons (Pérez-Uribe, 1998). In the case of ANNs, the function to minimize is the difference between the target and the output of the model (supervised learning). One example of a gradient descent strategy is the Back-propagation algorithm (Rumelhart et al., 1986; Bishop, 1995) used to train MLPs in a supervised way.

3.2.6 INTERPRETATION

When the training process is successfully accomplished, the model describing the underlying process that generates the training data lies in the connections between the artificial neurons. At this stage, it is possible to write the equation describing the ANN. However, the resulting formula is hardly interpretable, and this is the reason why ANNs are considered to be black-box models (Ljung, 1999). Nevertheless, the information from the synaptic weights could be processed using strategies aimed at evaluating the relevance of each input in the model.

In agricultural systems, it is very important to identify the underlying relationships between input and output data in order to guide farmers in their decision making processes. ANNs, as black-box models, have to be scrutinized using methodologies that allow model interpretation. In the present research, we use input relevance and input profile plots in an attempt to get better insights in the process being modelled. To accomplish this, the relevance of the inputs is calculated by measuring the output sensitivity with respect to each input. As part of the SSATF project outlined in the introduction, at the HEIG-VD, six strategies for assessing the electiveness of relevance metrics were evaluated.

The efficiency of each one of the relevance metrics was tested and classified into two groups: metrics only using network parameters, and metrics using network parameters together with input patterns. It was found that the sensitivity matrix technique presents the best behaviour in most cases (equation 3.4; Satizábal and Pérez-Uribe 2007);

$$R_{ik} = \sum_{p \in patterns} \left|\frac{\partial o_k^p}{\partial I_i}\right| and \frac{\partial o_k^p}{\partial I_i} = \frac{\partial f_k}{\partial a_k^p} \sum_{j=0}^{nh} \left(w_{jk} \frac{\partial f_i}{\partial a_j^p} w_{ij}\right) \quad (3.4).$$

where nh is the number of hidden neurons, w_{ij} is the weight between input *i* and hidden neuron *j,* w_{jk} is the weight between hidden neuron *j* and output k, $\frac{\partial f_j}{\partial a_j^p}$ is the first derivative of the activation function of hidden neuron *j* and $\frac{\partial f_k}{\partial a_k^p}$ is the first derivative of the activation function of the output neuron (Bishop, 1995).

3.2.7 VALIDATION

The training process reduces the error between the output of the model and the target. However, when this process is finished, there is a need to assess the behaviour of the model using unknown input data. After having been trained, a neural network should be able to reproduce proper responses even when new input data is presented. This feature is called generalization. A low generalization is achieved when there is not enough training or when the neural network is overtrained. In an attempt to assess the generalization capabilities of the model, performance is tested over different validation datasets.

There are a number of strategies that can be followed to conduct this validation step. For large datasets, validation strategies that split the training set into several datasets are used. However, these approaches require datasets large enough to be split into new datasets. They are not recommended for small datasets such as the ones we obtained for Andean blackberry and lulo used here. Therefore, we conducted a split-sample validation.

In a split-sample or hold-out validation, training and validation datasets are created before training the neural network. In the case of Andean blackberry and lulo (as it will be shown in chapters 4 and 5), 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 may assess predictive model performance, but in the case of Andean blackberry and lulo, as they are small datasets, it is not recommended in its simplest form (Goutte, 1997). Nonetheless, the split-sample procedure can be improved for small datasets by repeating the split-sample procedure many times,

and by calculating the resulting performance as the average of all the tests made over the different validation subsets. The latter include methods for estimating generalization error such as: cross-validation, leave-one-out validation, and bootstrap validation (Jiménez et al., 2009; Satizábal, 2010). Models presenting a lower validation error are preferred (Bishop, 1995; Ripley, 1996; Hsieh, 2009).

3.2.8 SELF-ORGANIZING MAPS

A Self-Organizing Maps (SOM) or Kohonen maps (Kohonen, 1995), can be seen as data visualization techniques that reduce high-dimensional datasets through the use of a self-organizing clustering algorithm. One of the problems with data visualization is that humans cannot visualize high dimension representation of data. SOM techniques can be used to better understand high dimensional data by visualizing information in a low dimensional space (generally a grid of two dimensions) (Figure 3.5).

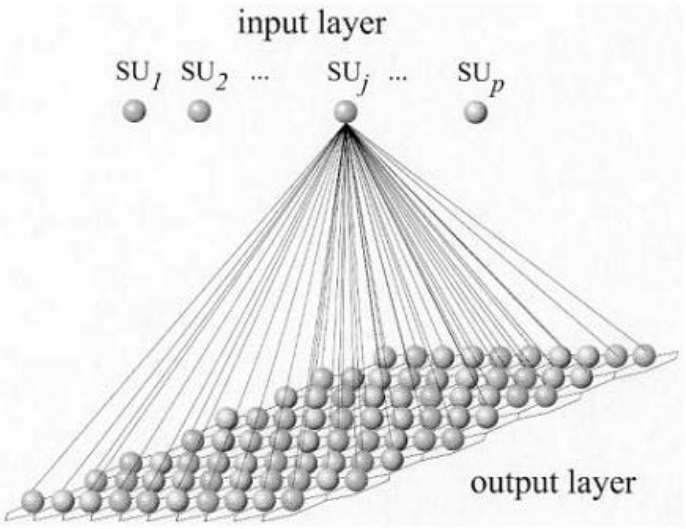


Figure 3.5. A two-dimensional SOM. Each sphere symbolizes each unit (neuron) at the high-dimensional input layer, and is mapped in a two-dimensional grid (output layer – Kohonen map) (Giraudel and Lek, 2001)

A SOM is formed by artificial neurons situated on a regular low-dimensional grid (Figure 3.6). This grid can be in one, two or more dimensions, but generally two are used. The neurons in the grid have rectangular or hexagonal forms. Each neuron i represents a n-dimensional prototype vector mi = [mi1,…, min], where n is equal to the dimension of the input space (Figure 3.7a). In the beginning of the training process, prototype vectors are initialized with random values. With each step of the training process, a data vector (observation) x from the input data is selected and presented to the SOM. The map's m_c unit closest to x is called: the Best-Matching Unit (BMU). The

BMU and its neighbouring prototype vectors on the grid are moved in the direction of the sample vector (equation 3.5):

$$m_i = m_i + \alpha(t)\, h_{ci}\,(t)(x - m_i) \qquad (3.5).$$

where $\alpha(t)$ is the learning rate and $h_{ci}\,(t)$ is a neighbourhood kernel centred on winner unit *c*.The learning rate and neighbourhood kernel radius decrease with time. Through iterative training, SOM organizes the neurons so that neurons that represent similar vectors in the input space are located on the map in contiguous zones. In this way, SOMs try to conserve the linear or non-linear relations of the input space (Barreto and Pérez-Uribe, 2007).

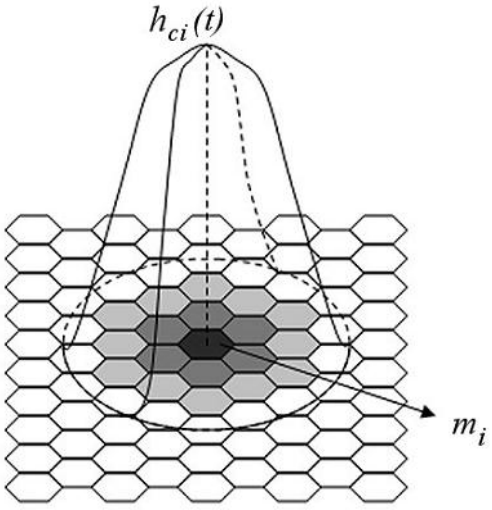


Figure 3.6. Self-Organizing Map with a hexagonal neuron lattice. The neighbourhood function h_{ci} (t) is centered on the best-matched neuron *mi*, which is shown as a black cell in the centre. Neighbouring neurons that have their weights recalculated by this best match are shown in gray surrounding the *mi* neuron. Other neurons are not affected

3.2.9 SOM AS A DATA EXPLORATION TOOL

Visualization tools are widely used for exploring and analysing data used in agro-ecological modelling. As far as the exploration of data is concerned, these tools allow an easy way to visualize the variables to be modelled, their dependencies, and their structure.

Traditional visualization techniques include the use of scatter plots to detect dependencies between variables, and the use of a scatter plot matrix, when these are more than two variables. In this latter technique, one generates a matrix, composed of several sub-plots where each variable is plotted against each of the other variables. However, in this technique, the number of pair-wise

scatter plots increases quadratically with the number of variables (Figure 3.7d) (Himberg, 1998). Therefore, this type of visualization is not practical in applications where we need to analyse many variables.

To improve the analysis of the dependencies between variables and/or their influence on the outputs of the system, it is possible to slice SOMs in order to visualize their so-called component planes (Kohonen, 1995) (Figures 3.7b and 3.7c). Nonetheless, like many clustering approaches, SOMs share the problem of deciding on boundaries for clusters. In order to address this problem, standard clustering methods are used to cluster pattern vectors (prototypes) of the SOM grid (Vellido et al., 1999). For this aim, the K-means algorithm is used to group prototypes into a given number of K clusters. However, one of the limitations of using K-means is that the number of clusters has to be defined before starting the analysis. To deal with this drawback, different K values are tested whereupon different groups with different number of clusters can be calculated. The optimal K number is then derived using the relative index of cluster validity known as Davies-Bouldin index (equation 3.6; Davies and Bouldin, 1979; Vesanto and Alhoniemi, 2000; Park et al., 2003);

$$DB = \frac{1}{n} \sum_{i=1, i \neq j}^{n} max \left(\frac{\sigma_i + \sigma_i}{d\,(c_i, c_j)} \right) \qquad (3.6).$$

where n is the number of clusters, σ_i is the average distance of all patterns in cluster i to their cluster centre c_i, σ_j is the average distance of all patterns in cluster j to their cluster centre c_j , and d (c_i, c_j) is the distance between cluster centres c_i and c_j. Small values of DB correspond to clusters that are compact, and whose centres are far away from each other. Consequently, the number of clusters that minimizes DB is taken as the optimal number of clusters.

Each component plane shows the relative distribution of one data vector component, e.g., each variable of the respective input variables. Unlike the traditional method where the number of pair-wise scatter plots increases quadratically with the number of variables, when using such SOM component plane-based visualization, the number of sub-plots grows linearly with the number of variables (Figures 3.7b and 3.7c). In addition, it is possible to cluster variables with a similar pattern. After plotting all component planes, relations between variables can be easily elucidated, as dependencies can be found by organizing the component planes in such a way that analogous planes are positioned near to each other. This method facilitates data visualization of input-input and input-output dependencies. An example of this technique with data of one of the under-researched crops under study is illustrated in chapter 4.

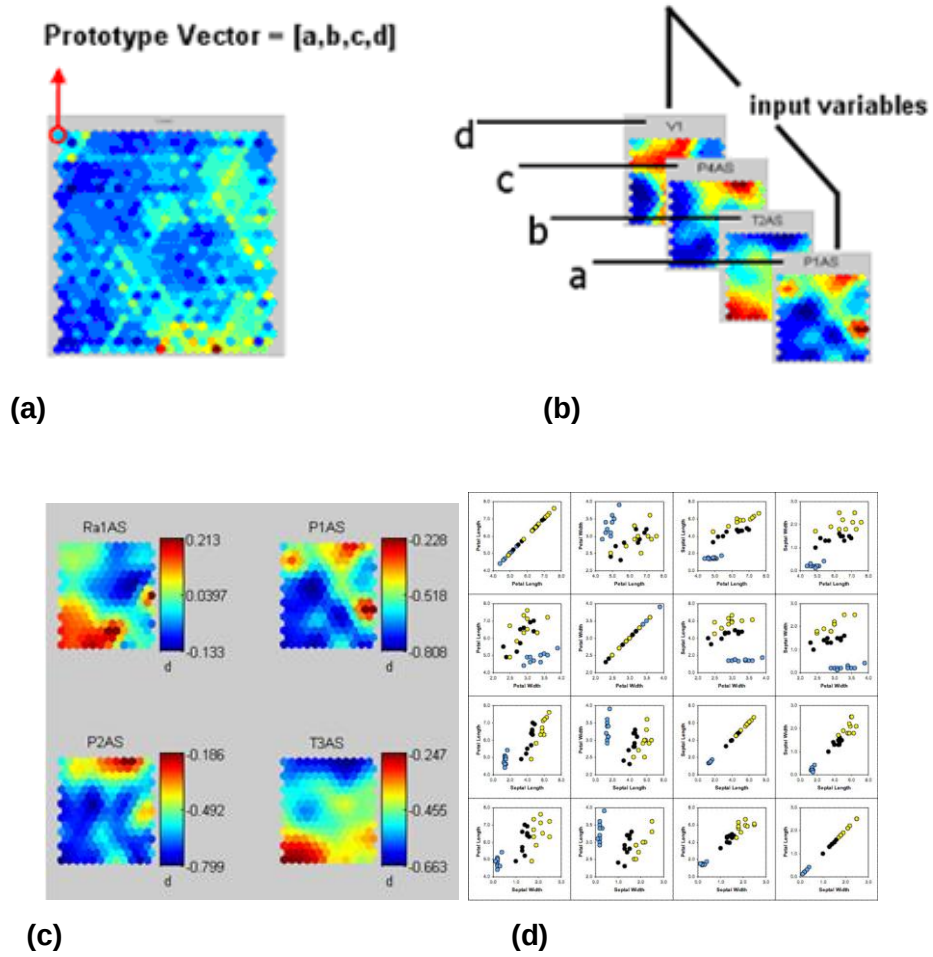


Figure 3.7. (a) Self-Organizing Map (SOM). Each neuron *i* represents a n-dimensional prototype, where n is equal to the dimension of the input space, in this case n = 4. (b) Component planes. It is possible to slice the Self-Organizing Maps in order to visualize their so-called component planes, where each component plane represents one input variable. (c) Component planes. Using component planes for analysis of relations between 4 variables, 4 plots are needed. (d) Scatter plot matrix. Using the scatter plot matrix for analysis of relations between 4 variables, 16 plots are needed

3.2.10 LITERATURE REVIEW OF APPLICATIONS OF ANNS IN AGRICULTURE

In this section, we present an extensive literature review of research and application articles reporting the uses of ANNs in agriculture. This section was first published as a book chapter (Jiménez et al., 2008). The studies surveyed provided us with a general overview of the exact nature of the applications and the context in which the respective studies were performed. Most of the articles surveyed, are part of the literature review made for the two articles published in agronomic peer-review journals and that are presented in this thesis. Likewise, links to these references can be found throughout the present document.

The literature review has been organized in such a way as to group the articles according to the application in agriculture.

The studies deal with grading fruits, weather forecast, weed control, pest and disease management, yield prediction, natural resource management, irrigation and fertilization, crop physiology, control of greenhouse environments, soils, and field operations and agro-industrial processes. In our review, we start by describing the application domain and then we give some details of the contribution via tables (see Table 3.3).

Table 3.3. Explanation of table content

Prob.	Particular problem that was developed/studied/addressed
Input	Data researchers used as inputs for the models
Alg.	Techniques and algorithms used
Res.	Main results of the research
Ref.	Literature references

3.2.10.1 GRADING FRUIT

Grading fruit is an important operation after harvesting. Fruit quality identification is commonly a manual task carried out by labourers based on their experience and empirical knowledge. However, these methodologies have some drawbacks such as inconsistency, cost, subjectivity and tediousness. Artificial neural network approaches have been used in order to overcome these drawbacks. In Table 3.4, we show some examples of how fruit can be graded using external or internal features.

Table 3.4. Applications of ANNs in grading fruits

Prob.	To grade apple colour
Input	Image data
Alg.	Improved Back-propagation
Res.	Successful classification of quality
Ref.	Nakano, 1997
Prob.	To grade oranges (*Miyauchi iyokan*) according to their acid or sugar content
Input	Data on colour, shape, roughness of skin surface, and weight
Alg.	The Kalman filter algorithm
Res.	Sweet (not the acid) oranges were those that were reddish, had low weight, were glossy and medium-sized
Ref.	Kondo et al., 2000
Prob.	To classify strawberry varieties (*Fragaria virginiana*)
Input	Different chemical signatures associated with growing year, place of production and fresh and frozen state of samples
Alg.	SOM
Res.	Growing year was one of the most relevant factors in differentiating varieties
Ref.	Boishebert et al., 2006
Prob.	To detect defects of sweet cherries *(Prunus avium)*
Input	Spectral information on different cherry tissues
Alg.	EGAN (feed-forward ANN)
Res.	Accuracy of 73% in classifying sweet cherry defects
Ref.	Guyer and Yang, 2000

3.2.10.2 WEATHER IN AGRICULTURE

Weather is one of the most important factors influencing agricultural systems. This factor influences many agro-ecological processes and variables, as it affects soil properties, pest and disease dynamics, agricultural practices, production, yield, etc. Variability in yield in rain-fed agricultural production systems can be attributed to the variability in weather conditions for more up to 90% (Hoogenboom et al., 2000). Therefore, it is desirable to develop accurate weather models capable of estimating climatic variability and its impact on agricultural systems or the environment, in an attempt to support decision making processes in agriculture. Table 3.5 provides some examples of weather modelling in agricultural systems.

Table 3.5. Applications of ANNs for weather conditions

Prediction of weather variables	
Prob.	To forecast solar radiation
Input	Daily values of radiation, daily temperature range, precipitation, cloudiness and relative sunshine duration
Alg.	Back-propagation
Res.	Modelling process using neural networks successfully estimated daily solar radiation
Ref.	Bocco et al., 2006
Prob.	To estimate daily maximum and minimum air temperature and total solar radiation
Input	Maximum air temperature, daily minimum air temperature, daily total solar radiation, difference in elevation, difference in directions and the straight line between the target and input station location
Alg.	Back-propagation, traditional spatial analysis
Res.	Neural network models were more accurate than other models in estimating maximum and minimum temperatures and solar radiation for a single location
Ref.	Li, 2002
Prob.	To analyse trends in rainfall over long periods
Input	Rainfall data corresponding to a period from 1893 to 1933
Alg.	Back-propagation-Adaptive Basis Function Network (ABFNN)
Res.	ABFNN performed better than Back-propagation in predicting long–term rainfall behaviour
Ref.	Philip and Joseph, 2003
Prob.	To estimate ozone concentrations
Input	Meteorological data
Alg.	SOM, two-stage neural networks, multiple linear regression, two-level clustering, MLP neural networks
Res.	Two-stage neural network had the best performance, explaining at least 60% of ozone concentration variance
Ref.	Chung Lu et al., 2006
Prob.	To estimate relative air humidity (testing MATLAB and STATISTICA software)
Input	Measurements of relative air humidity taken during 100 days in the year 1988
Alg.	Back-propagation
Res.	Neural models built with MATLAB estimated and predicted relative air humidity with higher accuracy than those created with STATISTICA.
Ref.	Białobrzewski, 2008

Forecasting temperature and critical duration periods affecting plant growth	
Prob.	Frost forecast in peaches (*Prunus persica*) during critical growth periods
Input	Temperature, relative humidity, rainfall, wind speed and solar activity
Alg.	Feed-forward neural network with different activation functions, trained with Back-propagation
Res.	Best frost prediction was achieved using relative humidity, solar activity and wind speed from 2 to 6 hours before the frost event
Ref.	Jain, 2003
Studying impact of climate change on the potential distribution of vegetation	
Prob.	To model vegetation distribution in past, present and future climates, in tropical forests
Input	Seven climate variables, nine soil parent material classes and seven terrain classes
Alg.	Back-propagation
Res.	Certain locations were occupied by a forest class in some climates while others continued occupying the same class despite changes in local climate
Ref.	Hilbert and Ostendorf, 2001
Prob.	To predict functional characteristics of ecosystems
Input	Data regarding six functional traits derived from the normalized difference vegetation index (NDVI)
Alg.	Back-propagation, regression models
Res.	Correlation between predicted and observed values for each functional trait was higher for the model developed with ANN than when using a regression model
Ref.	Paruelo and Tomasel, 1997

3.2.10.3 WEED CONTROL

Weed control is an essential activity in agriculture. Weeds compete with commercial crops in terms of space, nutrients, water and light. Weed control generally involves spraying herbicides which is an undesirable practice due their negative environmental impact, and because of the cost involved. Farmers and researchers have been interested in minimizing environmental impacts and reducing costs associated with weed control. Neural network classifiers have been shown to be a useful tool for discriminating between different kinds of weeds using databases of pictures.

A number of research activities have used this approach to develop a site-specific weed management system which helps farmers to spray herbicides in a selective way, thereby decreasing the amounts of these substances remaining in the crops and the environment subsequently decreasing costs. Table 3.6 presents several articles dealing with weed control; most of them focused on recognizing weeds in field situations. However, a particular case, where weed seeds were recognized in a crop seed selection process, is also shown.

Table 3.6 Applications of ANNs in weed control

Identification of weeds in field	
Prob.	To classify weed species in crop fields
Input	Image datasets of 33 texture features in a six class dataset of foxtail, crabgrass, common lamb's quarter, velvet leaf, and morning-glory, and clear soil surface.
Alg.	Back-propagation, Counter propagation, Radial Basis Function (RBF) networks
Res.	Feed-forward neural network trained with the Back-propagation algorithm had the best classification performance
Ref.	Burks et al., 2005
Prob.	To differentiate sunflowers (*Helianthus annuus*) from weed and soil
Input	Field images taken between two and three weeks after planting sunflower
Alg.	Back-propagation
Res.	Maximum number of correct differentiations of weeds from sunflower plants was 71 (out of 86), 82 in separating sunflower from bare soil and 74 in distinguishing weed images from bare soil
Ref.	Kavdır, 2004
Prob.	Site-specific herbicide applications using automatic weed localization
Input	Digital images of more than 30 common weeds present in the research area
Alg.	Neural networks and fuzzy logic
Res.	With this approach, site-specific weed densities could be determined and thus herbicide applications could be better targeted (managed, monitored)
Ref.	Yang et al., 2003
Prob.	To differentiate between weeds and seedlings of carrot (*Daucus carota*)
Input	Measures of leaf shape, digital images
Alg.	SOM
Res.	Neural networks discriminated species without predefined plant descriptions, however, this method required image processing
Ref.	Aitkenhead et al., 2003
Prob.	To differentiate weed species from corn (*Zea mays*) and sugar beet (*Beta vulgaris*) crops
Input	Spectral proprieties
Alg.	Back-propagation, probabilistic neural network, learning vector quantization, SOMs, local linear mappings
Res.	With local linear mapping, an identification accuracy of 90% was obtained
Ref.	Moshou et al., 2001 and 2002
Recognition of weed seeds in a crop seed discrimination process	
Prob.	To discriminate weed species present in commercial seed lots (in Argentina)
Input	Image dataset of morphological, colour and texture features
Alg.	Back-propagation
Res.	Morphology was the most important feature for identifying weed seeds
Ref.	Granitto et al., 2000

3.2.10.4 PEST AND DISEASE MANAGEMENT

Pest and disease management is an essential component of any agricultural management system. Pest and disease damages to crops are known to be important factors causing yield losses. These yield losses are commonly tackled with strategies which involve the use of a range of pesticides, thereby increasing both production costs and the danger of high toxic residue levels on agricultural

products. Correct pest and disease management requires an accurate identification of pests and diseases; it also requires knowledge about the organisms' life cycles, as well as the effect of environmental variables on pest and disease development and population dynamics. Some models based on ANNs have been developed to support these management practices, in an attempt to help decision makers in agriculture to avoid unnecessary pesticide applications, and to identify various phenomena leading to yield losses. In the following Table 3.7, we show some examples of predicting pests and diseases in agricultural systems, by using different sources of data (environmental, images, cultural, etc.).

Table 3.7. Applications of ANNs in pest and disease management

	Diseases
Prob.	To predict the severity of maize gray leaf spot (*Cercospora zeae-maydis*) in corn (*Zea mays*)
Input	Environmental, cultural, and location-specific data, temperature, relative humidity, surface wetness, cumulative hours of surface wetness and of daily temperature, cumulative hours of nightly relative humidity
Alg.	Back-propagation
Res.	Best variables predicting severity were hours of daily temperature and hours of nightly relative humidity
Ref.	Paul and Munkvold, 2005
Prob.	To compare and forecast diseases in wheat (*Triticum* spp.)
Input	Environmental data
Alg.	Back-propagation, logistic regression and multivariate linear discrimination
Res.	Neural network and statistical models showed similar performance
Ref.	Franck, 2004
Prob.	To detect the development of "yellow rust" (*Puccinia striiformis* f. sp. *tritici*,) in wheat (*Triticum* spp cv. Madrigal). To compare a hybrid neural network + spectral reflection method with a fluorescence remote sensing system
Input	Spectral images (wavebands)
Alg.	SOM
Res.	The waveband centered at 861 nm was the variable which best discriminated healthy from diseased leaves. The hybrid approach showed the best performance
Ref.	Moshou et al., 2004 and 2005
Prob.	To detect and classify Phalaenopsis (*Phalaenopsis* spp.) seedling diseases: bacterial soft rot (*Erwinia carotovora),* bacterial brown spot (*Burkholderia cattleyae)* and phytophthora black rot (*Phytophthora parasitica)*
Input	Images of 18 texture features of the lesion area and Red, Green, Blue (RGB) colour features
Alg.	Back-propagation
Res.	*Phalaenopsis* seedling diseases were successfully detected and classified
Ref.	Huang, 2007
Prob.	To identify Cucumber Green Mottle Mosaic and Tobacco Rattle Viruses
Input	Prototypes of virus reactions
Alg.	Back-propagation, genetic algorithms
Res.	The method proved its potential to identify CGMMV and TRV
Ref.	Glezakos et al., 2010
	Pests
Prob.	To predict pod borer pest (*Helicoverpa armigera*) attack on chickpea (*Cicer arietinum* L.)

Input	Data on climate, location and pest incidence: date, minimum and maximum temperature, humidity, rainfall, larvae**/**plant, eggs**/**plant, light and pheromone trap, location, season, area surveyed, plant protection type
Alg.	Bayesian regularization + Levenberg-Marquardt algorithm
Res.	Pest attack activities were successfully predicted for one week in advance, by using weather and pest surveillance data
Ref.	Gupta et al., 2000
Prob.	To predict the presence or absence of greater flamingo (*Phoenicoterus rubber roseus*) damages in rice fields in the Mediterranean
Input	Ecological variables of rice paddies
Alg.	Back-propagation
Res.	Neural networks successfully predicted flamingos incursions from a reduced set of ecological variables
Ref.	Tourenq et al., 1999

3.2.10.5 CROP YIELD PREDICTION AND OTHER ESTIMATIONS

It has long been accepted that farmer incomes increase or decrease depending on crop yields (Kaul et al.,2005). In an attempt to support farmers in their decision-making processes, it is important to understand the relationships between factors that result and explain crop yield. These factors are extremely complex in time and space. Success of management decisions lies in understanding the combined influence of soil, landscape, climate, genotype, and water availability on crop yield, as well as identifying the moments or variables that could be modified by farmers in pursuit of improving their crop yields. The search for modelling techniques capable of recognizing these influences is an essential first step into understanding and identifying these processes.

It has been shown that ANNs are a powerful tool to tackle these problems. ANNs have been used in predicting yield in crops such as: corn, sugar beet, soybean and winter wheat, using databases of environmental data, plant features, and hyperspectral images. ANNs have not only been used in crop yield prediction but also in predicting the volume of harvestable pine barks, exploring the contribution of weather and other variables to some properties in winter cereal, as well as estimating the concentrations of pollutants in grass plant species. The following table 3.8 presents a more detailed description of these research activities.

Table 3.8. Applications of ANNs in yield predictions and estimations

Corn (*Zea mays*) and soybean (*Glycine max*) yield prediction	
Prob.	To predict corn and soybean yields
Input	Data from different locations, on soil type and multiple combinations of monthly or weekly precipitation
Alg.	Back-propagation, multiple linear regression models
Res.	Neural network model showed the best yield prediction for both crops using data of weekly precipitation
Ref.	Kaul et al., 2005

Corn (*Zea mays*) yield prediction	
Prob.	To predict corn yield under spatial and temporal variations of land management and soil conditions
Input	Topographic, climatological and soil properties data
Alg.	Several variations of Back-propagation (standard, batch, momentum, weight decay, quickprop, and resilient Back-propagation), forward search stepwise multiple linear regression, and pursuit regression
Res.	Performance of neural networks and linear methods was similar
Ref.	Park et al., 2005
Prob.	To identify the most important factors influencing corn yield (quantity) and quality
Input	Data on soil (electrical conductivity, organic matter, pH, chemical elements availability), landscape (slope, elevation) and genetic seed hybrid characteristics
Alg.	Back-propagation
Res.	Genetic characteristics (seed hybrid) was the factor which best explained variability in corn quality and yield variability
Ref.	Miao et al., 2006
Prob.	To predict corn yield
Input	Hyperspectral images, vegetation indexes
Alg.	Back-propagation, stepwise multiple linear regression
Res.	Neural networks and stepwise multiple linear regression performed better than models based only on vegetation indexes
Ref.	Uno et al., 2005
Prob.	To predict spatial variability in corn yield
Input	Soil and landscape characteristics (fertility, elevation, electrical conductivity and satellite imagery)
Res.	Neural networks successfully predicted spatial yield variability using fertility, elevation, electrical conductivity and spectral satellite image features
Ref.	Shearer et al., 2000
Sugar beet (*Beta vulgaris*) yield prediction	
Prob.	To predict sugar beet yield
Input	Physical and chemical characteristics of plants
Alg.	Predictive modular network (PRENOM)
Res.	The PRENOM network performed better than previous prediction approaches in estimating sugar beet yield
Ref.	Kehagias et al.,1998
Wheat (*Triticum aestivum*) yield prediction	
Prob.	To predict dry land winter wheat (*Triticum aestivum* L.) grain yield by using topographic attributes
Input	Landscape topographic attributes – spatial coordinates
Alg.	Spatial analysis neural network algorithm, whereby a so-called “kernel function” accounts for the influence of neighbouring points in the input space
Res.	Spatial analysis neural network algorithm successfully estimated spatial patterns of dry land crop yield
Ref.	Green et al., 2007
Predicting the volume of pine (*Pinus brutia*) bark	
Prob.	To estimate volume of pine bark
Input	Measures of: outside bark diameter at breast height, inside bark diameter at breast height, bark volume of the tree
Alg.	Kalman’s learning algorithm for training and Cascade Correlation, non-linear regression approaches (logistic mode, Gompertz, Metcherlich, Morgan, Mercer, Florin, Verhulst)
Res.	Neural networks using the algorithms mentioned successfully estimated pine bark volume

Ref.	Diamantopoulou, 2005
Exploring the contribution of weather and other variables to some properties in winter cereal	
Prob.	To predict phenological development, matter increase and soil moisture in winter cereal stands (wheat, rye and barley)
Input	Meteorological data
Alg.	Back-propagation
Res.	Neural networks successfully predicted phenological development
Ref.	Schultz and Wieland, 1997
Estimating the concentrations of pollutants in grass plant species	
Prob.	To estimate deposition and accumulation of pollutants (lead) in a grass plant species (*Cynodon dactylon*)
Input	Vegetation height, wind velocity, height of buildings, distance to adjacent street, traffic volume
Alg.	Back-propagation, multiple linear regression and stepwise multiple linear regression
Res.	Distance to adjacent streets, density, and height of buildings were the most important variables influencing lead concentration; ANN approach was more accurate than regression models
Ref.	Dimopoulos et al., 1999

3.2.10.6 NATURAL RESOURCE MANAGEMENT

When studying natural resource management, it is important to know the different interactions that take place between organisms, the environment and the changes occurring in the ecosystem. For instance, it is particularly useful to assess changes in an organisms' distribution, and temporal change in abundance or composition, that enable us to plan strategies addressed to put in practice a more rational use of natural resources. Therefore, there is a need to develop approaches that support the decision making process in natural resource management. In the following table 3.9, there is a summary of a number of studies that model population dynamics in ecosystems. We also show a case of soybean variety identification in the area of preservation of genetic resources.

Table 3.9. Applications of ANNs in natural resource management

Prob.	To classify and ordinate groups of vegetation
Input	10 types of vegetation expressed as presence or absence
Alg.	SOM
Res.	SOM was a feasible tool in classifying (grouping similar samples) and ordering (arranging samples in an ordered manner) in ecology
Ref.	Foody, 1999
Prob.	To predict the presence of a honeysuckle (*Lonicera morrowi*)
Input	Physical site characteristics and soil data
Alg.	Back-propagation
Res.	Predictions of neural networks were closely matching field observations
Ref.	Deadman and Gimblett, 1997
Prob.	To predict richness and species composition in a tropical forest
Input	Satellite remote sensing data

Alg.	Back-propagation, radial basis function (RBF), generalized regression neural networks, Kohonen SOMs
Res.	It was possible to estimate both richness and species composition by using remotely sensed data
Ref.	Foody and Cutler, 2006
Prob.	To evaluate intra- and inter-specific variations using soybean leaflets (*Glycine max*)
Input	Leaf shape images from 38 varieties
Alg.	Hopfield model, simple perceptron
Res.	The model was a suitable tool in varietal discrimination of soybean leaflets
Ref.	Oide and Ninomiya, 2000

3.2.10.7 IRRIGATION AND FERTILIZATION

Water and plant nutrients are very important factors for generating crop yield, because they are limiting factors in plant growth. While irrigation supplies the water required by crops according to their hydric requirements, when any other natural water source is limited, fertilization provides plants with the necessary elements for their healthy growth. According to Raju et al. (2006), water resources are becoming scarce due to factors resulting from human activities that reduce water availability for irrigation. This problem is more important in developing countries, where the population is growing faster than elsewhere, and where there is more contamination of water resources. In pursuit of a solution to this problem, a more accurate irrigation plan is needed in order to optimize water use. According to Broner and Comstock (1997), traditional knowledge-based crop management expert systems include fertilization and irrigation, and knowledge from field experts and growers. This knowledge is commonly acquired from different regions, which may differ in climate, soils and crop. There is a necessity to "tune-up" these systems. ANNs are able to tackle this problem by building models exploiting training sets containing site-specific data. Table 3.10 summarizes two studies dealing with irrigation and fertilization modelling in agricultural systems.

Table 3.10. Applications of ANNs in irrigation and fertilization

	Fertilization
Prob.	To provide nitrogen fertilizer recommendations for growing malting barley
Input	Data of phosphorous and nitrogen recommendations generated by an expert system
Alg.	Back-propagation
Res.	The neural network adjusted itself to site-specific conditions with fewer than 5% site-specific patterns in the training set
Ref.	Broner and Comstock, 1997
	Irrigation
Prob.	To select among available alternatives for irrigation planning
Input	Labour, agricultural production, economic data
Alg.	SOM
Res.	SOM-based integrative approach, was a successful tool for modelling a multi-objective irrigation planning
Ref.	Raju et al., 2006

3.2.10.8 ECOPHYSIOLOGY

Plant physiology is the study of the functions, or physiology, of plants. It comprises the study of processes such as plant water relations and plant responses to different stimuli. On the one hand, evapotranspiration and transpiration constitute a part of the water relation processes. Evapotranspiration includes: evaporation of water from soil and transpiration. This process is an important component of the hydrology cycle and decision makers in agriculture should be able to estimate more accurately irrigation water requirements using information on evapotranspiration and other water balance factors in order to maximize yield. On the other hand, plant growth is a process that can be influenced by many physiological factors, as well as by many environmental stimuli. The following table 3.11 presents a series of research results that used ANNs to estimate evaporation and transpiration, as well as three studies of neural networks emulating plant growth.

Table 3.11. Applications of ANNs in ecophysiology

Evaporation	
Prob.	To estimate evapotranspiration
Input	Data of meteorological variables and the physical basis involved in the evapotranspiration process and estimates provided by empirical models
Alg.	Quickprop, empirical models (Hargreaves-Samani and Blaney-Criddle)
Res.	Neural networks showed their potential for accurately modelling evapotranspiration
Ref.	Arca et al., 2001
Emulating plant growth	
Prob.	To model plant growth by means of characterizing plant processes. Neural networks used to model transpiration process
Input	Air, canopy temperature, relative humidity, and plant type
Alg.	Back-propagation
Res.	Neural network model successfully modelled transpiration. However, it was necessary to identify certain plant physiology processes such as assimilation, allocation, and nutrient update for better modelling plant growth
Ref.	Zee and Bubenheim, 1997
Prob.	To analyse lettuce (*Lactuca sativa*) growth characteristics under reduced gravity
Input	Biomass, chlorophyll content, plant width, and height
Alg.	Back-propagation
Res.	Neural networks were a useful technique for modelling plant growth under reduced gravity
Ref.	Zaidi et al., 1999
Prob.	To estimate biomass growth in winter cereals
Input	Site information, real observations and measurements, and weather information
Alg.	Back-propagation
Res.	The best network generating the desired biomass estimations used as inputs information of: temperature, field capacity of site, sum of precipitation since sowing, sum of global radiation since sowing and soil moisture in the upper soil layer
Ref.	Schultz et al., 2000
Forecasting maturity of fruit	
Prob.	To forecast maturity index (MI) of green peas (*Pisum sativum*)
Input	Historical harvest information with weather and climate forecasts
Alg.	Back-propagation

Res.	The model allows to harvest peas closer to their ideal harvest moment
Ref.	Higgins et al., 2010

3.2.10.9 GREENHOUSE

Greenhouses production is an agronomic practice where plants are grown under more controlled conditions than in conventional open field agriculture. This practice has some advantages such as increase in crop yield, whereas indoor climate factors can be controlled, pests and diseases can be drastically reduced, plants are healthier. It is also an ecological choice because herbicides are not required given that there are normally no weeds, and because it uses less water than growing plants outdoors. To generate an adequate environment for plant growth is a challenge in greenhouse growing. To address this problem, it has become necessary to use models able to simulate and predict greenhouse environment behaviour. These models should be able to identify those indoor conditions that need specific managing, resulting in the most suitable environment for plant growth. In this manner, decision makers could improve their crop yields. Table 3.12, summarizes some examples of prediction and simulation of greenhouse environments using ANNs.

Table 3.12. Application of ANNs in greenhouse climate control

Prob.	To study the compensation of external climate disturbances on the basis of input-output linearization and decoupling, in the operation of ventilation and moisturizing of greenhouses
Input	Combination of biological and physical models
Alg.	Feedback-feed-forward linearization
Res.	ANN approach achieves input-output linearization and decoupling in the moisturizing and cooling of greenhouses
Ref.	Pasgianos et al., 2003
Prob.	To model greenhouse climate dynamics
Input	Outside weather conditions
Alg.	Bottleneck neural network in input reduction
Res.	Bottleneck neural network was useful to control greenhouses climate
Ref.	Seginer, 1997
Prob.	To optimize cultivation and storage of tomato (*Solanum lycopersicum*)
Input	Cultivation process: nutrient concentration of the growing solution. Storage optimization process: storage temperature
Alg.	Back-propagation, genetic algorithm
Res.	The expert system provided practically the same advice on strategy for cultivation and storage provided by a skilled grower
Ref.	Morimoto and Hashimoto, 2000
Prob.	To simulate and predict greenhouse environment at any moment in the production process
Input	Inside air temperature, humidity and carbon dioxide concentration values
Alg.	Physical models and black-box linear parametric models
Res.	The neural network was not an adequate technique to predict the inside climate
Ref.	Boaventura, 2003
Prob.	To identify lettuce (*Lactuca sativa*) growth and greenhouse temperature
Input	Values of daily averaged CO_2 concentrations

Alg.	NUFZY (a hybrid neurofuzzy approach), orthogonal least squares
Res.	NUFZY model proposed coupled with an orthogonal least squares training algorithm successfully predicted both lettuce growth and greenhouse temperature
Ref.	Tien and van Straten, 1998

3.2.10.10 SOILS

Soils provide plants with the physical support and essential elements for growing. Depending on chemical and physical characteristics, a given soil could either be suitable for healthy growth of plants or not. Understanding of phenomena associated with soil properties is particularly useful in making decisions about adequate management of environmental resources and improvement of productivity. Models simulating soil processes will help to understand important procedures and clarify problems related to agricultural activities. Several studies in soils have been done through neural networks. The following Table 3.13 presents some studies related to modelling soil processes such as: rainfall, run-off, soil temperature, soil water retention and pesticide concentrations, and on predicting chemical properties and classifying physical soil conditions.

Table 3.13. Applications of ANNs in soils

Classification of physical properties	
Prob.	To classify soil texture
Input	Combinations of different classifications of soil particles according to size, and other soil parameters
Alg.	Back-propagation
Res.	Neural networks using as inputs soil properties such as: silt, clay, and organic carbon content, had a soil structure classification accuracy of around 79% in the training and validation phases
Ref.	Levine et al., 1996
Prob.	To classify soil texture
Input	Satellite aerial remote sensing and soil structure data
Alg.	Back-propagation
Res.	Spectral radiance and information of the most relevant variables for soil texture classification
Ref.	Zhai et al., 2006
Prob.	To predict soil texture
Input	Soil maps combined with hydrographic parameters derived from a digital elevation model
Alg.	Levenberg–Marquardt optimization algorithm, Back-propagation
Res.	ANN model can be used in the areas where the model was calibrated
Ref.	Zhao et al., 2009
Classification of land cover	
Prob.	To outperform a traditional land cover classification, known as National Land Cover Data (NLCD)
Input	Visible bands (blue, green, and red) of satellite images, textural information
Alg.	Back-propagation, decision trees
Res.	Overall accuracy of the proposed method had better classification results than the National Land Cover Data classification
Ref.	Arellano, 2004

Modelling soil processes	
Soil temperature	
Prob.	To model soil temperature by testing different neural network topologies
Input	Universal Transverse Mercator (UTM) coordinates and elevation
Alg.	Levenberg-Marquardt, resilient and standard Back-propagation
Res.	Conjugate gradient, Levenberg-Marquardt, and resilient algorithms predicted soil temperature with a smaller error than Back-propagation. Best performance was achieved by a multi-layer perceptron with a single hidden layer
Ref.	Veronez et al., 2006
Rainfall / run-off	
Prob.	To model rainfall run-off by testing several traditional rainfall/run-off models
Input	Rainfall, historical seasonal and nearest neighbour information
Alg.	Gradient algorithm, traditional rainfall/run-off models (simple linear model, seasonally based linear perturbation and nearest neighbour linear perturbation model)
Res.	The neural network had higher efficiency values than traditional rainfall/run-off models
Ref.	Shamseldin, 1997
Prob.	To estimate soil erosion, dissolved P (DP) and NH_4–N concentrations of rainfall-runoff from a land application site
Input	Rainfall/run-off, pH, conductivity (EC)
Alg.	Back-propagation
Res.	The ANN models derived from measurements of rainfall/run-off, electrical conductivity, EC and pH provided reliable estimates of DP and NH_4–N concentrations
Ref.	Kim and Gilley, 2008
Soil water retention	
Prob.	To predict soil water retention by implementing three neural networks (A,B,C) varying in inputs and outputs
Input	Neural network A: data of topsoil, bulk density, organic matter, clay, silt and sand. Neural network B: metric potential (amount of work required to bring water into a soil from outside) Neural network C: soil structure
Alg.	Back-propagation
Res.	Neural network model A which used topsoil, bulk density, organic matter, clay, silt and sand as inputs had better prediction performance of soil water retention than the other models
Ref.	Koekoek and Booltink, 1999

3.2.10.11 FIELD OPERATIONS AND AGRO-INDUSTRIAL PROCESSES

Field operations carried out in agricultural systems, imply a large amount of capital investment and often have negative environmental implications. Field operations often use tractors and chemical applications which are not desirable in agricultural systems in terms of agricultural management and ecology. Sustainable agriculture deals with the preservation and management of natural resources such as soil, water and processes essential for maintaining adequate crop productivity. Sustainable agriculture can be regarded as a potential solution to reduce the negative environmental impact caused by field operations. The use of intelligent robot tractors or autonomous vehicles could be an alternative that could favour the principles of sustainable agriculture. These robot vehicles could outperform traditional field operations with regard to

efficiency, energy consumption and soil preservation. ANN together with other machine learning approaches have been applied in describing the motions of mobile robots and helped in guiding autonomous vehicle field operations. As far as agro-industrial processes are concerned, and when dealing with cereals, grain drying is an important process after harvesting. During this process, it is important to maintain seed quality and to avoid the development of diseases that can be promoted by humidity. There is a necessity to find methodologies to support decisions makers in agro-industrial process, but also to determine the moisture content in the drying process of any agricultural system.

Table 3.14. Applications of ANNs in field preparation/management operations and agro-industrial processes

Field operations	
Prob.	To create an optimal path for an agricultural mobile robot
Input	Information of location and velocity from a robot tested on an asphalt surface
Alg.	Back-propagation, genetic algorithm
Res.	The approach was suitable for finding an appropriate path. The neural network was accurate enough in simulating the robot path
Ref.	Noguchi and Terao, 1997
Prob.	To develop an intelligent vision system for autonomous vehicle field operations.
Input	Classification of crops and weeds
Alg.	Back-propagation, genetic algorithm, fuzzy logic
Res.	The mixture of techniques and algorithms was fully appropriate for guiding autonomous mobile robots used in precision agriculture.
Ref.	Noguchi et al., 1998
Agro-industrial processes	
Prob.	To determinate the relation between initial moisture of barley seed (*Hordeum vulgare*) to dried and physical parameters
Input	Air flow rate, inlet and outlet air temperatures and humidity levels
Alg.	Back-propagation (slightly modified for the specific research)
Res.	ANN was a viable technique for modelling the grain drying
Ref.	Farkas et al., 2000

3.3 BUILDING OF ANALYTICAL APPROACHES TO UNDERSTAND YIELD VARIABILITY

Adapted from:

Jiménez, D., Satizábal, H.F., Pérez-Uribe Andrés. (2007). Modelling Sugarcane Yield Using Artificial Neural Networks. In: Proc. of the 6th European Conference on Ecological Modelling (ECEM'07), Trieste, Italy, pp. 244-245.

And

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

3.3.1 METHODOLOGY

This part of the research focuses on (a) finding the most relevant variables needed to model sugarcane yield through a supervised approach (MLP trained Back-propagation algorithm) (Bishop, 1995); and (b) improving visualization of the input–input (dependent variables) and input-output (sugarcane yield) relationships by means of SOMs.

3.3.1.1 DATABASES

The sugarcane databases used in the following experiments are based on the type of information described in chapter 2.

Two different datasets were implemented. In the case of the experiment conducted to determine the variables that contribute most to predict tons of sugarcane per hectare, the dataset included information from the year 2005. Due to the incapability of the Back-propagation algorithm to process missing data, we used this year's data as they presented less missing data. Variables are described in Table 3.15.

As far as the SOM approach is concerned, observations over a total period of seven years (1999 to 2005) were considered as SOM can process datasets having missing values. This particularity of SOM also enabled us to additionally integrate more variables into this model (Table 3.16).

Sugarcane is harvested between 11 and 18 months after planting or ratooning depending on the variety or local practices. Experts at CENICAÑA pointed out that the most critical periods for sugarcane growth are initial and final stages of plant development. The first months are essential

for vegetative structure formation, whereas during the last months the plant accumulates most of its saccharose that will ultimately be processed from the harvested plants. Therefore, in both experiments, the period of five months before harvest and the first five months after the preceding harvest were taken into account. Hence models were trained with monthly averages and all variables were scaled between [-1,1] in order to allow a comparison in magnitude.

As far as the variables related to climate are concerned, only the data from a group of *x* months after sowing (denoted by xAS) and *y* months before harvest (denoted by yBH) were used. In this case study, five months after sowing and five months before harvest were taken into account. Thus, creating a set of ten variables for each climate variable. For instance, in the case of radiation the set is composed of: Ra1BH (radiation-the-first-month-before-harvest),..., Ra5BH (radiation-the-fifth-month-before-harvest) and Ra1AS (radiation-the-first-month-after sowing),..., Ra5AS (radiation-the-fifth-month-after-sowing).

The output target value of the MLP model was yield in tons of sugarcane per hectare (TCH). This variable, together with plant age, was provided by sugarcane mills. Water balance of the production zone was expressed in a scale from 0 to 8. This factor expresses the change in soil moisture, through the difference between water gains and water losses (equation 3.7). For a more detailed description of this procedure see Torres (1998).

$$\text{Change in soil moisture} = \text{water gains} - \text{water losses} \qquad (3.7).$$

Information of water gains and losses was provided by the network of pluviometers coupled with the estimated evapo-transpiration data.

Finally, the weather variables used were extracted from the 34 weather stations. A total of 42 variables and 861 events were included in the model.

Table 3.15. Variables considered in the identification of factors that contributed most to predict tons of sugarcane per hectare

Variable/input	Source	Unit	Classes	Total of inputs
Climate variables				
Monthly accumulated precipitation	Weather stations	mm	AS, BH*	10
Monthly average temperature	Weather stations	°C	AS, BH	10
Monthly average relative humidity	Weather stations	%	AS, BH	10
Monthly radiation average	Weather stations	cal/cm²/day	AS, BH	10
Other variables				
Plant age	Sugarcane mills	Month		1
Water balance	Regional maps	-	0 - 8	1
Output				**42**
TCH	Sugarcane mills	Ton/ hectare		

*In the climate group, acronym AS indicates months After Seeding, and BH months Before Harvest

In the case of the experiment aimed to the improve the visualization input–input and input-output dependencies by means of SOMs; variables related to soil and sugarcane variety were ordered using a presence/absence coding, where 1 represents presence and 0 absence. Water balance was used in a scale from 0 to 8 as aforementioned. As a result, total number of samples was 1328 (Figure 3.8) whereas the vector which defines an agro-ecological event is composed of 47 variables (Table 3.16).

Table 3.16. Variables included in the intelligent visualization experiment

Variable	Source	Acronym	Classes	Unit	Total Variables
Climate variables					
Monthly average temperature	Weather stations	T	AS, BH*		10
Monthly relative humidity	Weather stations	RH	AS, BH	%	10
Monthly average radiation	Weather stations	Ra	AS, BH	cal/cm²/day	10
Monthly precipitation	Weather stations	P	AS, BH	mm	10
Other variables					
Order	Soil maps	Ord	Ord1, Ord2, Ord3	-	1
Texture	Soil maps	Tex	-	-	1
Soil depth	Soil maps	Dee	-	-	1
Landscape	Soil maps	Ls	Ls1, Ls2, Ls3	-	1
Slope	Soil maps	Sl	-	-	1
Water balance	Regional maps	WB	0-8	-	1
Variety	Sugarcane mills	V	V1, V2, V3	-	1
Productivity	Sugarcane mills	P		Ton/ hectare	**47**

*In the climate group, acronym AS indicates months After Seeding, and BH months Before Harvest.
All climate variables and productivity are continuous, whereas other variables are categorical

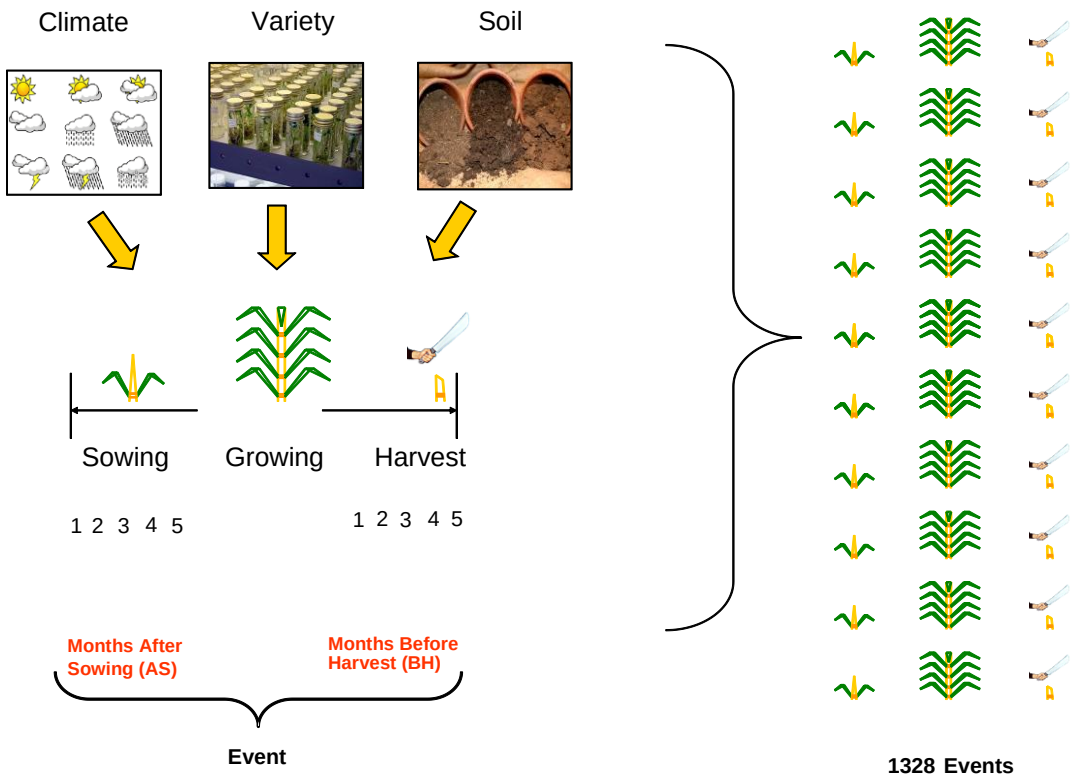


Figure 3.8. Illustration of the growth periods of sugarcane and the events included in the SOM

3.3.2 RESULTS

3.3.2.1 MLP (SUPERVISED APPROACH)

The best neural network configuration was found after varying the number of hidden neurons. Finally, it was a MLP with a single hidden layer of ten neurons which gave the best results. In order to obtain a more consistent model, a group of 100 ANNs was trained using the Back-propagation algorithm. A more detailed and graphic explanation of these procedures is provided in chapters 4 and 5 with the databases of Andean blackberry and lulo. In every of the 100 runs, the sensitivity matrix explained in chapter 3 was used to determine the factors that contribute most to predict tons of sugarcane per hectare. All the variables ordered by relevance are shown in Figure 3.9.

The analysis of input relevance shows that plant age (months) and water balance and are the most important variables for the construction of the model. Figure 3.9 also shows that in this respect, precipitation is the less relevant agro-ecological variable.

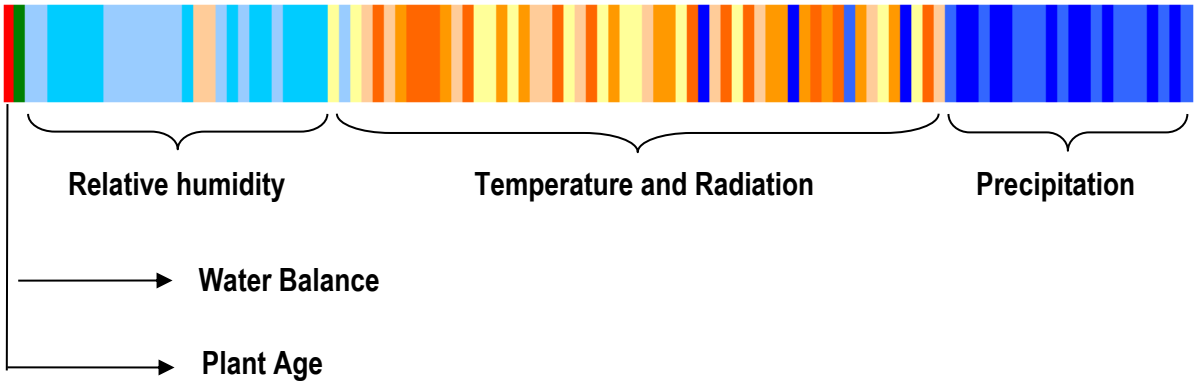


Figure 3.9. Input variables of the neural network-based model ordered by relevance (the most relevant variables are shown on the left)

Figures 3.10 and 3.11 show two profile graphics plotted using the two most relevant variables: plant age and water balance. Figure 3.10 shows that sugarcane plants gain weight with increasing plant age. According to experts at CENICAÑA this is logical if we consider that sugarcane is a plant that grows continuously until reaching a maximal height. Conversely, Figure 3.11 shows a decrease in weight with more water availability. In a previous work, plant age was also found as the most important variable when modelling sugarcane yield (CENICAÑA, 2006). Our analysis further identified water balance as a variable having strong yield modelling relevance. Indeed, this variable is involved in almost all processes we wanted to analyse (soil-plant-water-atmosphere relations).

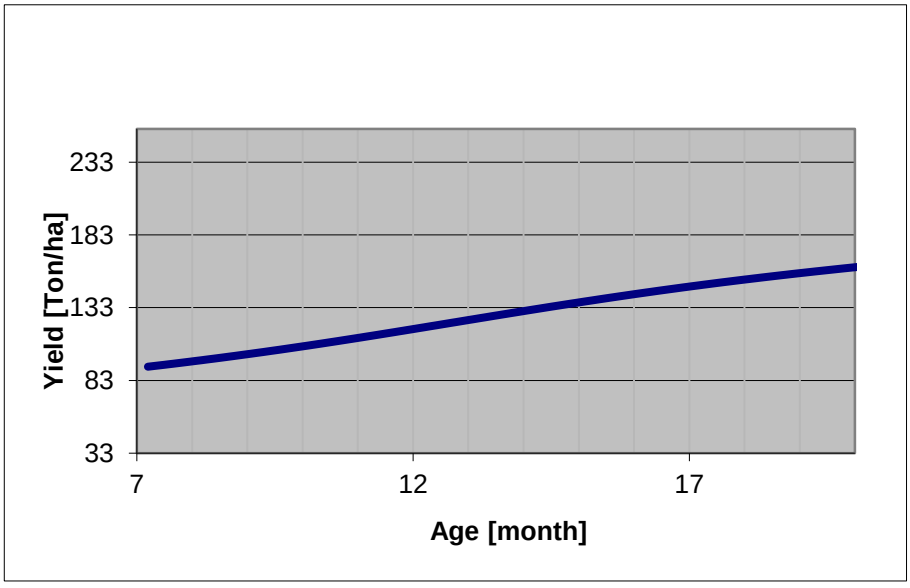


Figure 3.10. Input profile for the variable “plant age” using a MLP

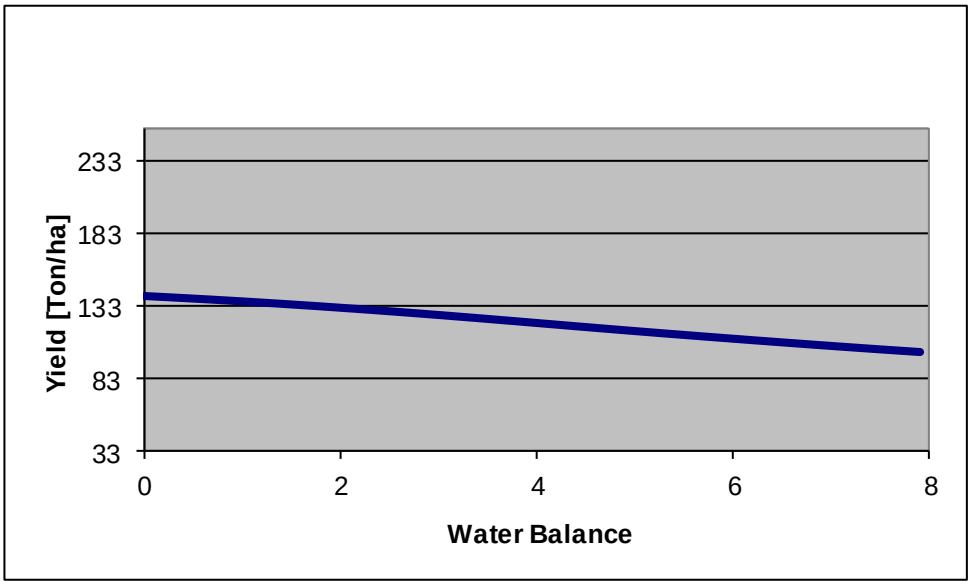


Figure 3.11. Input profile for the variable water balance using a MLP

3.3.2.2 INTELLIGENT VISUALIZATION (UNSUPERVISED APPROACH)

In the case of SOM, an input matrix with 1328 vectors was created, corresponding to each event, and its 47 associated agro-ecological variables. It should be noted that the output of this sugarcane model is productivity (see section 3.3.2.1.1). Nevertheless, in this approach productivity was used as input in order to obtain its component plane to be compared with the component planes of the agro-ecological variables. Vesanto (1999) suggested this technique in an attempt to find associations between inputs and outputs.

The matrix composed by events and agro-ecological variables was used to train a SOM of 400 neurons (20x20). Component planes were projected into a new SOM composed of 400 neurons (20x20). Finally, the tree-structured component planes representation was applied to the last SOM, so then, obtaining the structure shown in Figure 3.13

The SOM was sliced in order to visualize each component plane (additional information of this technique is provided in chapter 4), in an attempt to improve the analysis of the relationships between variables and/or their influence on the outputs of the system. Component planes show the relative distribution of each input variable (Kohonen, 1995). As mentioned previously, unlike the traditional approach where the number of pair-wise scatter plots increases quadratically with number of variables, using a SOM component planes-based visualization method, the number of sub-plots grows linearly with the number of variables (Himberg, 1998). In addition, it is possible to cluster variables with a similar pattern. After plotting all component planes, the relationship between variables can be easily interpreted.

The task of organizing similar component planes in order to find correlating components is called correlation hunting (Vesanto, 1999). However, when the number of components is large, it is difficult to determine which planes are similar to each other. Different techniques can be used to reorganize the component planes in order to perform this correlation hunting. The main idea is to place correlated components close to each other (Barreto and Pérez-Uribe, 2007).

With the aid of the tree-structured representation, it was possible to analyse planes' groups at several detail levels and to find relations between variables. The radiation during the first month after seeding (Ra1AS); radiation of the first month before harvest (Ra1BH); and presence of sugarcane variety two (V2) are more related to productivity than the other variables (Figure 3.13). In addition, a local correlation is observed between a majority of high radiation values and high productivity during these months (Figure 3.12).

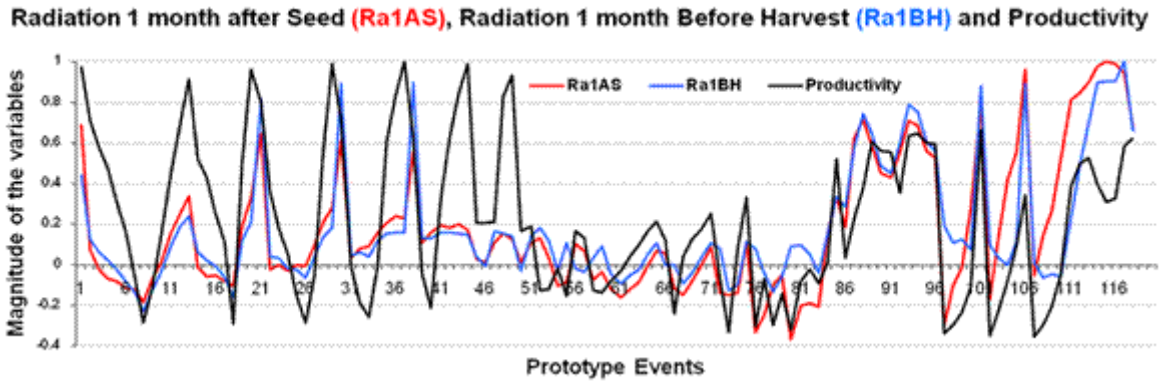


Figure 3.12. Lines displaying the patterns (prototypes) of the component planes: productivity, radiation of the first month after seeding (Ra1AS) and radiation of the first month before harvest (Ra1BH)

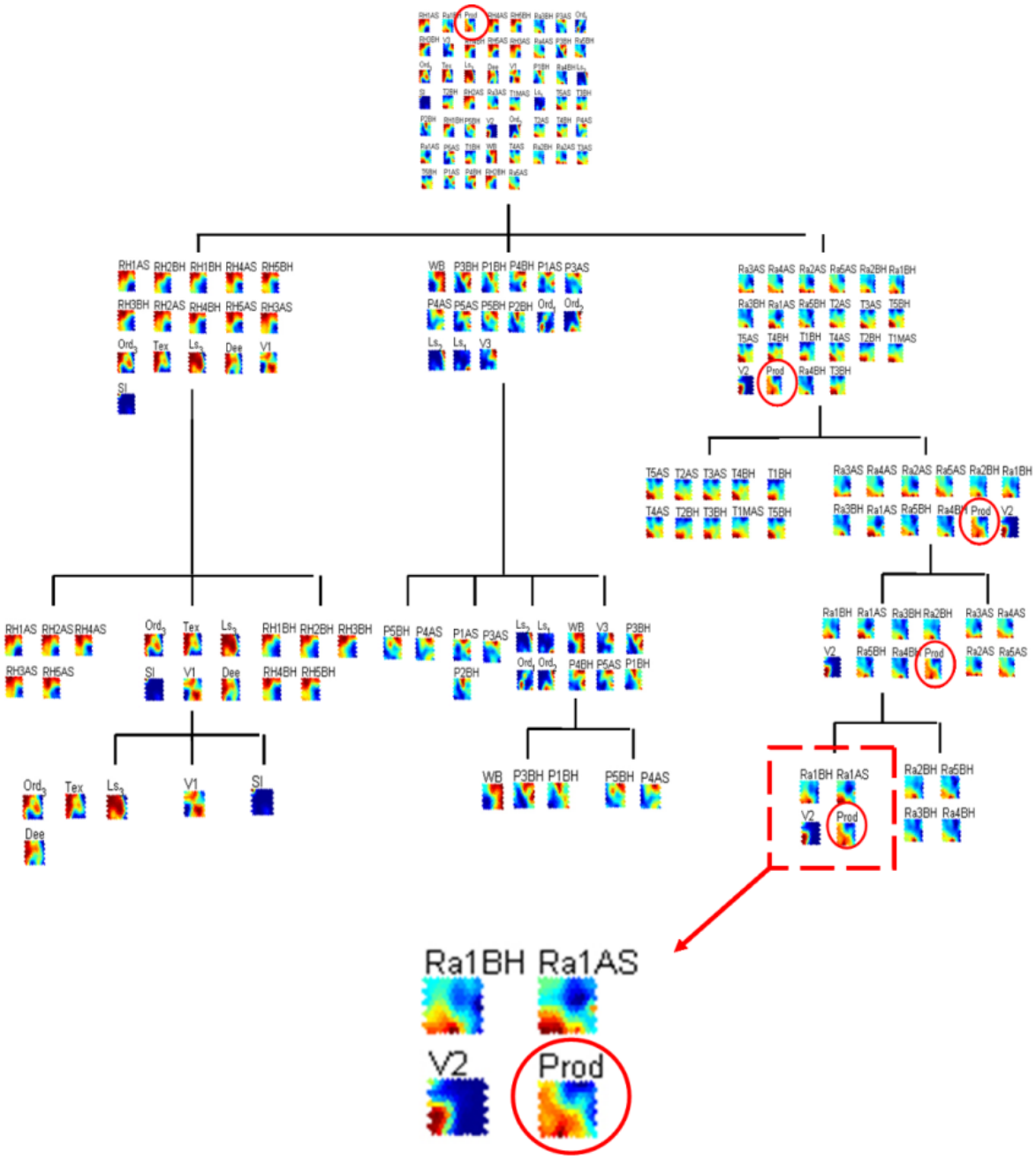


Figure 3.13. Tree-structured Self-Organizing Map component planes. Each plane represents a variable from the input space and its distribution before a self-organizing process is started. Red colours represent high variable values, blue colours low values. Radiation of the first month after seeding (Ra1AS), radiation of the first month before harvest (Ra1BH) and sugarcane variety two (V2) were most related to productivity

3.4 DISCUSSION AND CONCLUSIONS

According to the literature review of applications of ANNs in agriculture, ANNs have been used to solve diverse regression or classification modelling problems in many agricultural domains. In general, ANNs performed better than traditional approaches. In several case studies, however, hybrid approaches gave more accurate predictions (Tien and van Straten, 1998; Boaventura, 2003; Francl, 2004; Park et al., 2005; Uno et al., 2005).

Many data sources can be used to train ANNs to build an agricultural model, there were studies that used satellite images, pictures, weather and soil data, morphological descriptions, landscape characteristics, and land management information. An important characteristic of ANNs is their ability to use diverse data from multiple sources. The most frequent model uses simple feed-forward multi-layer perceptrons, trained with the Back-propagation algorithm or a variant of it.

Neural network models are criticised for their “black-box” nature, including the tendency to overtrain, the difficulty of interpreting relations between the inputs and outputs, and the need for large enough datasets for correct training (Schultz et al., 2000; Sargent, 2001; Paul and Munkvold, 2005).The resulting models are based on the interconnection weights between the neurons, which can be numerous, making it difficult to extract concrete conclusions. Thus, neural network models perform well on either classification or regression tasks, but it is not always clear how they use the input data to produce the outputs, which can complicate identifying the most relevant input variables. Nevertheless, the sensitivity matrix illustrated in section 3.2.6, based on the network parameters and input patterns, identified the variables that contributed most to predict sugarcane yield. The sensitivity matrix therefore provides tool to extract relations from neural network models.
SOM is also a promising tool for “intelligent visualization” of data and for data exploration. It needs fewer pre-processing steps, can estimate missing data, and can extract information from databases of growers’ production experiences. (Kohonen, 1995; Himberg, 1998; Vellido et al., 1999; Vesanto and Ahola, 1999; Aitkenhead et al., 2003; Moshou et al., 2004; Boishebert et al., 2006;Barreto and Pérez-Uribe, 2007; Barreto, 2012).

With sugarcane, sensitivity metric analyses showed that plant age and water balance were the most important variables for predicting yield (figures 3.9 to 3.11). SOM showed that variety and solar radiation during the first month after planting and during the last month before harvest were related to yield (figures 3.12 to 3.13). It is important to show that both results used different datasets. The unsupervised neural network, which was trained with 1999 to 2005 data (some of which were missing), with a more complete set of explanatory variables, found relations between radiation and yield. The supervised neural network was trained with climate data only for 2005

(because it is unable to process missing data) and did not identify radiation as controlling yield. In sugarcane, the non-supervised neural network was found to be a successful way to determine clusters with similar environments where there is insufficient data to define homogeneous clusters of agro-ecological zones (Barreto, 2012).

Previous studies conducted by CENICAÑA using cropping events in order to explain yield, indicated that the response of sugarcane crop to variation in growing environment or management is frequently non-linear (Cock et al., 2011).

In the case of sugarcane, expert knowledge provided modellers with guidelines for choosing the functions to analyse the non-linear yield responses to environmental and/or management variables. In the case of Andean blackberry and lulo there was no expert knowledge to provide guidelines for the functional response to be used in models that associate yield with variation in environmental or management parameters. Furthermore, there was no prior evidence to support the normality of the datasets. When normality tests were applied there was no evidence that for both datasets residuals were normally distributed (Figure 3.1).

The experience with the sugarcane databases provided the basis for applying these advanced modelling techniques to analyse productivity in Andean blackberry and lulo in highly heterogeneous conditions of both environment and management.

The feasibility of using both MLP and SOM to identify the most relevant variables, clustering, and visualization of input-input and input-output dependencies are illustrated in the next chapters. Databases were constructed using operational and participatory research methodologies, coupled with publicly-available environmental data for Andean blackberry and lulo. Both studies were published in high quality agronomic peer-review journals.

4 APPLYING MODELLING TECHNIQUES REFINED ON A SUGARCANE DATABASE TO IDENTIFY KEY PREDICTORS OF ANDEAN BLACKBERRY (*Rubus glaucus* Benth) YIELD

Adapted from**: Jiménez, D.,** Cock, J., Satizábal, F., Barreto, M., Pérez-Uribe, A., Jarvis, A. and 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 (2): 198–208.

ABSTRACT

Site-specific information recorded by small-scale producer groups of Andean blackberry (*Rubus glaucus* Benth.) on their production systems and soil characteristics coupled with publicly-available environmental data was used to develop models of these production systems. MLPs and SOMs were used as computational models in the identification and visualization of the most important variables for modelling production of Andean blackberry. ANNs were trained with information from 20 sites in Colombia where Andean blackberry is cultivated. MLPs predicted with a reasonable degree of accuracy the crop's production response. Soil depth, average temperature, external drainage, and 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 large differences in production could be assigned to management practices. The information obtained can be used to determine sites that are suitable for Andean blackberry production, and to propose management practices from sites with high productivity to sites with similar environmental conditions but which currently have lower levels of productivity.

4.1 INTRODUCTION

Research on Andean blackberry (*Rubus glaucus*) is limited. With the current levels of research intensity based on traditional plot experimentation varying individual factors that affect crop production, it is unlikely that technological packages can be developed to be used by growers Heterogeneous growing conditions and continuous production throughout the year of many tropical crops mean that a large number of experiments or treatments would be required to draw firm conclusions concerning the optimum management of these crops under their diverse production conditions. As it has been mentioned in the introduction (chapter 1), the situation of a tropical crop such as Andean blackberry contrasts strongly with that of, let us say, raspberries in a temperate

climate, where there is a well-defined harvest period and all management is focusing on generating optimal production in that period.

In tropical perennial crops that are harvested throughout the year, the number of possible combinations of management practices 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 continuous experimentation by producers of these crops, as it is the case for Andean blackberries.

Experience with sugarcane and coffee in Colombia has shown that by collecting farmers' production experiences generated with the naturally occurring variation in management and environment, crops' responses can be modelled (Isaacs et al., 2007; Niederhauser et al., 2008, Cock et al., 2011). Given the high degree of heterogeneity in growth conditions, the lack of detailed information, and the structure of the data shown in chapter 3 (Figure 3.1a), we opted for a data-driven modelling approach to provide information to growers on how to choose suitable sites for and to better manage their crops.

Crop models are basically of two types which can roughly be described as mechanistic simulation models, and best fit or statistical models.

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 observed variations in crop response to changes in 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 for most tropical fruit species, 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 growing conditions. 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. 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 determine 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; the latter 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 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.

Following on the premises of this research, (a) if it were possible to characterise production system in terms of management and environmental conditions; and (b) if we were able to collect information on the harvested product of a large number of harvesting events under varying conditions, and based on the approaches of operational and participatory research, it should be possible to develop best fit models for such production system. Hence, a 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 behaviour (Basso et al., 2001; Jiménez et al., 2008, Satizábal et al., 2012). Moreover, the available information describing these systems frequently includes both qualitative and quantitative data. The former are often difficult to include in traditional modelling approaches. We surmised that models based on ANNs, are an appropriate alternative for developing models that can be used to improve production systems.

ANNs have been successfully used to model agricultural systems (Hashimoto, 1997; Schultz and Wieland, 1997; Schultz et al., 2000). According to the survey presented in the preceding chapter, 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 with our Andean blackberry dataset.

ANNs do not require prior assumptions concerning data distribution or the form of the relationships between inputs and outputs (Sargent, 2001; Paul and Munkvold, 2005; Nagendra and Khare,

2006). In general, ANNs provide pattern recognition capabilities that are superior to 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 visualization of data, and facilitate data interpretation (Barreto et al., 2007).

Nevertheless, and as it has been mentioned in the conclusions of chapter 3, there are a number of disadvantages concerning the use of ANNs. Some of them are: its "black-box" nature, which makes it difficult to interpret relations between inputs and outputs; the difficulty of directly including knowledge of ecological processes, the tendency to overtrain; and the need for an adequate number of 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 collection of data that are required as inputs into the model and provide important insights about critical factors affecting yield.

Several studies identify the most relevant variables, and explain responses in agriculture through the use of MLPs. For instance, Miao et al. (2006) implemented a neural network for identifying the most important variables explaining 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 in corn quality and yield. In another study, Jain (2003) reported that the best frost prediction was obtained from relative humidity, solar activity and wind speed from 2 to 6 hours 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 daily temperature, nightly relative humidity, and nightly temperature. More recently, Jiménez et al. (2007) modelling sugarcane yield, suggested that crop age and water balance were most relevant factors for the modelling process.

SOMs have also been used to improve visualization of input-input and input-output dependencies. Thus, for example Moshou et al. (2004) found that a waveband centred 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). As another example, Boishebert et al. (2006) pointed out that growing year was an important factor in differentiating yield of strawberry varieties.

Extension officers, expert crop advisers and growers of Andean blackberry have reached a general consensus that optimum conditions for the crop are: soils with high organic matter content and a loamy texture, altitude between 1800 and 2400 Meters Above Sea Level (MASL); average relative humidity between 70 to 80 %, average temperature between 11 and 18 ° C, and 1500 and 2500 mm of rainfall per year (Franco and Giraldo, 2002).

The goal of this study was to demonstrate that collection of data from farmers' production experiences (small-scale commercial producers) of Andean blackberry and its analysis by means of ANNs can provide growers with useful information to increase their productivity.

4.2 MATERIALS AND METHODS

4.2.1 DATA COLLECTION AND COMPILATION

Corporación Biotec together with local Colombian Andean blackberry producers developed a simple aid based on a calendar (see chapter 2 and Appendix A2) which was used by farmers to record information on the production of each plot planted to blackberries on their farm. As farmers neither have the knowledge nor the wealth to assess their soil and terrain through formal methodologies, and as there is a lack of approaches, guidelines, books and field manuals for farmers or extension workers to characterise soils and terrain *in situ*, the RASTA system explained in chapter 2 was used to characterise soil conditions. Farmers were provided with RASTA kits and used these to characterise their soil and terrain (Alvarez et al., 2004) for 20 different sites in the departments of Nariño and Caldas (Colombia) (Figure 4.1).

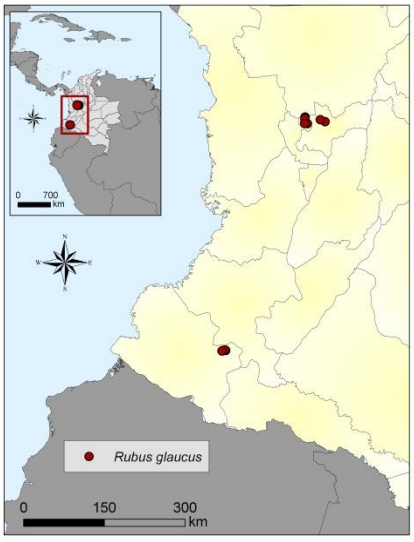


Figure 4.1. Map of the study area. Dots indicate the sites of production of Andean blackberry where data was collected

Information collected by farmers on the calendars and with RASTA was then transferred to the SSAFT project database illustrated in chapter 2. This database includes information on location, varieties, yield, and harvest time, and data on soil characteristics. A total of 488 yield records taken from the database were included in the analysis. These farmers' production experiences or "events" provided producers' estimates of the quantity (kg) of fruit harvested per plant for week (see figures 4.3 and Appendix A2).

Weather stations in Colombia are often not close to the fields where under-researched crops like Andean blackberry are grown (see chapter 2).Therefore, the information provided by these stations rarely represents the climate of individual production sites, largely due to the large variation in altitude in the region. Hence, environmental information on each site was obtained from the coordinates (latitude and longitude). As it was indicated in chapter 2, with this GPS information, it is possible to extract environmental data from publicly-available environmental databases (Tables 2.3 and 4.1) and to estimate the climatic conditions of any site that has been geo-referenced (see chapter 2).

Thus, topography and landscape data was extracted from the Shuttle Radar Topography Mission (SRTM) (Farr and Kobrick, 2000) using version 3 dataset available from Consortium for Spatial Information (CSI-CGIAR). Long-term averages for monthly temperature and precipitation were obtained from WorldClim (Hijmans et al., 2005), whereas daily rainfall was extracted from the 3b42 product of the Tropical Rainfall Measuring Mission (TRMM) database (Bell, 1987; Huffman et al., 1995; Kummerow et al., 1998) using the Version 3 dataset available from CSI-CGIAR.

4.2.2 VARIABLE SELECTION

Variables were selected taking into account as much information as possible on the available environmental conditions of each production site. As mentioned previously there is no reliable information about the climate associated with each production area. The scenario for soil data is even worse, as there does not exist a reliable soil map to obtain information for sites (see chapter 2).

Therefore, the process of defining variables was not only guided by expert knowledge (agronomic information provided by extension officers on those variables that were considered likely to influence production) but used the information recorded by farmers, available publicly-environmental data and RASTA. Nonetheless, not all variables collected from these sources are useful to our models. Some of the source constraints include: the 1-km spatial resolution of the environmental databases that makes it difficult to get information of smaller areas; and the difficulty

to measure basic properties in RASTA such as pH, presence of carbonates, and hardpans (see Table 2.1).In addition other soil characteristics were not included into our modelling because we failed to see variability in them. For example, all soil structures in our study were granular (see Appendix A1). Hence, variables were chosen on a pragmatic basis, not only using the agronomic knowledge available but considering that they could also be readily recorded, and processed by the model.

As far as productivity is concerned, for under-researched crops, it is typical to have more data for low yields than for high yields, which is a constraint from the data collected by small-scale farmers without access to information to increase productivity. As a consequence, in the resulting dataset, this dependent variable displayed more information on low than on high values of productivity of Andean blackberry (Figure 4.2).

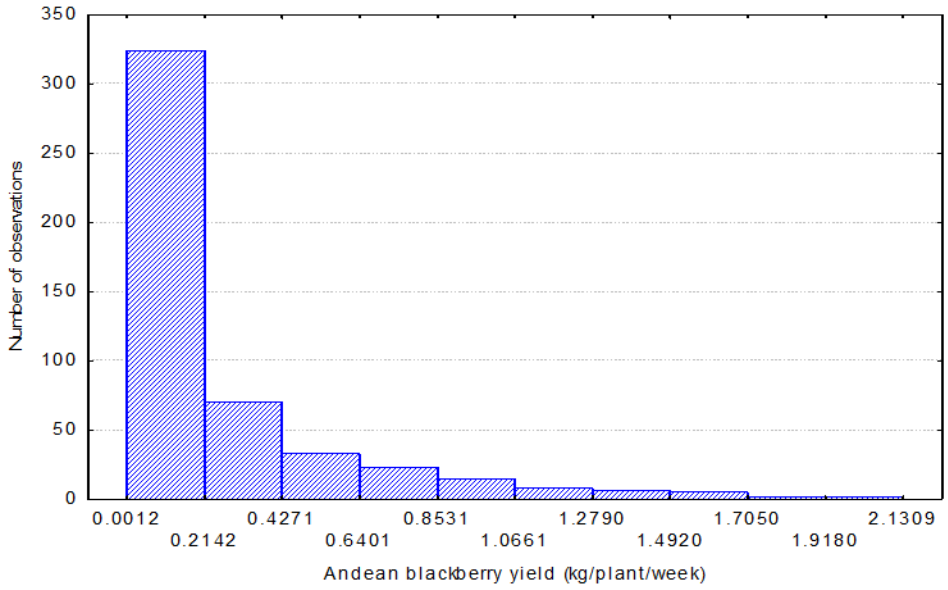


Figure 4.2. Histogram displaying yield data distribution of Andean blackberry

The information compiled in the database for Andean blackberry consisted of 28 variables (Table 4.1). This information included binarized categorical variables (see chapter 3) describing geographical position (large areas for departments, specific areas for particular localities within departments) and variety (thorny blackberry or thornless blackberry), and environmental variables based on landscape, soil and climate (Table 4.1). The climate data was chosen to represent the critical period for yield formation which is from first appearance of flowers to fruit ripening (see chapter 2). Experienced agronomists and extension officers had warned us of frequent outbreaks of *Botritis* (*Botrytis cinerea*) in the stage of flower initiation, which affect yield and are related to weather conditions at that time. Each yield observation was associated with the environmental variables taking into account the date of harvest (Figure 4.3).

Scatter plots representing continuous variables and yield were produced in order to facilitate visualization (Figures 4.4a to 4.4o).

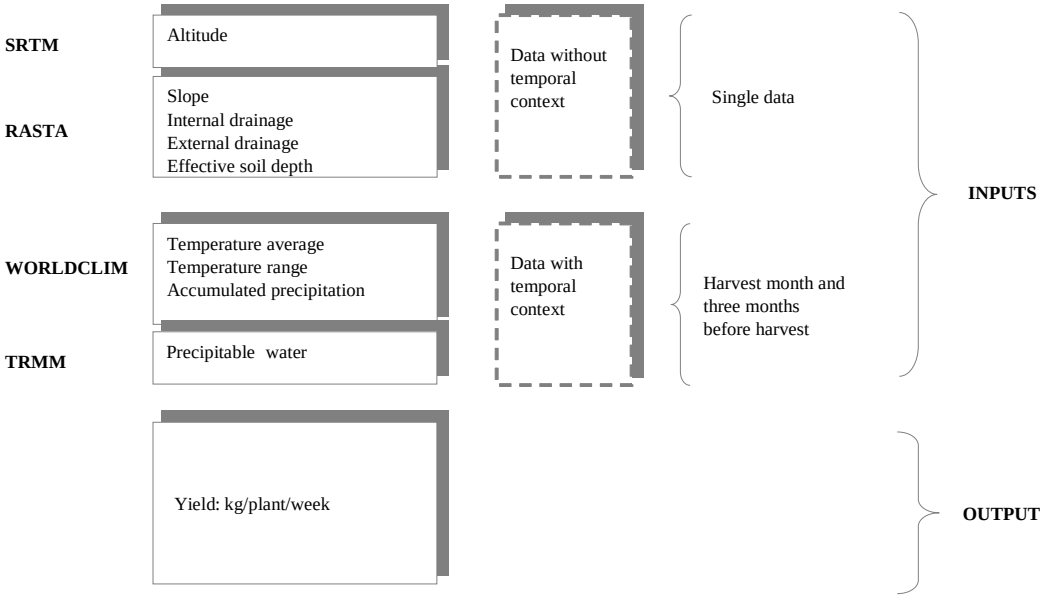


Figure 4.3. Variables selected for the construction of the Andean Blackberry yield model

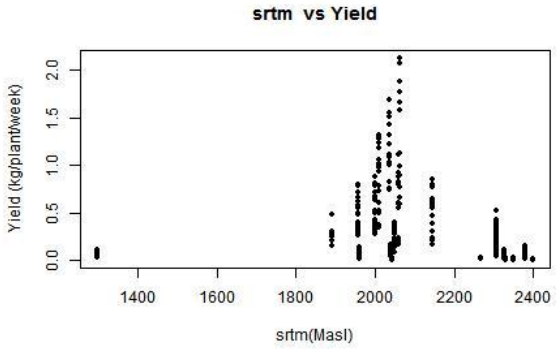
Table 4.1. Inputs used for development of Andean blackberry yield model

Input	*Variable*	*Units*	*Abbreviation*	Source	Ranges
1	[a]Thorn or Thornless blackberry	-	AB_Thorn_N	AEPS	-
2	[a]Nariño – Caldas (Large geographic area)	-	Nar-Cal	AEPS	-
3	[a]Nariño, la union, chical alto (specific geographic area)	-	Na_un_chical	AEPS	-
4	[a]Nariño, la union, cusillo alto (specific geographic area)	-	Na_un_cusal	AEPS	-
5	[a]Nariño, la union, cusillo bajo (specific geographic area)	-	Na_un_cusba	AEPS	-
6	[a]Nariño, la union, la jacoba (specific geographic area)	-	Na_un_lajac	AEPS	-
7	[b]Caldas Riosucio zona rural (specific geographic area)	-	Cal_riosu_zr	AEPS	-
8	[b]Altitude	MASL	Srtm	SRTM	1297-2399
9	[b]Slope	degrees	Slope	SRTM	2-50
10	[b]Internal drainage	-	IntDrain	AEPS	1-3
11	[b]External drainage	-	ExtDrain	AEPS	1-3
12	[b]Effective soil depth	cm	EffDepth	AEPS	25-80
13	[b]Precipitable water of the harvest month		Trmm_0	TRMM	0.4 -8[c]
14	[b]Precipitable water of the first month before harvest	mm	Trmm_1	TRMM	0.5-8[c]
15	[b]Precipitable water of the second month before harvest	mm	Trmm_2	TRMM	0.5-12[c]
16	[b]Precipitable water of the third month before harvest	mm	Trmm_3	TRMM	0.5-12[c]
17	[b]Average temperature of the harvest month	°C	TempAvg_0	WORLDCLIM	13-26
18	[b]Temperature range of the harvest month	°C	TempRang_0	WORLDCLIM	6-12
19	[b]Accumulated precipitation of the harvest month	mm	PrecAcc_0	WORLDCLIM	43-360
20	[b]Average temperature of the first month before harvest	°C	TempAvg_1	WORLDCLIM	13-25
21	[b]Temperature range of the first month before harvest	°C	TempRang_1	WORLDCLIM	6-11
22	[b]Accumulated precipitation of the first month before harvest	mm	PrecAcc_1	WORLDCLIM	120-360
23	[b]Average temperature of the second month before harvest	°C	TempAvg_2	WORLDCLIM	13-25
24	[b]Temperature range of the second month before harvest	°C	TempRang_2	WORLDCLIM	7-10
25	[b]Accumulated precipitation of the second month before harvest	mm	PrecAcc_2	WORLDCLIM	117-360
26	[b]Average temperature of the third month before harvest	°C	TempAvg_3	WORLDCLIM	13-24
27	[b]Temperature range of the third month before harvest	°C	TempRang_3	WORLDCLIM	7-12
28	[b]Accumulated precipitation of the third month before harvest	mm	PrecAcc_3	WORLDCLIM	90-358

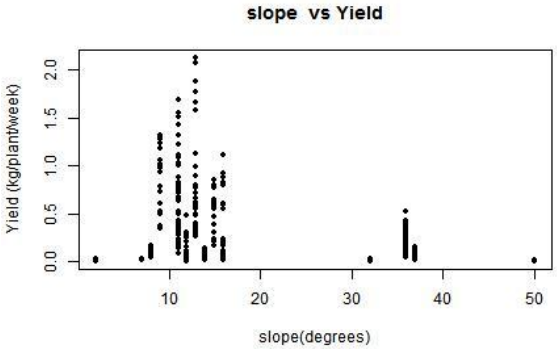
[a]Categorical variables- units are classes of each variable

[b]Continuous variables

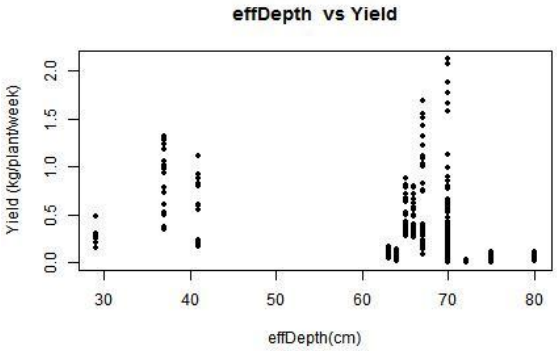
[c] Values divided by 10 according to the data provided by the satellite


srtm vs Yield
Yield (kg/plant/week)
0.0
0.5
1.0
1.5
2.0
1400
1600
1800
2000
2200
2400
srtm(Masl)

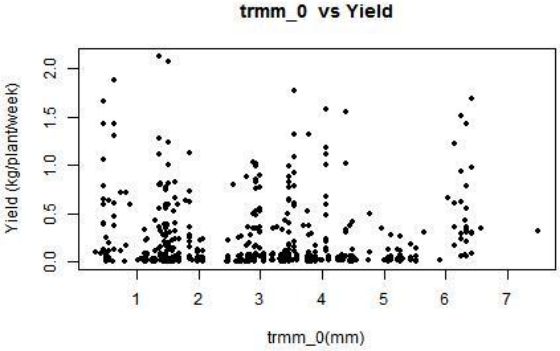
(a)


slope vs Yield
Yield (kg/plant/week)
0.0
0.5
1.0
1.5
2.0
10
20
30
40
50
slope(degrees)

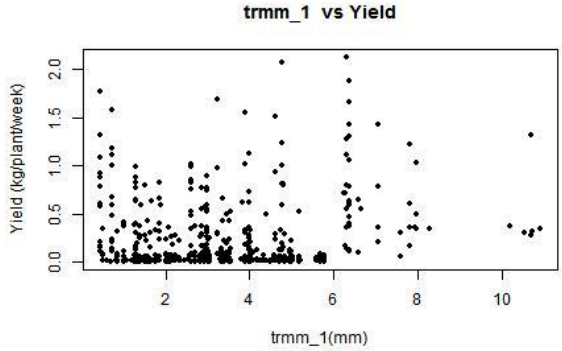
(b)


effDepth vs Yield
Yield (kg/plant/week)
0.0
0.5
1.0
1.5
2.0
30
40
50
60
70
80
effDepth(cm)

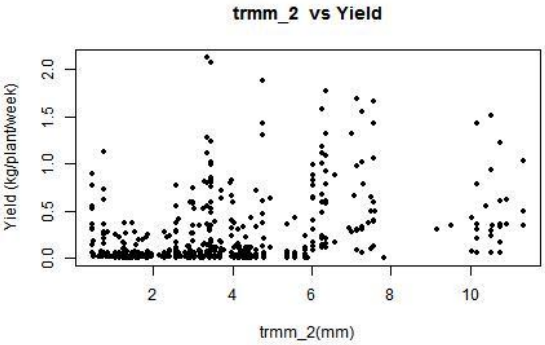
(c)


trmm_0 vs Yield
Yield (kg/plant/week)
0.0
0.5
1.0
1.5
2.0
1
2
3
4
5
6
7
trmm_0(mm)

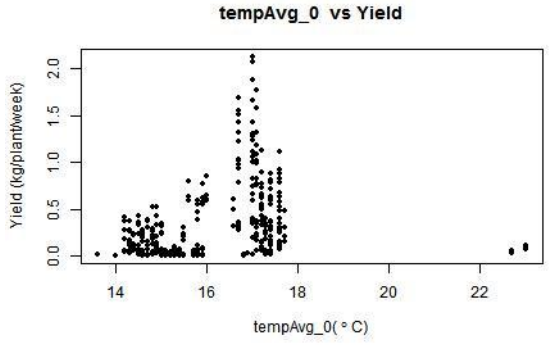
(d)


trmm_1 vs Yield
Yield (kg/plant/week)
0.0
0.5
1.0
1.5
2.0
2
4
6
8
10
trmm_1(mm)

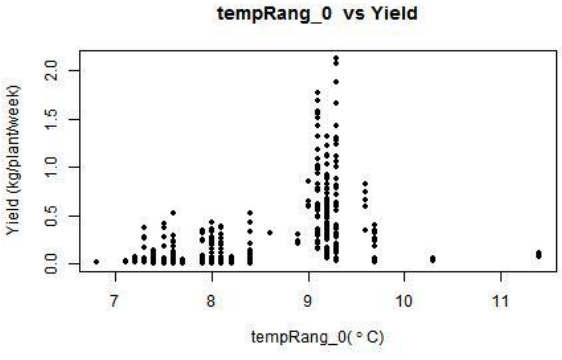
(e)


trmm_2 vs Yield
Yield (kg/plant/week)
0.0
0.5
1.0
1.5
2.0
2
4
6
8
10
trmm_2(mm)

(f)


tempAvg_0 vs Yield
Yield (kg/plant/week)
0.0
0.5
1.0
1.5
2.0
14
16
18
20
22
tempAvg_0(° C)

(g)


tempRang_0 vs Yield
Yield (kg/plant/week)
0.0
0.5
1.0
1.5
2.0
7
8
9
10
11
tempRang_0(° C)

(h)

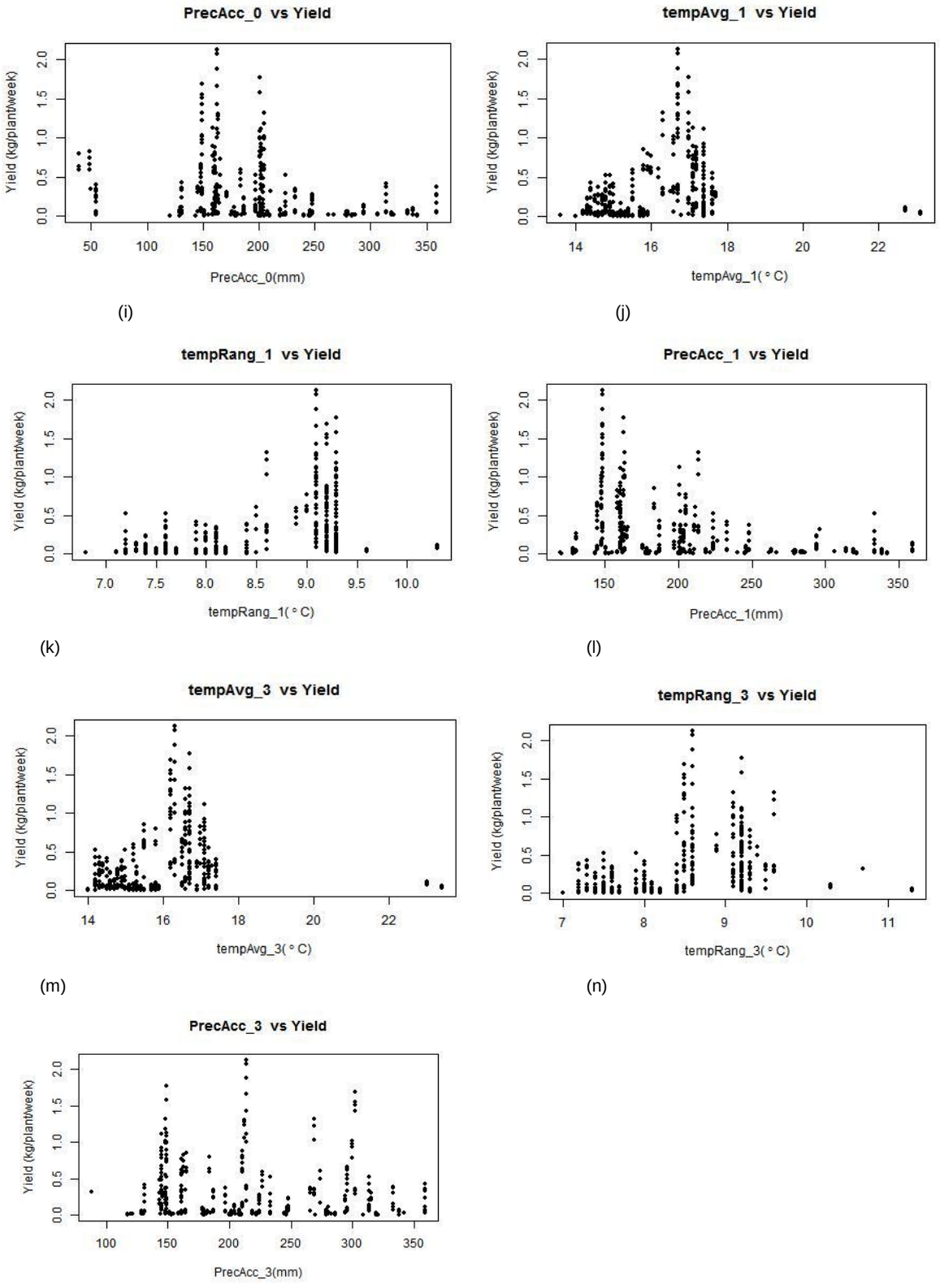


Figure 4.4. Scatter plots of continuous variables vs. yield. (a) srtm, (b) slope, (c) effective soil depth, (d) precipitable water of the harvest month, (e) precipitable water of the first month before harvest, (f) precipitable water of the second month before harvest, (g) average temperature of the harvest month, (h) temperature range of the harvest month, (i) accumulated precipitation of the harvest month, (j) average temperature of the first month before harvest, (k) temperature range of the first month before harvest, (l) accumulated precipitation of the first month before harvest, (m) average temperature of the third month before harvest, (n) temperature range of the third month before harvest, (o) accumulated precipitation of the third month before harvest

4.2.3 COMPUTATIONAL MODELS

4.2.3.1 MLP

ANNs software FENNIX was used in order to train a feed-forward neural network. This software is a graphical interface for a neural network simulator which allows fast experimentation on analysis. It can be downloaded from http://fennix.sourceforge.net/. FENNIX has been implemented and designed at the HEIG-VD the SSAFT project. The software is used by statisticians, students, and agronomists, as well as experts on artificial neural network modelling in many countries (http://sourceforge.net/projects/fennix/files/FENNIX.zip/stats/map?dates=2012-01-25%20to%202012-11-11).

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 yield variable, is determined by the 28 variables we used as inputs. The Back-propagation algorithm was employed in order to train the neural networks (Bishop, 1995). The algorithm is a descent-based optimizer that minimizes the difference between the desired output of the model (in the training dataset) and the actual output of the network, e.g. the Mean Square Error (MSE) (see chapter 3).

The mechanism for testing model performance was the split-sample approach explained in chapter 3. 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 to 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, whereupon the different values of the averaged MSE were compared in order to determine the network having the best performance based on error criteria. The topology with the lowest MSE over the validation subset had 5 units in the hidden layer neural network (Figure 4.5) and was then selected for further development.

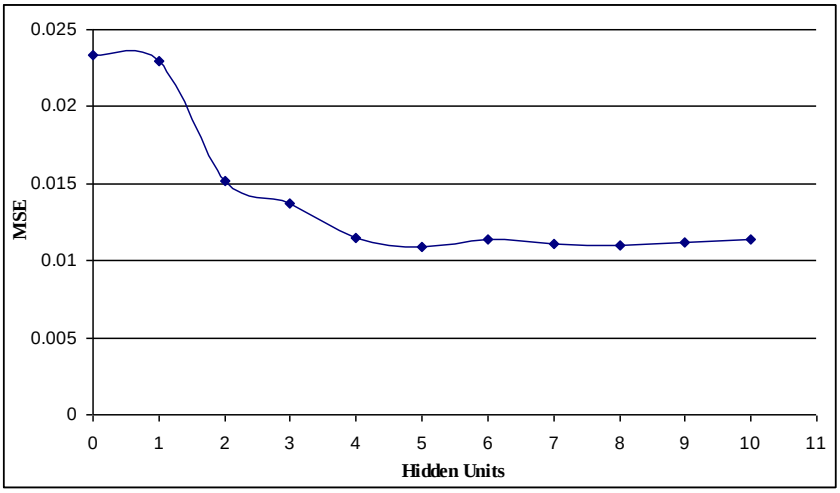


Figure 4.5. Validation error (MSE) of artificial neural networks with different numbers of neurons in the hidden layer

In FENNIX, it is possible to work with so called ensemble networks. Thus, 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, 2005). Neural networks ensembles are less affected by local minima, and have been shown to outperform their single components (Yao, 1998). In our case, the source of diversity among models was the starting point of the Back-propagation algorithm (random initialization). The resulting model output was then calculated by averaging the outputs of the 100 individual networks. 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 chapter 3.

4.2.3.2 SOM

SOM (Kohonen, 1995) is a non-supervised algorithm which combines clustering and visualization. SOM maps high-dimensional datasets in a low-dimensional output space (generally a grid of two dimensions). Observations with similar characteristics appear clustered together in the low-dimensional map produced. Such a map facilitates exploratory, visual analysis of the clusters and 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 separated, and standard clustering methods are applied to the SOM prototype vectors. Then, the clusters are displayed on a lattice (Vesanto and Ahola, 1999; Barreto and Pérez-Uribe, 2007).

In order to implement SOM as one of the computational models in the present research, the software package MATLAB 7.0 and function package SOM toolbox were used. This latter package, developed by the Laboratory of Computer and Information Science at Helsinki University of Technology. The package is documented and downloadable from

http://www.cis.hut.fi/somtoolbox/download/. The software was used to (a) train a Kohonen map with a total of 488 yield records, associated with each event and its 28 variables, creating an input matrix trained with a SOM of 100 neurons (10 x 10); (b) cluster the prototypes of the resultant bidimensional map through the K-means algorithm and Davies-Bouldin index mentioned in chapter 3; and finally (c) visualize dependencies between clusters shown in the Kohonen map by a "component plane" representation, where several lattices, one for each variable, are shown side by side.

The variables are visualized in a lattice called a component plane with a variable-specific colouring. 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 (Figures 4.9 - 4.16). 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). As the SOM toolbox was built using MATLAB script language, scripts for each of these a, b, and c steps are shown in appendix A3.

4.3 RESULTS AND DISCUSSION

4.3.1 MODEL PERFORMANCE

The neural network model was evaluated to ensure that its performance was acceptable for our purpose of determining the relationship between yield of the Andean blackberry and characteristics of the sites where it is 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 only using the data from the "hold-out" validation dataset (Figure 4.6). The coefficient of determination (0.89) indicates that the model explained close to 90% of total variation, which we considered sufficient to proceed to the next step of determining input relevance.

The fit between real yield values and predicted values taken from the validation data was close at low yield levels, but was poor over the range of high levels (between 69 and 93, see Figure 4.7). At the same time, model accurately predicted expected yields at high yield levels. The model can be used to determine *ex ante* conditions and management associated with high yields, which can be used to provide guidelines for farmers on how to obtain high yields. It can also be used to identify site characteristics that are inevitably associated with poor crop performance and can thus be used to indicate to farmers that a particular site and management combination is not a viable option.

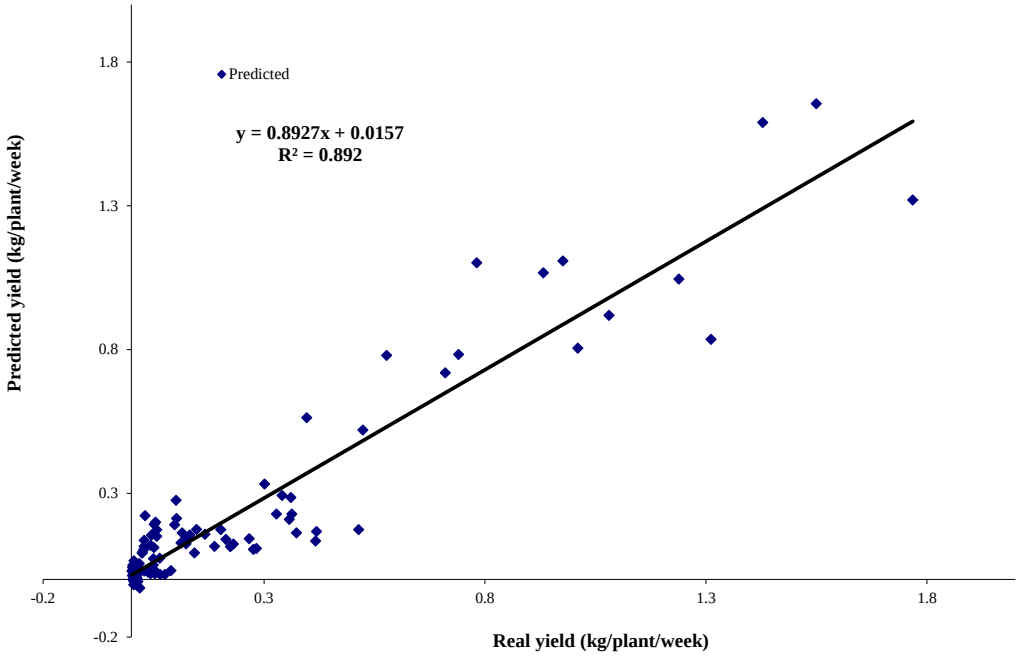


Figure 4.6. Scatter plot displaying multilayer perceptron predicted yield versus real Andean blackberry yield, using only the validation dataset

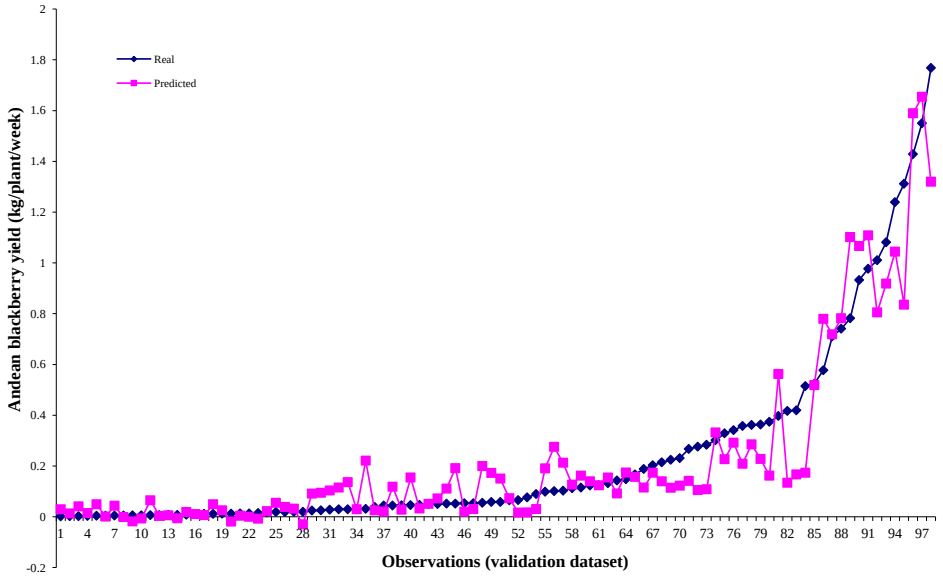


Figure 4.7. Line with markers displaying the fitness of the predicted and real Andean blackberry yield through the observations from the validation dataset (yield values upwardly ordered)

4.3.2 MODEL INTERPRETATION

We assessed yield response to changes in the 28 variables used in the model by calculating the sensitivity of the model output with respect to each of the inputs through the sensitivity metric described in chapter 3. This metric expresses the amount of change of the output with variations of the inputs. The nine most important variables identified by the sensitivity metric were: soil depth;

average temperature of the first month before harvest; specific geographical areas Nariño-La Union-Chical Alto and Nariño-La Union-Cusillo Bajo; average temperature of the harvest month; average temperature of the second month before harvest; average temperature of third month before harvest; external drainage and accumulated precipitation of the first month before harvest (Figure 4.8). As Figure 4.8 also shows, there was a moderately sharp drop in the sensitivity after the ninth variable. A Wilcoxon test at an alpha level of 5% (Table 4.2) indicated that the means of this group of nine variables were significantly different (p=0.0001) from the rest of variables. Hence, these nine most important variables were selected for further analysis.

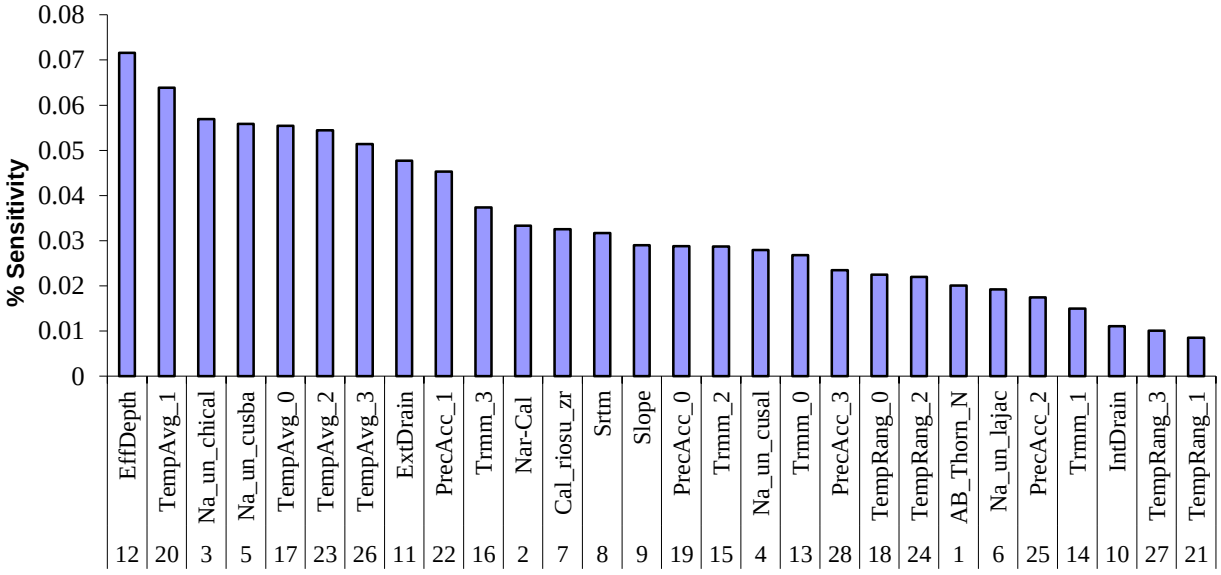


Figure 4.8. Sensitivity distribution of the model with respect to the inputs

Table 4.2. 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**

4.3.3 VISUALIZATION OF THE RELATIONS BETWEEN THE VARIABLES FOUND AS RELEVANT BY THE SENSITIVITY METRIC AND CLUSTERS WITH SIMILAR PRODUCTIVITY OF ANDEAN BLACKBERRY

To further analyse the effects of the nine variables, the Kohonen map was trained with the same observations we employed to train the MLP. The resulting bidimensional map is composed of vector prototypes which associate topological information of the original 28 variables for Andean blackberry yield (Figure 4.9a). 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 (Figure 4.9b).

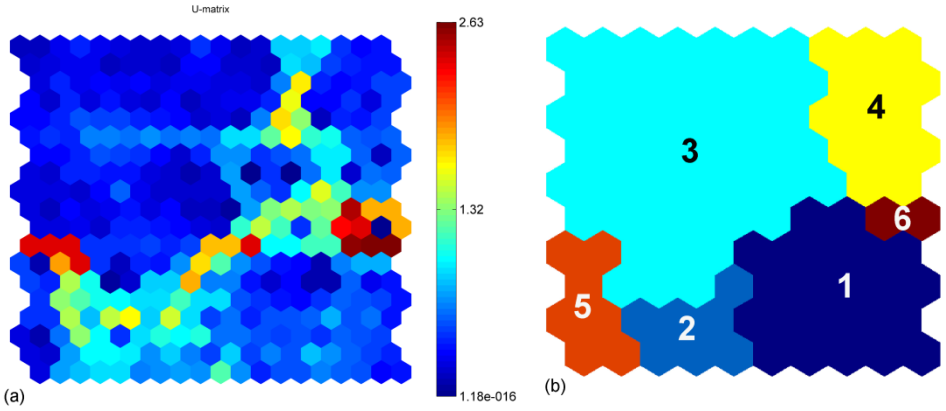


Figure 4.9. Kohonen map showing the resulting clusters: (a) U- matrix displaying distance among prototypes. The scale bar (right) indicates the values of distance. The upper side exhibits high distances, whilst the lower displays low distances. (b) Kohonen map displaying the 6 clusters obtained after using the K-means algorithm and the Davies-Bouldin index

4.3.3.1 COMPONENT PLANES AND VARIABLE DEPENDENCIES

In order to improve the visualization of the dependencies between the clusters shown in the Kohonen map (Figure 4.9b), the "component planes" of Andean blackberry productivity (Figures 4.10a and 4.10b), and the variables previously identified as the most relevant for modelling Andean blackberry yield: effective soil depth (Figure 4.11), average temperature of the harvest month, average temperature of the first, second and third months before harvest, (Figures 4.12a - 4.12d), two specific geographic areas (Figures 4.13 and 4.14), external drainage (Figure 4.15), and accumulated precipitation of the first month before harvest (Figure 4.16), were separated from the Kohonen map and displayed as lattices.

4.3.3.1.1 PRODUCTIVITY PLANE

Yields greater than 1.16 Kg/plant/week were associated with regions in cluster 2 on the Kohonen map (Figures 4.10a and 4.10b).Yield values between 0.018 and 1.16 Kg/plant/week correspond to clusters 1, 3, 4, 5 and 6 in the Kohonen map. Whereas clusters 3, 4, and 6 were the clusters with lowest yields. In summary, the range of yield for low values was between 0.018 g and 0.920 Kg (clusters 3, 4, and 6) for medium 0.920 kg and 1.16 kg (cluster 5, and 6) and for high between 1.16 Kg and 2.08 Kg (cluster 2).

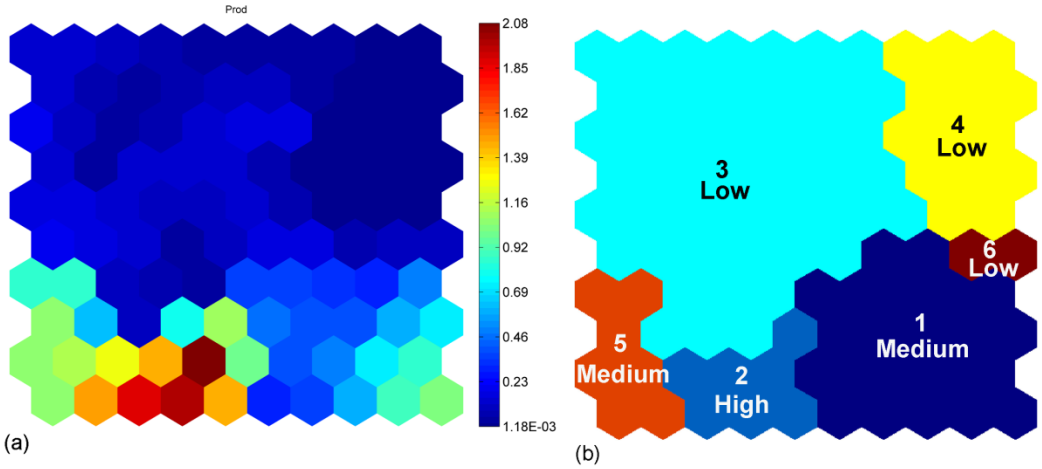


Figure 4.10. (a) Component plane of Andean blackberry yield, the scale bar (right) indicates the range value of productivity in kg/plant/week. The upper side exhibits high yield values, whereas the lower displays low values; (b) Kohonen map displaying the resultant 6 clusters and their labels according to yield values

4.3.3.1.2 EFFECTIVE SOIL DEPTH

Inspection of Figure 4.11 indicates that high yields are obtained when effective soil depth is greater than around 65 cm (cluster 2). Low yields were also found on soils with depths greater than 65 cm (clusters 3, 4 and 6) suggesting that other soil factors not included in the analysis were affecting productivity, presumably soil characteristics such as presence of rock fragments, soil structure or salinity and sodicity. As it was aforementioned, in this study there is absence of soil variables that were difficult to measure by means of RASTA and therefore were not integrated into the model. Without having these data it is not possible to draw firm conclusions on the factors that might affect yield in soil depth deeper than 65 cm.

Most roots of Andean Blackberry are concentrated in the first 30 cm of soil, with some roots found at depths of 50 to 105 cm (Franco and Giraldo 2002). Combining the expert opinion of Franco and Giraldo (2002) with the information from farmers' field points, there is a minimal effective soil depth for high yields of Andean Blackberry of about 65 cm. Below that soil depth yields will be medium (clusters 1 and 5).

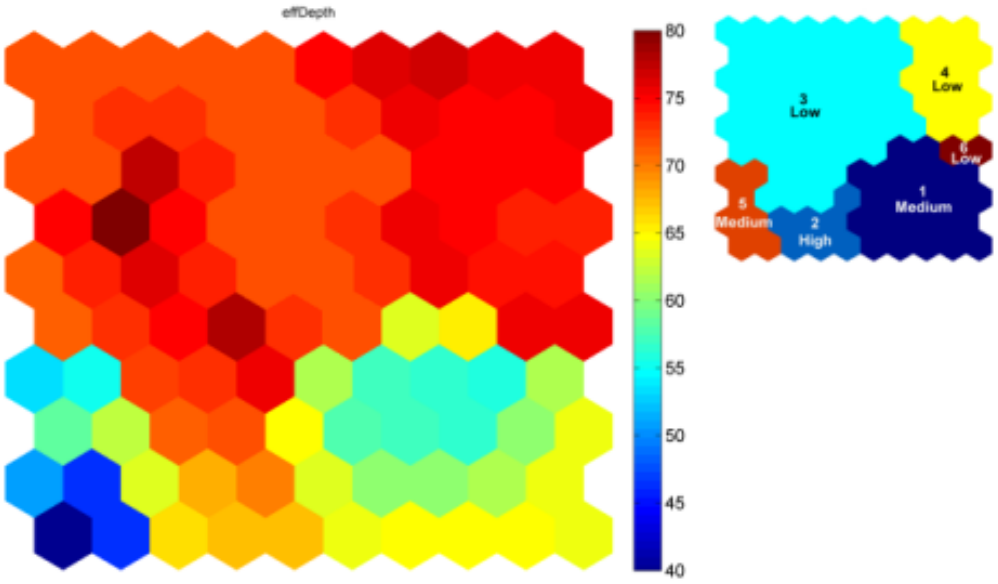


Figure 4.11. Component plane of effective soil depth. The scale bar (right) indicates the range value in cm of soil depth: the upper side of the scale exhibits high values, whereas the lower displays low values

4.3.3.1.3 AVERAGE TEMPERATURE OF THE HARVEST MONTH AND AVERAGE TEMPERATURES OF THE FIRST, SECOND AND THIRD MONTHS BEFORE HARVEST

Kohonen maps for temperature of the first, second and third months before harvest were similar (Figure 4.12). MLP showed that average temperature of the first month before harvest was more important than the other temperatures (that occurs due to small differences captured to better fit the output). However, in the tropical environment the monthly variation in temperature is small, particularly in comparison with the spatial variation in temperature which is largely associated with differences in altitude in the Andean region.

We consider that the differences in input relevance in terms of temperature of the different time periods before harvest captured by the multilayer perceptron are not necessarily causal, but that they were a product of the process of learning in order to better fit the output. Taking into account both the process of determining relevance and the similarity of temperatures for different months in each site, temperatures were analysed as the three months average rather than separately for each month. From this analysis, it is immediately evident that cluster 6 with temperatures of about 24 °C is not well suited to allow high yields of blackberries (Figure 4.12). Clusters 1, 2 and 5 with medium to high yields are related to temperatures between 16 and 18°C (Figures 4.12a - 4.12d), whereas 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 to 18 °C to be

associated with high yields, with lower yields coinciding with temperatures above or below this range.

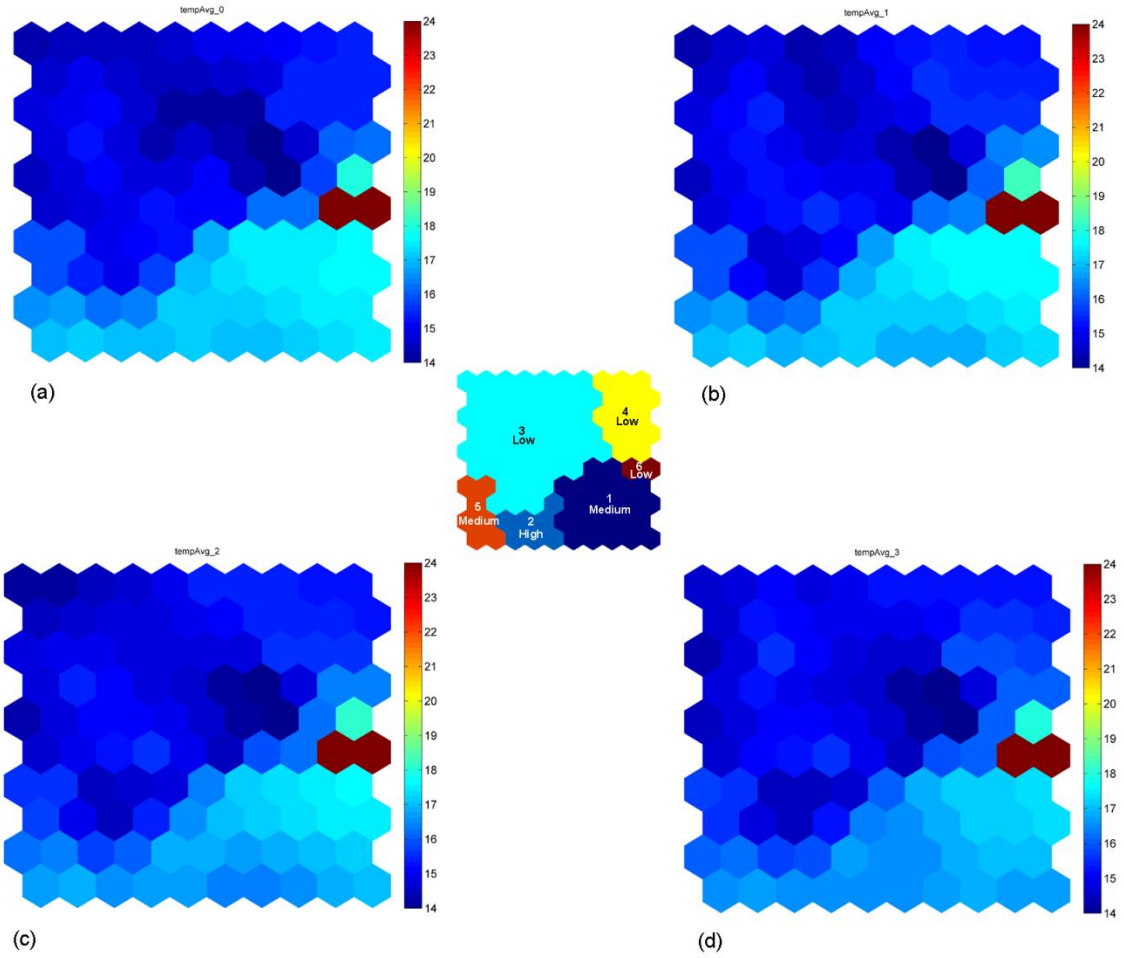


Figure 4.12. Component planes of the average temperature: (a) temperature of the harvest month; (b) average temperature of the first month before harvest; (c) average temperature of the second month before harvest; and (d) average temperature of the third month before harvest. In all figures, the scale bar (right) indicates temperature range value in °C, the upper side exhibits high values, whereas the lower displays low values

4.3.3.1.4 GEOGRAPHICAL 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 directly measured. 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 of importance and related to the geographic location of a site.

In sugarcane, certain farmers consistently obtain higher yields than others even in the same edaphic-climatic conditions; the farmers that obtain the higher yields belong to a particular cultural and socio-economic group that apply better management practices (Isaacs et al., 2007; Cock et

al., 2011). The localities Nariño-La Union-Chical Alto (Figure 4.13), and Nariño-La Union-Cusillo Bajo (Figure 4.14), were associated with cluster 2 which is characterized by the highest yields. As will be shown in Chapter 5, some farmers in a specific locality obtained higher yields than others, even under similar environmental conditions. This implies that the differences are due to management practices associated with the particular locality

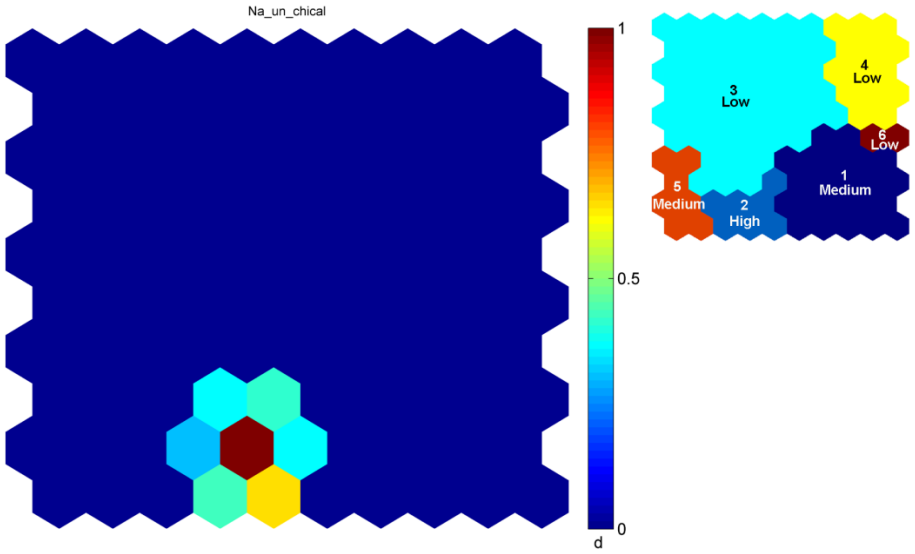


Figure 4.13. Component plane of the specific geographic area Nariño-La Union-Chical Alto. The highest values indicate presence and the lowest absence as they are categorical variables

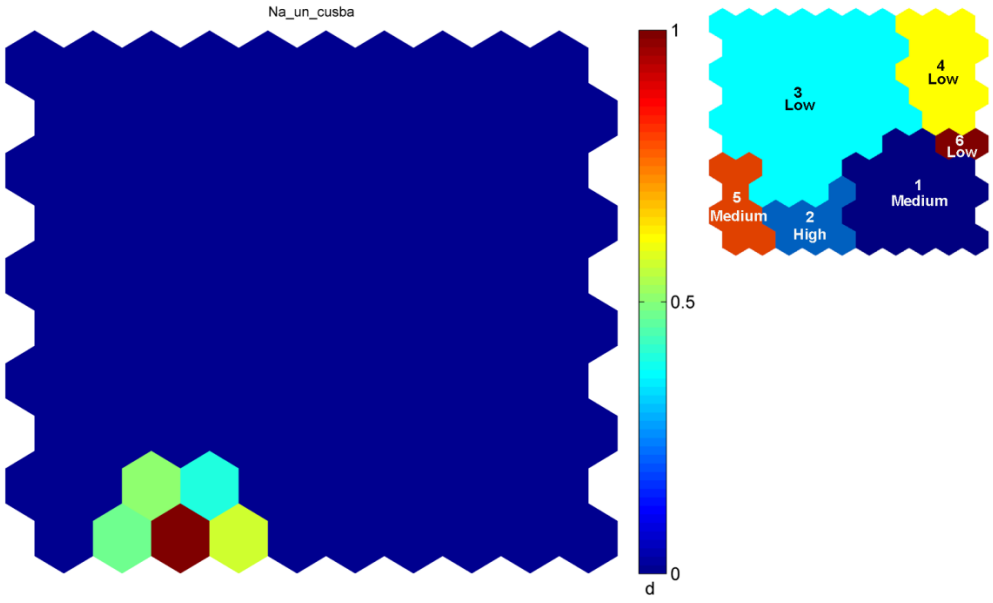


Figure 4.14. Component plane of the specific geographic area Nariño-La-Union-Cusillo Bajo. The highest values indicate presence and the lowest absence as they are categorical variables

4.3.3.1.5 EXTERNAL DRAINAGE AND ACCUMULATED PRECIPITATION OF THE FIRST MONTH BEFORE HARVEST

Scrutiny of the external drainage lattice (Figure 4.15) gave no obvious clues as to how drainage affects yield of blackberries. In appendix A1, it can be seen how external drainage is categorized as, excessive, good moderate, and slow. Continuous values have been assigned to this variable for each category. Thus, according to the dataset, which was used in this study 3 indicates good or fast drainage, 2 moderate drainage, and 1 poor or slow drainage.

In fact, medium yield in cluster 5 is associated with poor external drainage; whereas cluster 2 which has high yields 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 (Figure 4.16): This does not only appear to be true from the Kohonen maps, but it also makes agronomical sense. Good external drainage is evidently more important when rainfall is higher.

This example clearly indicates how visual inspection of Kohonen maps can assist in understanding how various factors affect growth and development of a crop and the interactions between them.

Further inspection of Figures 4.15 and 4.16 indicates 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 a number of exceptions. However, when the two variables, e.g., precipitation of the first month before harvest and external drainage are taken together, it is clear that low rainfall accompanied with varying external drainage conditions can provide good yields, but that heavier precipitation during the first month before harvest with poor drainage is not conducive to high levels of productivity.

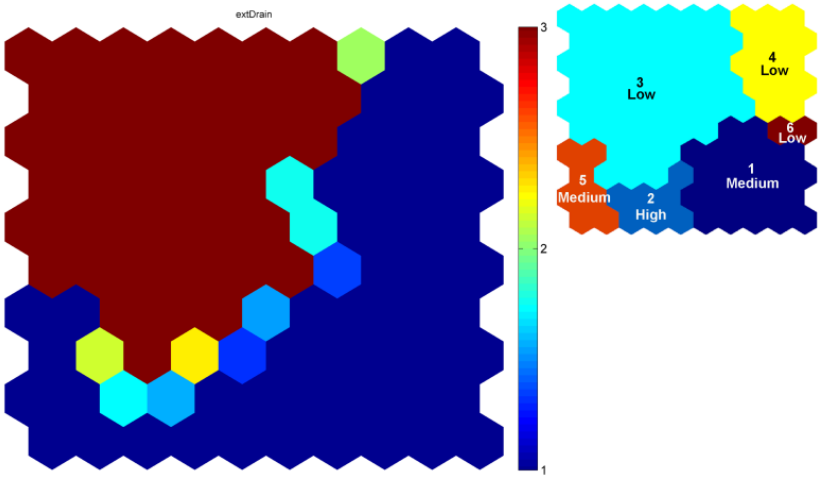


Figure 4.15. Component plane of external drainage. In the scale bar (right), the highest value 3 indicates good or fast drainage, 2 moderate drainage, and 1 poor or slow drainage

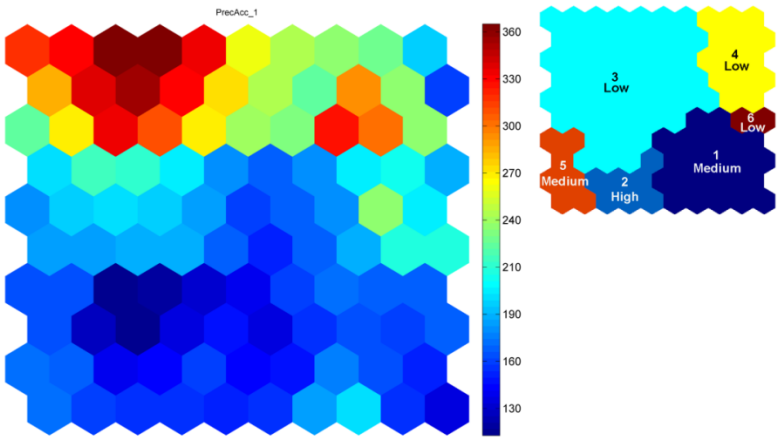


Figure 4.16. Component plane of the accumulated precipitation of the first month before harvest. The scale bar (right) indicates the range value in mm of rainfall; the upper side of the scale exhibits high values, whereas the lower displays low values

4.4 CONCLUSIONS

Farmers' production experiences collected in the Andes coupled with information from publicly-available environmental databases were successfully used to characterise specific production events for Andean blackberry and to relate production to site and time specific events. Analysis focused first on identifying those variables that explain most yield variability by means of MLP neural networks, then using the SOM as a tool for dimensionality reduction and visualization of input-input and input-output dependencies.

ANNs 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 existing publicly-available environmental databases. MLPs were used to develop a model based on a dataset with 28 variables. This model explained close to 90% of variation in a validation set. The relevance metric based on the use of network parameters and input patterns described in chapter 3, was used to identify the most important variables in determining variation in yield. SOM was 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 combination of factors for high yields of Andean blackberry are an average temperature between 16°C and 18°C, minimal effective soil depth of about 65 cm, and low rainfall (between 120 and 210 mm) during the first month before harvest in poor external drainage locations, or moderate to high rainfall (between 210 mm and 360) in better-drained areas. Nonetheless, given the nature of the modelling tool being used (black-box model), these results have to be taken as exploratory, and should be interpreted by experts in the specific field of application of the data.

The identification of geographic areas with higher yields than those that would be expected solely from environmental conditions suggests that 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 these 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.

5 COMBINATION OF DATA-DRIVEN APPROACHES TO INTERPRET VARIATION IN COMMERCIAL PRODUCTION OF LULO (*Solanum quitoense* Lam)

Adapted from**: Jiménez, D.,** Cock, J., Jarvis, A., Garcia, J., Satizábal, H.F., Van Damme, Pérez-Uribe, A., and Barreto, M., 2010. Interpretation of Commercial Production Information: A case study of lulo, an under-researched Andean fruit**.** Agricultural Systems. 104 (3): 258-270

ABSTRACT

Years of agronomic experimentation have led to a wealth of knowledge on crop responses to variation in growth environments. This knowledge has been used to develop empirically-based crop models which quantify crop response to variations in growing conditions. The detailed 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 currently not available. Moreover, it would take years of experimentation using traditional approaches 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 following the top-down model could consist of observing crops under varying management and different environments in the field.

Analysis and interpretation of farmers' production experiences 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 lulo (*Solanum quitoense* Lam), and the associated environmental conditions in Colombia. This information was used to develop and evaluate procedures for the interpretation of variation of farmers' production experiences. The most effective procedures for interpreting yield variation depended on expert guidance: it was not possible to develop a simple effective one-step procedure, but rather an iterative approach was required to analyse and interpret commercial production data of lulo.

The most effective procedure was based on the following steps. First, highly correlated independent variables were evaluated and those that were shown to be duplicates were eliminated. Second, regression models identified the environmental factors most closely associated with the dependent variable of fruit yield. Environmental factors associated with variation in fruit yield were then used for more in depth analysis, whereas environmental variables not associated with yield were excluded from further analysis.

Linear regression and multilayer perceptron regression models explained 65-70% of total variation in yield. Both models identified three of the same factors as being important, whereas the multilayer perceptron based on a neural network approach 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 analysed 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 total variation in yield with 61% associated with HECs and 19% with farm. Location only 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 or farmers' production experiences are available, and also to identify those farms that have superior management practices for a given set of environmental conditions.

5.1 INTRODUCTION

In the case of tropical fruit crops, farmers' production experiences could be used to interpret crop responses to variation in growing conditions caused both by inherent variation in growth environment and also by variation in farm management practices. As mentioned in the previous chapters, our premise is that if it were possible to describe the management and environmental conditions that characterise a production system, and if information of farmers' production experiences occurring under varying conditions were available, it would be possible to develop data-driven models for the production system (Jiménez et al., 2009). The experience acquired with Andean blackberry in the preceding chapter, in addition to studies conducted for sugarcane and coffee, demonstrated that recording farmers' production experiences occurring within the naturally variation in management and environment under which these tropical species are growing, crops response can be modelled using best fit models (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 crop response, interactions and non-linearity of the responses, and the inevitable errors in data collection by farmers, a large number of datasets are required to be able to make sense of the data. With the large datasets required to draw conclusions, it is likely that novel analytical approaches will be necessary. Grimm (1999) suggested that modellers should experiment more

with their mathematical models. In this chapter, we experiment with various analytical modelling approaches and compare their efficacy in explaining yield for lulo.

In the present chapter a range of approaches is used to interpret variation in information of farmers' production experiences 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. The efficacy of the different approaches is compared with examples from non-parametric and parametric methods and combinations of the two approaches.

ANNs were selected as the non-parametric approach, and multiple linear regression and mixed models as the parametric approaches. Furthermore, based on the recommendation of Schultz et al. (2000) the parametric approach was combined with ANNs methods in order to benefit from the advantages of both. Multiple linear regression and mixed models combined with Best Linear Unbiased Prediction (BLUP) are frequently used to understand 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, the results of 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 farmers' production experiences, categorical variables such as agro-ecological zone and location are likely to be important. ANNs, as non-parametric approaches, and as was previously mentioned, 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 modelling (Hashimoto, 1997; Schultz and Wieland, 1997; Paul and Munkvold, 2005; Miao et al., 2006). However, and as it has also been mentioned in conclusions of chapter 3, among the disadvantages of artificial neural networks are: their "black-box" nature, they are computationally exhaustive, they can be over-trained and give false expectations of their predictive capacity (Schultz et al., 2000; Sargent, 2001; Paul and Munkvold, 2005; Ozesmi et al., 2006).

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

method could be 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., 2011) and shrimp production (Gitterle et al., 2009), suggests that one of the most effective means of analysing farmers' production experiences information is first to establish clusters of events with similar environmental conditions, and then to determine the effects of variation in 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 for identifying clusters of events with similar environmental conditions (Tuma, 2007). Many variables recorded for commercial crops production are likely to be categorical (e.g. weed control, land preparation practices).

Furthermore, categorical variables such as the presence of a given 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 incorporate 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., 2011).

We chose the example of lulo to evaluate different data-driven approaches to develop predictive models. We selected lulo as it is an under-searched 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 it normally produces during the whole year, with high variability in yield in both space and time.

5.2 METHODOLOGY

Farmers' production experiences data were collected on farms, and environmental conditions of farmers' plots were characterized using both, data collected on-farm and publicly-available climate databases as explained in chapter 2. Data was compiled in SSAFT databases for analysis. The resulting database, which is the result of merging information from different sources, was analysed in an iterative way with the aim of finding both the most apposite dataset and approach to model it.

5.2.1 FARMERS' PRODUCTION EXPERIENCES

The calendars developed by *Corporación BIOTEC* in collaboration with lulo producers in the department of Nariño, Colombia, were used to keep on-farm records based on a calendar that also provided them with useful information on lulo production (BIOTEC, 2007). Twenty-one lulo producers (Figure 5.1) recorded information on these calendars over 2 years period (January 2006 to December 2007). Records of individual farms provided a data series on lulo production for each farm with farmers' estimates of the quantity (gr) of fruit harvested per plant per week (see chapter 2 and Appendix A2). Data collected in the database included information on location, variety, yield, and harvest time for a total of 254 records In addition, each site was geo-referenced using a hand-held GPS (Garmin®-Etrex).

5.2.2 BIOPHYSICAL CHARACTERIZATION OF SITES

Similar to the process followed for Andean blackberry in the previous chapter, the generation of climatic, landscape, and topographic information for each site was extracted from interpolated publicly-available databases as explained in chapter 2, through the use of automated algorithms that enable us to estimate environmental conditions of lulo production sites, as they were geo-referenced. Thus, environmental information was obtained from WorldClim, TRMM, and SRTM databases. In the case of lulo, extension officers reported the presence during flowering and fruit setting of *Picudo de la flor* (*Anthonomus* spp.) in the department of Nariño. Expert opinion states that this insect has an important effect on lulo yield. Climate information was included for the period of yield formation (harvest month, first, and second month before harvest) (see chapter 2) which also encompassed the period when the *Picudo de la flor* attacks the flower buds.

With regard to soil characteristics, the simple, easy to learn methodology called RASTA, presented in chapter 2, was distributed to lulo growers to characterise their soil conditions. According to the results obtained for Andean blackberry, we were aware that a number of management variables, that we were not able to measure, could have a major impact on the outputs of the models. In order to estimate the effect of these variables, we integrated location into the analysis in a similar

manner as with Andean blackberry 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.

5.2.3 VARIABLES

In the present study, four locations with 21 different lulo producing sites (Figure 5.1) were characterized. In this case variables were chosen on the same pragmatic basis followed for Andean blackberry, taking into account factors easily recorded by farmers as well as expert opinion on which factors are likely to affect yield. Similar to Andean blackberry, the dataset of lulo has more data for low yields than for high yields (Figure 5.2).

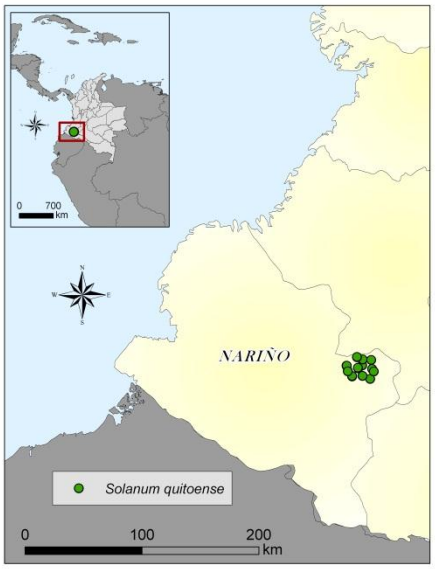


Figure 5.1. Map of the study area; the dots indicate the sites of lulo production where data was collected

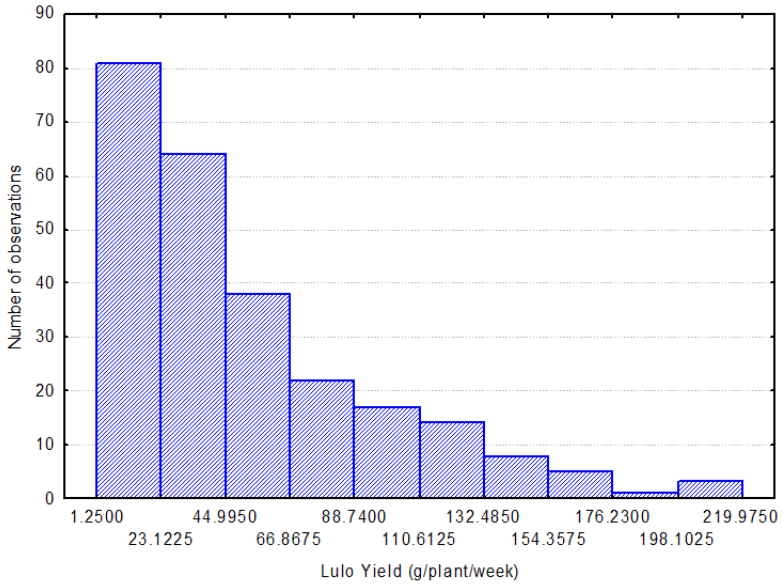


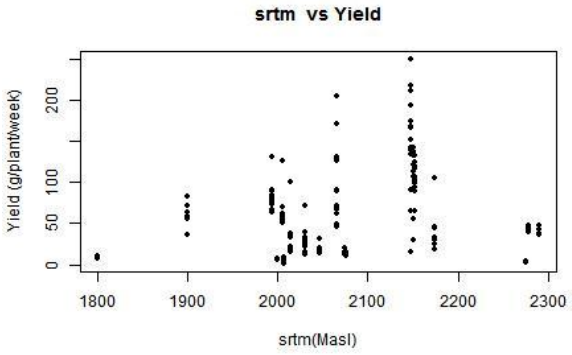
Figure 5.2. Histogram displaying lulo yield data distribution

The information was compiled in the database for lulo with 254 records, 19 independent variables providing information for each site and the dependent variable, e.g. productivity of lulo (Table 5.1). The independent and dependent variables of the multiple linear regression approaches correspond to the supervised ANNs 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 5.1). Scatter plots considering continuous variables and yield were calculated in order to facilitate visualization (Figures 5.3a to 5.3j). Each yield observation was associated with climate variables taking into account date of harvest.

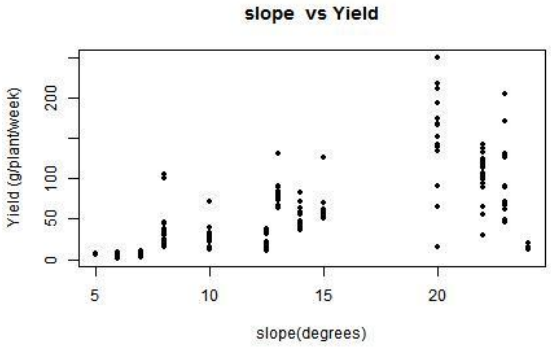
Table 5.1. Variables recorded in the lulo database

Input	Variable	Units	Abbreviation	Ranges
1	**Location**			-
1[a]	[a]Nariño, cartago, san isidro	-	Na_ca_san *	-
1b	[a]Nariño, la union, buenos aires	-	Na_un_ba *	-
1c	[a]Nariño, la union, la jacoba	-	Na_un_jac *	-
1d	[a]Nariño, la union, chical alto	-	Na_un_chical *	
	Variety, landscape, topography, edaphic conditions			
2	[a]Thorn or no thorn	-	Nar_Thorn_N	-
3	[b]Altitude	MASL	Srtm *	1800-2290
4	[b]Slope	degrees	Slope *	5-24
5	[a]Internal drainage	-	IntDrain	2-3
6	[a]External drainage	-	ExtDrain *	2-3
7	[a]Effective soil depth	cm	EffDepth *	21-70
	Climate			
8	[a]Precipitable water of the harvest month	mm	Trmm_0 *	0.2-12 [c]
9	[a]Precipitable water of the first month before harvest	mm	Trmm_1 *	0.2-14 [c]
10	[a]Precipitable water of the second month before harvest	mm	Trmm_2 *	0.3-13 [c]
11	[a]Average temperature of the harvest month	°C	TempAvg_0 *	14-19
12	[a]Average temperature of the first month before harvest	°C	TempAvg_1	12-20
13	[a]Average temperature of the second month before harvest	°C	TempAvg_2	13-18
14	[a]Accumulated precipitation of the harvest month	mm	PrecAcc_0	30-320
15	[a]Accumulated precipitation of the first month before harvest	mm	PrecAcc_1	3-310
16	[a]Accumulated precipitation of the second month before harvest	°C	PrecAcc_2	3-300
17	[a]Temperature range of the harvest month	°C	TempRang_0 *	8-11
18	[a]Temperature range of the first month before harvest	°C	TempRang_1 *	7-13
19	[a]Temperature range of the second month before harvest	°C	TempRang_2 *	8-13
Output	[a]lulo yield		g/plant/week	1.25-220

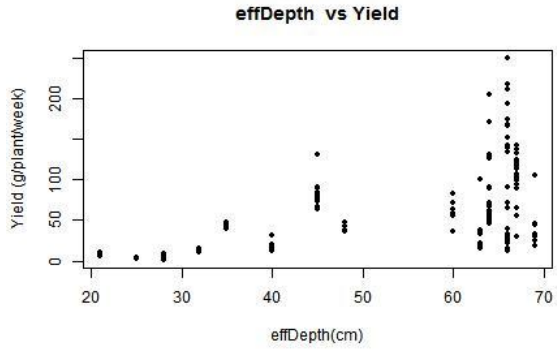
[a] Categorical variables. Units are classes of each variable
[b] Continuous variables
[c] Values divided by 10 according to the data provided by the satellite
* Final set of inputs/dependent variables used in the development of lulo yield regression models

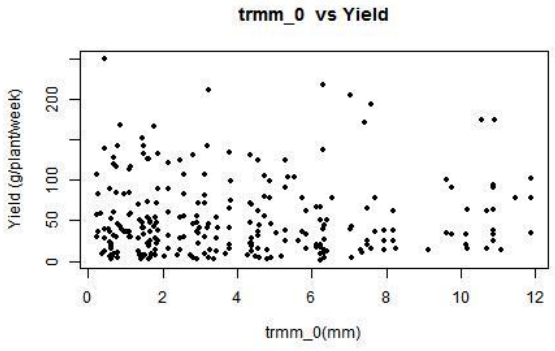


(a)



(b)





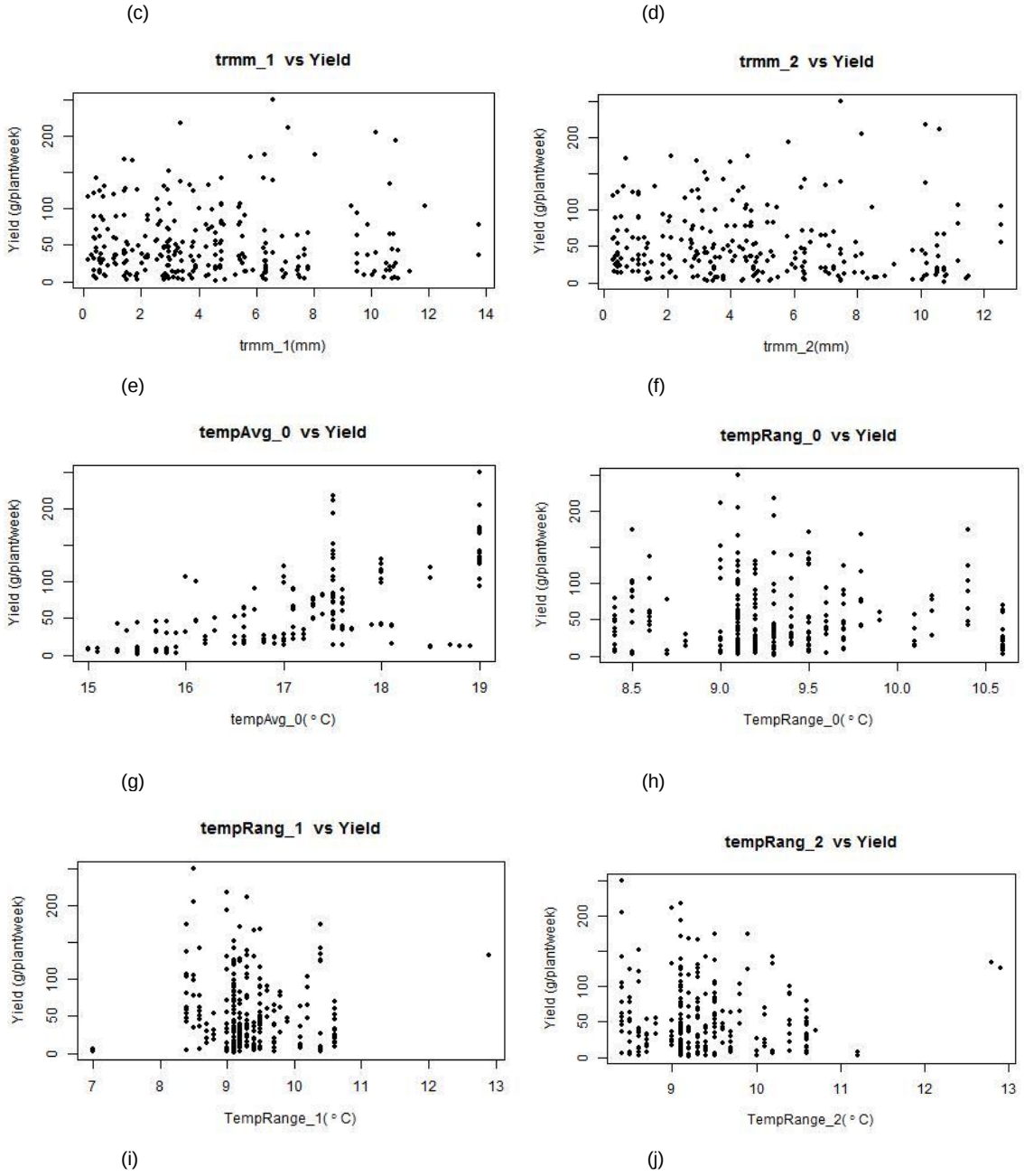


Figure 5.3. Scatter plots of the final set of variables used to model lulo yield. Scatters plots of yield vs: (a) srtm; (b) slope; (c) effective soil depth; (d) precipitable water of the harvest month; (e) precipitable water of the first month before harvest; (f) precipitable water of the second month before harvest; (g) average temperature of the harvest month; (h) temperature range of the harvest month; (i) temperature range of the first month before harvest; and (j) temperature range of the second month before harvest

5.2.4 MODELS

Figure 5.2 shows that there is a preponderance of low productivity data in the dataset with only a few cases of high yields. Statistical analyses often take the drivers of these high productivities as outliers, and yet it is precisely factors associated with high yield that are of most interest.

In the study conducted for Andean blackberry, two geographical areas were identified as highly productive and it was suggested that farmers in these localities were managing their crops particularly effectively. Nonetheless, there was not sufficient information to determine which management factors led to these high yields.

In the case of lulo, we wanted to know further details despite the scarcity of management information. Thus, taking into account (a) that outliers need to be included when modelling under-researched crops; (b) there is a need to know more in detail insights about farmers` knowledge on-farm, and (c) the experience with sugarcane, and coffee determining the effects of variation of management practices, within and between events with similar environmental conditions; we opted for three data-driven approaches. Two regressions were implemented: non-linear and linear regressions and an iterative approach which used linear robust regression, mixed models and a non-supervised ANN combined with expert guidance. In this work, ANNs 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 SOM.

Parametric techniques were employed in order to construct a robust linear regression and to determine the effects of location, management and groups of homogeneous environmental conditions with mixed models in an iterative approach guided by expert opinion.

5.2.4.1 ROBUST LINEAR REGRESSION

STATA statistical package was used to implement multiple regressions. We selected the robust linear regression approach, which exploits as much information as possible without removing outliers, exceptional records or events. This technique is the most adequate 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 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 distance values, whereupon any observation with a Cook's D value greater than 1 was dropped in an iterative process. Using the matrix notation, the Cook’s distance is defined in equation 5.1:

$$D_i = \{\hat{\beta} - \hat{\beta}\ (i)\}X^t X^t\ \{\hat{\beta} - \hat{\beta}\ (i)\}/ps^2 \qquad (5.1).$$

Where: $\hat{\beta}$= is the usual least square estimator vector of *p* by 1 dimension; $\{\hat{\beta} - \hat{\beta}\ (i)\}$ = is the difference between the two *p* by 1 vectors, also *p* by 1; $\hat{\beta}\ (i)$= is the least squares estimator after the *i* th data point has been omitted from the data, also *p* by 1 dimensions; *X*= the matrix containing the values of the independent variables, X^t= the transpose of *X*, *p*= the number of independent variables plus one; and s^2= is the estimate of variance provided by residual mean square error from using the full dataset. A large D_i corresponds to an influential observation; that is, an observation that has more than the average influence on the prediction of the parameters (StataCorp., 2005; Castelló-Climent, 2008).

5.2.4.2 MULTILAYER PERCEPTRON (MLP) REGRESSION

For 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 datasets (Nagendra and Khare, 2006).

Similar to the study conducted for Andean blackberry, a MLP was implemented in FENNIX software to make a non-linear regression. The Back-propagation algorithm was used to train the neural network and minimize the difference between the estimated output of the model and the real output through MSE. Model performance was tested in a similar manner to Andean blackberry whereas the number of neurons in the hidden layer was made comparing neural networks with 1, to 10 hidden units using the bootstrap validation scheme (Efron, 1983) and testing each network by performing split-sample validations 100 times, comparing the different values of averaged MSEs, and then determining the network having the best performance. In the case of lulo, 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 (Figure 5.4).

One hundred networks with the selected topology were built and tested in order to improve the model generalization capabilities (Dietterich, 2000; Brown, 2005).

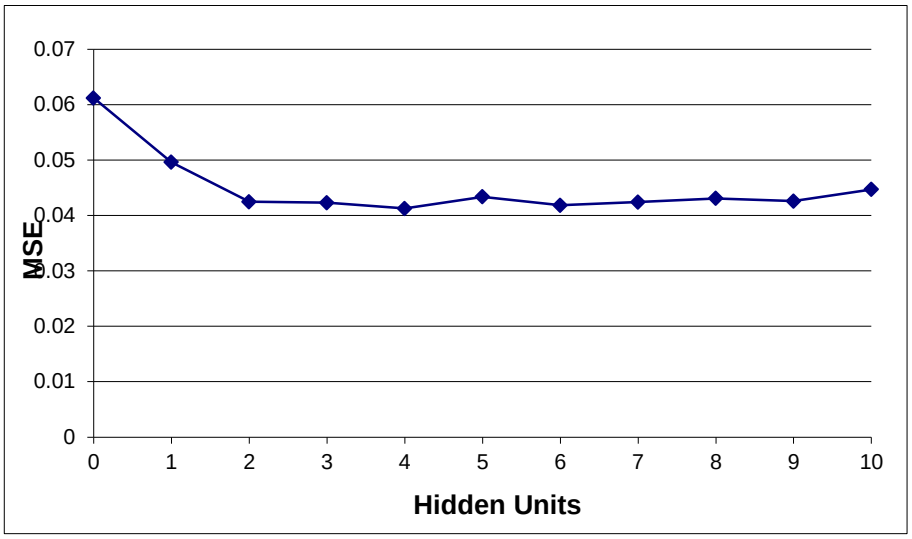


Figure 5.4. MSE of artificial neural networks with different number of neurons in hidden layers

5.2.4.3 ITERATIVE MODEL APPROACH

The iterative approach was based on robust linear regressions, non-linear ANNs regression, and a combination of a non-supervised ANNs known as SOM with mixed models with best linear unbiased prediction. The iterative approach first identified the most important variables associated with yield; second, it used this information to identify homogeneous environmental conditions, and third, analysed differences in productivity related to variation between HECs and due to management variation within HECs.

5.2.4.4 SELF-ORGANIZING MAPS (SOM)

SOMs were used to map high-dimension datasets in a lattice of two dimensions. Observations with similar characteristics, in the high-dimensional space appear grouped together in the two dimensional map. In the same manner as with Andean blackberry, software package MATLAB 7.0 and its function package SOM toolbox were used to train the Kohonen map and to group observations into a given number of K through the K-means algorithm, and Davies-Bouldin index. MATLAB's scripts for training a Kohonen map and clustering prototypes are shown in appendix A3.

SOMs were used to define Homogeneous Environmental Conditions (HEC) based on first, the original set of selected environmental variables by training a Kohonen map with a total of 255 yield records, creating an input matrix trained with a SOM of 96 neurons (12 x 8); and second on those variables identified as important by the robust regressions and ANN non-linear regressions by generating an input matrix trained with 78 neurons (13 x 6). HECs take into account both temporal and spatial variability; thus a particular farm may fall into different HECs according to changes in weather conditions.

5.2.4.5 MIXED MODELS

Mixed models were selected as they include both, random and fixed effects in the analysis. The STATA statistical package was used to develop a linear mixed model called Best Linear Unbiased Prediction (BLUP) 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 datasets with many categorical variables as a result of the exponential increase in volume linked to adding extra dimensions to a mathematical space (Bellman, 1961; O'Grady and Medoff, 1988), mixed models are well-suited to perform this task. BLUP are estimates of the realized values of an output as linear functions of the random variables; they are unbiased in the sense that the average value of the estimate is equal to the average value of the quantity being estimated; they are best in the sense that they have minimum MSE within the class of linear unbiased estimators; and predictors to distinguish them from estimators of fixed effects. The mixed models assumed linear effects of 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 datasets that included the categorical variables HEC, locations and farms.

5.2.4.6 REGRESSION MODEL TESTING

In order to provide a mechanism for testing model performance and to compare different models or network topologies, both training and validation datasets were created in FENNIX by random sampling without replacement from the whole dataset for both robust regressions and MLP. In this way, each robust regression or MLP model regression was performed using 80% of the whole dataset, the model performance was assessed on the remaining 20%. The method is the same as the one we used for Andean blackberry and mentioned in chapter 3. This time, in order to compare the MLP model and the robust regression model, the split-sample procedure was run 100 times. The 100 yield estimates were then used to estimate the coefficient of determination (R^2) and the confidence limits of both the MLP and robust regression models.

5.3 RESULTS AND DISCUSSION

5.3.1 REGRESSIONS

5.3.1.1 SELECTION OF VARIABLES

In the iterative analysis process, input datasets 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).

A Pearson correlation calculated in XLSTAT for MS Excel, 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 5.2).

Table 5.2. 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 5.1

Similar to the procedure followed with Andean blackberry, the decision of which variable to retain was made on the basis of expert knowledge. In the case of Nariño-Cartago-San Isidro, thorn or thornless (variety), the Nariño-Cartago-San Isidro categories for location were retained as the use of a thornless variety was considered to be just one of the several management factors that might be associated with that particular location (see chapter 2). External drainage was chosen over internal drainage as a single variable contains the whole information.

Variable average temperatures for the harvest month, first month before harvest and second month before harvest were strongly correlated: variable average temperature of the harvest month was retained. Likewise, accumulated precipitation of the harvest month, the first and second months before harvest were strongly correlated with temperature range throughout the different months; the variable temperature range was maintained instead of accumulated precipitation as our experts (experienced extension officers) on lulo suggested that it is most likely that temperature range has an effect on lulo yield, rather than 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 5.1).

5.3.1.2 PERFORMANCE ANALYSIS AND MODEL INTERPRETATION

Mean R^2 from the 100 validations subsets was 0.69 for MLP and 0.65 for the robust regression model (Table 5.3). Distribution of R^2 provided by each approach was similar (Figure 5.5) with a 95% confidence interval 0.67 - 0.70 for MLP regression, and 0.63 - 0.66 for the robust regression. Both models explained more than 60% of yield variability at P= 0.05. The R^2 of 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 MLP explained significantly more variation (69%) than the robust regression (65%).

Table 5.3. R^2 of predicted versus real lulo yield provided by both regressions, using 100 validation datasets

Regression	*R^2 (mean)*	*Confidence interval (95%)*
Robust (linear)	0.65	0.63 - 0.66
MLP (non-linear)	0.69	0.67 - 0.70

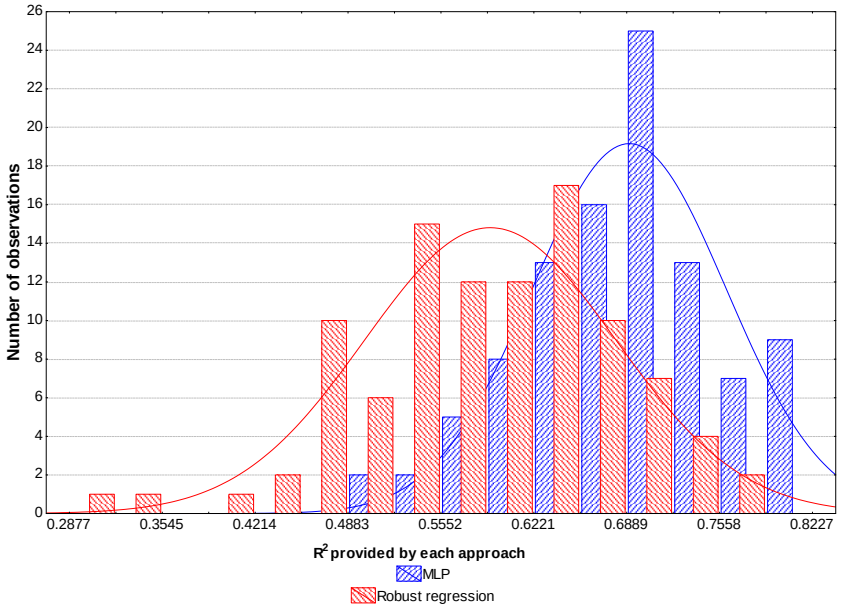


Figure 5.5. Distribution of R^2 obtained with each model

One of the steps followed to develop the robust regression, included the computation of a forward stepwise addition procedure (Tomassone et al., 1983). This method was used to add step-by-step one predictor and assess the change in MSE of the model. The change in 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 total yield variation, average temperature of the harvest month 11 % and effective soil depth 4% of total variation (Table 5.4).

In the case of MLP, and in order to identify the variables which contribute most to yield, we used the relevance metric described in chapter 3. This method assesses input relevance by calculating the partial derivative of the output of the neural network with respect to each of the inputs. Thus, the greater the partial derivate, the more relevant is the variable.

The sensitivity metric in the MLPs identified effective soil depth, average temperature of the harvest month, slope and locality Nariño-La Union-Chical Alto as the most important variables associated with yield variation (Figure 5.6). 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 modelling Andean blackberry yield.

Table 5.4. Variables explaining lulo yield according to a forward stepwise procedure

Variable Added	R^2	R^2 *due to variables*	*% 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

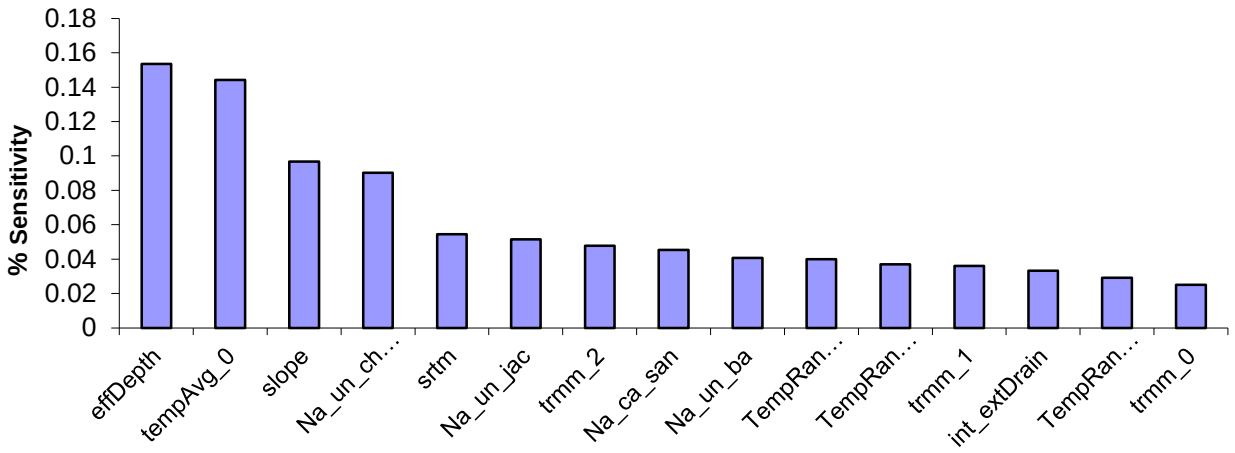


Figure 5.6. Sensitivity distribution of the MLP model with respect to inputs

5.3.1.3 MIXED MODELS AND SELF-ORGANIZING MAPS

Table 5.5. Variables integrated into the mixed model

Variable	*Abbreviation*
Biophysical data used in regressions [a b]	see Table 5.1
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

In various crops, attempts have been made to define the major environments where crops are grown and homogeneous or mega-environments in which similar varieties or crops could 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 by 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 analyse the effects of management practices on cane and sugar yield within and across AEZs using farmers' production experiences.

In the case of sugarcane, AEZs were based on expert opinion and an intimate knowledge of the crop and its response to variation in climate and soil conditions (see chapter 2). The idea behind the HECs, AEZs and mega-environments is that crop response in any one HEC, AEZ or mega-environment is uniform or homogeneous. With lulo, we were not able to define AEZs in the same manner as with sugarcane, cassava or wheat as there was not sufficient expert knowledge of the crops response to variation in soil and climatic conditions. Hence, we explored an iterative approach to defining HECs for lulo. The first step to identify production conditions that were homogeneous in terms of environment and 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 Kohonen map and identify clusters of HECs (Figure 5.7a). The Davies-Bouldin index indicated that there were six major HECs (Figure 5.7).

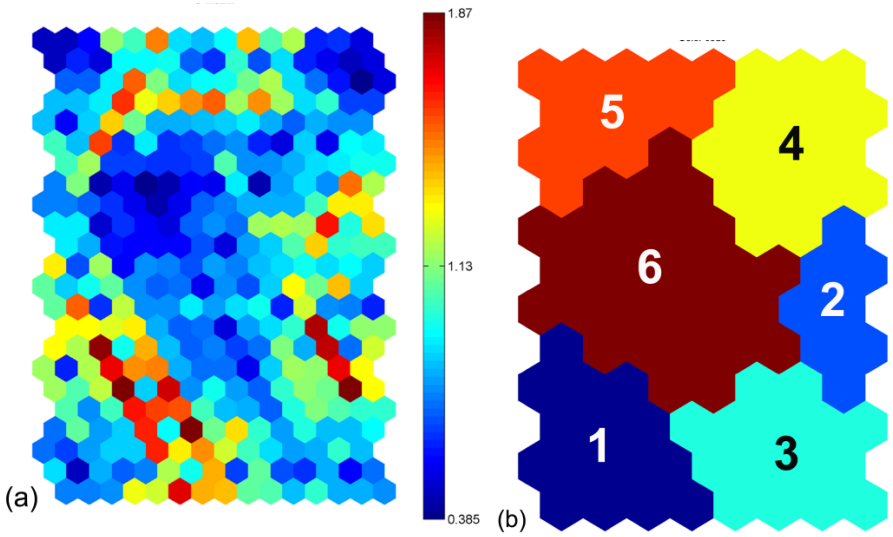


Figure 5.7. (a) U-matrix displaying the distance among prototypes. The scale bar (right) indicates the values of distance. The upper side exhibits high distances, whilst the lower displays low distances; (b) Kohonen map displaying the 6 clusters obtained by the K-means algorithm and the Davies–Bouldin index

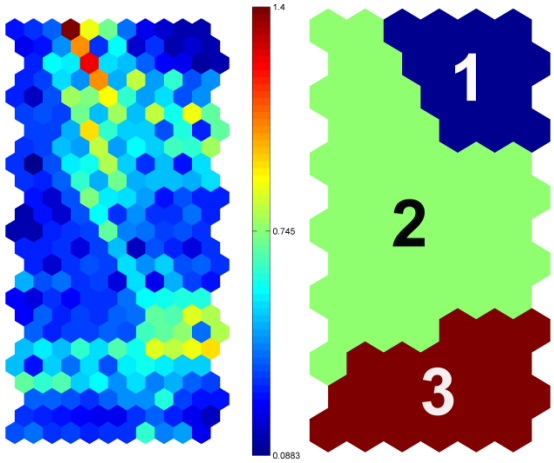
These six HECs were then incorporated into a mixed model together with the variables farm and location (Table 5.5). Farm and location were both incorporated as proxy variables for crop management on the assumption that HECs covered the variation due to environmental variation and that the remaining variation must be due to management. Furthermore, we hypothesized that (a) management in any one geographical location might be similar due to the sharing of ideas between farmers; and (b) even in the same location there would undoubtedly be managerial differences between farms. In previous studies, variable “location” or “site” were 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 variation (Table 5.6). However, the single variable farm explained 70% of variation, location 8% and the HECs a negligible amount (less than 1%).

Based on the MLP and robust regressions, in which environmental variables explained more than 60% of variation with 95% confidence limits, we had expected HECs to explain a much larger proportion of variation. This suggests 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 is that the variables used to develop the clusters were inappropriate with variation encountered in several variables being irrelevant in terms of crop development.

From both, MLP and 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 used previously. As a result three HECs were identified (Figure 5.8).



(a) (b)

Figure 5.8. (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 3 clusters obtained after using the K-means algorithm and the Davies–Bouldin index

A mixed model with the categorical variables of three HECs, location and farmer explained more than 80% of variation in lulo yield (Table 5.6). Variable HEC explained 61% of total variation indicating the extreme importance of environmental conditions in yield determination. Location explained less than 3% of variation in yield suggesting that differences in local practices between locations are of little importance in determining yield. This variable was not as relevant as we expected, considering the results suggested by the study with Andean blackberry. On the other hand, 19% of variation in yield, *ceteris paribus*, was attributed to farm suggesting that the management skills of individual farmers influenced yield. Furthermore, the high level of explanation of total variance by the HECs suggests that the method used to define them is effective.

Table 5.6. 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%

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, 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 5.8). The effect of the variable thorn or no thorn was not significant at the standard 5% level ($p = 0.168$) (Table 5.7). 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.

Table 5.7. Variance components of the mixed model estimations, including variety information

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

[a] Variable defined in Table 5.1
[n] Not statistically significant difference

Table 5.8. Variance components of the mixed model estimations, including variety information (Nar_thorn_N)

Random effects			
Parameters	**Estimate (g plant^{-1} wk^{-1})**	**Standard error**	**% variation of total**
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%

Inspection of Figures 5.9a and 5.9b gave clues as to how HECs, 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 5.10). 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 the combination of environmental conditions suitable for high productivity of lulo (Table 5.9).

Table 5.9. Environmental conditions for each HEC

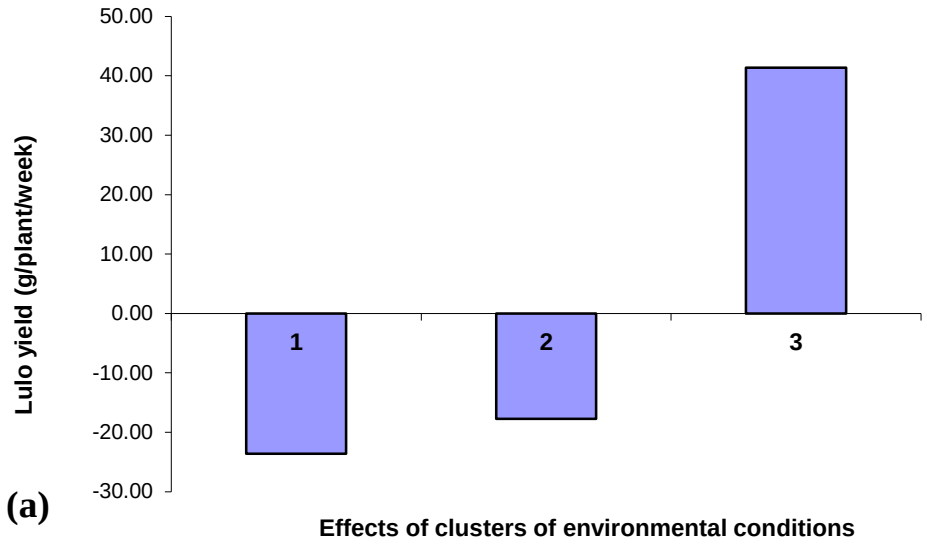
Variable ranges			*HEC*
Slope (degrees)	**EffDepth (cm)**	**TempAvg_0 (°C)**	
5-14	21-40	15 -16.5	1
8-15	32-69	15 -18.9	2
13-24	40-67	15.8 -19	3

Farms 5, 6, 16, 19, and 20 in HEC 2 and farms 7 and 9 in HECs had yields significantly different from the mean. A particular farm may fall into different HECs according to changes in environmental conditions such as: temperature or precipitation. 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 manage their crops effectively whereby those different environmental conditions do not greatly affect good management practices required to obtain higher than average yield. Farm 7 and 9 are in HEC 3 which in general presents 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 under similar environmental conditions 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 in yield variation associated with the farm, and little variation associated with location (Figure 5.7b).

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



(a)

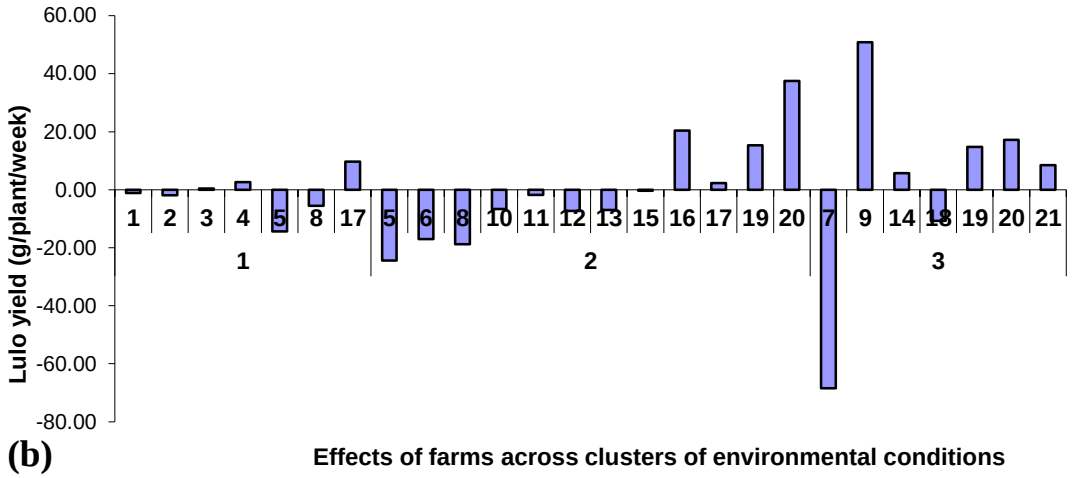


Figure 5.9. Clustered columns of the effects on lulo yield estimations: (a) effect of HEC, (b) effects of farms across HECs

Table 5.10. 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

5.4 CONCLUSIONS

Both parametric and ANNs models that used farmers' production experiences linked to characterization of the growing conditions explained more than 60% of variability in lulo yield. Multilayer perceptron neural network explained more variation (69%) than robust regression (65%).

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. 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 MLP was capable of discovering factors that were not identified to be important by robust regression.

Identification of HECs by taking all measured variables and using SOMs, 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 HECs were defined, it was possible to use a mixed model to analyse: 1) the effects of the environment using HECs as a categorical variable; 2) socio-economic conditions associated with geographic position of individual production units using location as a categorical variable; and 3) farm management skills using 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 farms.

The mixed model explained more than 80% of 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, farms with "good" management could readily be identified. Furthermore, observation of the range of conditions in 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 °C. It is also noteworthy that although in this dataset 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 analyse the data using a single methodology was much less powerful than an iterative guided approach using various methodologies. Both multiple linear regression and ANNs models were useful tools for analysing and interpreting 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 HECs had been determined. The mixed model has the advantage of effectively handling multiple categorical variables. Thus we suggest that when analysing farmers' production experiences, in order to interpret variation in yield, highly correlated independent variables should first be eliminated, and then either linear or non-linear 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 analyse variation in yield attributable to both environmental conditions (HECs) and social and management conditions.

In the particular case of lulo, proxies were used for social and management conditions. The analysis and interpretation of 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 for adequate analysis and interpretation of farmers' production experiences, especially for under-researched crops

6 GENERAL DISCUSSION AND CONCLUSIONS

6.1 GENERAL DISCUSSION

Experiences of farmers achieved under specific socio-economic and environmental circumstances, were analysed in this thesis. Through modelling tools developed by first using a database of sugarcane, it was possible to (a) identify the most relevant variables associated with productivity of Andean blackberry, and lulo, and (b) determine the effects on lulo yield, of location and variation within and between environmental clusters.

According to these results, optimal conditions that can led to high yields of Andean blackberry are: an average temperature between 16 and 18 °C, minimal effective soil depth of about 65 cm, and low rainfall during the first month before harvest in locations with poor external drainage, or with moderate to low rainfall in better-drained areas. In the case of lulo, the best conditions were 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°C. The models explained almost 90% of yield variation of Andean blackberry and more than 80 % in the case of lulo, through an iterative approach that considered the effect of location and environmental clusters.

Cock (2007) stated that "*The concept of observing crop response to variations in the environment and management is as old as agriculture itself*". That is essentially the approach of operational research (Operational Research Society, 2006), and this thesis supports that view. Hence, the observations made by farmers on their production experiences have been analysed to provide growers with useful information to optimize their production systems.

As far as we know, this is the first time that this methodology has been implemented for under-researched crops in general and in Colombia in particular, although it has been applied in other countries and to well-researched crops (Evans and Fischer, 1999; Schulz et al., 2001; Yan et al., 2002; Edwin and Masters, 2005; Lawes and Lawn, 2005; Welch et al., 2010; Erazo, 2011, Lacy, 2011).

The strategy of farmers collecting information on their own fields to be applied to their own production systems seems to be readily accepted by them compared to tryouts made under controlled conditions. The methodology used here supports the conclusions of others who have researched participatory research (Conroy et al., 1999; Rosenheim et al., 2011; Lacy, 2011). Nonetheless, the organization of the supply chain of each crop determines how data can be

collected and managed. Andean blackberry and lulo do not have strong growers' associations, so that, the collection of information required a different approach.

Cropping events in Andean blackberry and lulo were analysed based on the experience acquired in modelling events on sugarcane (Carbonell et al., 2001; Torres et al., 2004; Isaacs et al., 2007; Cock et al., 2011; Erazo, 2011). The studies conducted on sugarcane showed that it was feasible to estimate the environmental effects on crop performance by creating clusters with similar environmental conditions. In the case of sugarcane, data was available to determine these clusters and to establish both the effect and functions of the factors involved in crop performance. For Andean blackberry and lulo there was neither sufficient knowledge to indicate the likely non-linear response of any of the factors involved in production, nor expert criteria to define environmental clusters. Therefore, analytical approaches that were shown before allow to handle these conditions were used.

The most relevant factors linked to productivity were identified and then the effects within and between environmental clusters were estimated. The study conducted on Andean blackberry identified geographic areas with highest yields, and suggested that growers were managing their crops effectively in these geographical areas. In the case of lulo, the study illustrated the yield gaps between farms in similar environmental conditions.

Although there was insufficient information to determine precisely which management and social factors led to high yields, farmers who properly manage their fields were identified. This offers the chance for these farmers to spread their knowledge to other sites or farms with similar environmental conditions, so that they too can improve yields. In this manner, sharing production experiences between successful and less successful growers seems to be a powerful tool to increase productivity for under-researched crops in Colombia.

Tropical countries like Colombia are characterized by a lack of research on tropical fruit species. Information from weather stations does not represent the production areas well. There is no reliable soil data for the production sites, and the environment is extremely heterogeneous. Both the manner in which the data was collected and the analytical tools presented in this thesis seem to be promising tools to develop a SSCP program.

Technological packages for under-researched crops in Colombia have been developed by a handful of experienced agronomists who have documented these crops' performances under particular conditions. Often these packages are distributed as the official technological package for the crop, but given the extremely heterogeneous environment in which these crops occurs,

extrapolating them to other regions is hazardous. Typically, the optimal conditions for a particular crop defined in these packages are too general. For example, the ranges of temperature and precipitation are very wide, and often the specified soil texture and pH are those that most crops would tolerate (Franco et al., 2002; Franco and Giraldo, 2002).

Even if the results found here cannot be extrapolated outside the ranges of the variable values appearing in the collected datasets, the approach offers an adequate methodology to obtain more accurate information about the suitable conditions for growing under-researched crops in the tropics. Furthermore, the strategy of using the approaches of operational and participatory research, combined with farmers' production experiences, publicly-available environmental data and with data-driven models, is likely to provide growers with site-specific recommendations. They can use these to manage their crops better according to the specific conditions of their farms and hence develop a SSCP according to the socio-economic and environmental circumstances in Colombia.

6.2 CONCLUSIONS

6.2.1 FARMERS' PRODUCTION EXPERIENCES

Farmers showed that they can play a key role in the research process. They recorded their own data and estimated the characteristics of their soils. Producers of Andean blackberry and lulo were successfully trained in how to register information on their production sites.

By means of calendars developed by researchers working together with producers, 186 smallholder farmers registered information on 742 harvest events. Small-scale growers in Colombia record relatively little information on crop production. They do not have the habit of recording data on crop production and management factors such as disease and pest control and fertilizer application. It was therefore not possible to analyse which particular management practices were associated with high yields

My observations in the field suggest that small-scale farmers implement those practices that have rapid or immediate highly visible effects. These tend to be related to use of pesticides and fungicides but do not use fertilizer because it does not give immediate effects and moreover is costly. Hence, farmers pay little attention to this practice. According to my experience, a few do apply some fertilizer, commonly 15-15-15 (N-P-K). A few farmers also prune their crop, but do not take account of the phenological stage.

Soil information was provided by 41 growers with soil properties such as texture, structure, and surrounding terrain successfully recorded. Farmers could not measure other basic characteristics of soils such as slope, mottling, and pH, presumably because of the difficulty to obtain and manipulate the equipment required.

It is noteworthy that the inferred trait of effective soil depth, captured with RASTA, was integrated into the analyses made for both crops and found to be a relevant factor for both. This character has been identified by experienced Colombian extension officers working on fruit species, as one of the key factors explaining production of most tropical fruit species grown in the country (Franco et al., 2002; Franco and Giraldo, 2002; Orduz and Baquero, 2003; Bernal et al., 2005). This shows the great potential of RASTA to elucidate relevant factors that are likely to affect yield in other crops. Nevertheless, in order to draw firm conclusions of the effect of soil on yield, it is recommended that laboratory-based analysis of soil samples should be included in the future. This will provide more detailed information about soil chemical characteristics, which RASTA cannot provide.

Production data for sugarcane is accurately monitored by the sugar mills to calculate payment to farmers. In contrast, smallholder growers of Andean blackberry and lulo recorded information themselves for 742 cropping events. Figures 4.2 and 5.2 reflect the reality of production for these crops in Colombia. Clearly, collection of large numbers of farmers' production experiences over a wide range of conditions allows us to draw conclusions under the naturally-occurring variations that growers face. The analytical tools used in this research demonstrated that the analysis of farmers' production experiences is not the biggest challenge in working with under-researched crops. Rather, the collection of the information required to model the system is the major constraint. In general, tools used in this thesis showed that farmers' production information combined with publicly-available environmental data can be analysed as long as it is possible to collect sufficient data on how the growers manage their crop, and how much they produce. Nevertheless the smallholder producers who participated in this study and the experiences with CropCheck in Chile with small farmers show that they can provide useful information which can be analysed and used to improve their productivity (Lacy, 2011).

6.2.2 LINEAR MODEL OF RESEARCH AND TECHNOLOGY TRANSFER

Data generated at farmers' production sites where Andean blackberry and lulo crops are known to perform well was modelled to obtain insights about the combination of factors which contribute to both high and low yields. Research took into account the knowledge of both farmers and extension officers, and results are focused on combination of factors linked to productivity. This methodology complements the linear model of research, in the sense that it can also include formal experimental data where they are available. One example of the complementarity is when a new crop variety is produced in the linear model of research. In that case, the performance of that variety can be tested over a wide range of conditions, where factors vary naturally. But, this testing can be complemented, as in the case of lulo, to identify those producers whose yields are lower than those of the successful farmers, despite being under similar environmental conditions.

6.2.3 OPERATIONAL AND PARTICIPATORY RESEARCH

The conclusions of the research support others researchers' conclusions on the utility of operational and participatory research. From farmers' systematic observations, it was possible to obtain important insights to support them on how to be more efficient managing Andean blackberry and lulo. This approach of collecting data from observations made by farmers, and harvesting events, seems to be a feasible approach to develop SSCP for under-researched crops in Colombia.

In terms of participatory research, participatory on-farm research (Conroy et al. 1999) farmers participated in the research presented in this thesis in two different ways. In the consultative mode, farmers collected information on their own farms using tools developed by researchers. In the collaborative mode, as farmers participated and suggested ways to make the tools developed by the researchers easier-to-use. For example, the format of the calendars and the RASTA methodology were modified several times to take into account suggestions made by farmers.

6.2.4 PUBLICLY-AVAILABLE ENVIRONMENTAL DATA

Through the use of digital environmental surfaces of climate and topography, it was possible to characterise the environments of 742 production events. The effectiveness of the data provided by the publicly-available environmental data is demonstrated throughout chapters 4 and 5. In both chapters, temperature was found to be relevant to model yield of Andean blackberry and lulo. Temperatures between 16–18°C and 16–19°C were associated with high productivity of Andean blackberry and lulo, respectively, whereas temperatures above 24°C were associated with low productivity of Andean blackberry. In both cases, temperature ranges agree with the author's field observations, and the expert opinion of extension officers who have strong experience with both

crops (Franco et al., 2002; Franco and Giraldo, 2002). These publicly-available environmental data offer better coverage of many areas than the official weather stations in Colombia, which are typically neither representative of the production sites nor available to farmers. For SSCP, GPS data for a farmer's field is sufficient to extract a description of its environment. For PA, more accuracy is needed as the land units to be analysed are smaller than those used for SSCP (see chapter 2).

6.2.5 ANALYTICAL TOOLS

The sugarcane database was useful to explore modelling techniques aimed to (a) identify variables that contribute most to predict outputs, (b) visualize input–input and input–output dependencies, and (c) determine clusters with homogeneous environmental conditions. Analysis of the sugarcane database using sensitivity metric showed that plant age and water balance the most important variables to predict the crop's yield (Satizábal and Pérez-Uribe, 2007). Applying the same methodologies explained 89% of yield variation of Andean blackberry and 82% of yield variation of lulo. The nine most important factors for Andean blackberry were soil depth, average temperature of the first month before harvest, specific geographical areas, and average temperature of the harvest month, average temperature of the second month before harvest, average temperature of the third month before harvest, external drainage, and accumulated precipitation of the month before harvest. In lulo the most determinant factors were slope, average temperature, soil depth, and one locality

SOM is a data exploration tool with the capability to process datasets with missing data. Through its "component plane" representation, SOMs allowed the identification of variables on which productivity of Andean blackberry depends. The variables identified as the most relevant for modelling Andean blackberry yield were displayed as lattices, likewise it was illustrated the range values of each lattice associated with productivity (see chapter 4).The same non-supervised ANN successfully determined clusters with similar environments where there is not enough data to define them.

With Andean blackberry, there was an important location effect. Effect of location within and between clusters of environmental conditions was estimated in the lulo study in which data was divided into three clusters each with relatively uniform environments. The lowest-yielding farm 7 in cluster 3 produced 68 g $plant^{-1}$ wk^{-1} less than average, while the highest-yielding farm 9 in the same cluster produced 51 g $plant^{-1}$ wk^{-1} more than average. The difference is almost certainly due to management, suggesting that farm 9 managed the crop more efficiently than farm 7.

For crops like the Andean Blackberry and lulo there are no standard agronomic packages that have been developed. Farmers rely on trial and error on their own fields and also the experiences of their neighbours. In the case of lulo the locality was of relatively little importance suggesting that there was little shared local knowledge between farmers. This was reinforced by the large differences between farms, even in the same locality with similar climatic conditions, indicating a wide variety of practices used by individual farmers. On the other hand with Andean blackberry, the locality was of greater importance, suggesting that farmers were indeed sharing information or obtaining information from the same source, and using more uniform practices within a locality.

We conclude that farmers through trial and error have developed practices adapted to their own conditions. In the case of lulo, insights on best practices reside in the mind of a handful of farmers and there are large opportunities for sharing information of the best farmers with others growers. In the case of the Andean Blackberry there are major opportunities for transferring technologies from high yielding localities to others with similar climatic conditions. In this research it was not possible to identify the particular management practices associated with high yields in a particular environment. However, it was possible to identify those farmers that obtained high yields. In the sugarcane sector in Colombia surveys of the practices used by growers who obtain the highest yields have been compared to those used by the majority of growers to identify the individual practices associated with higher yields in a given AEZ (Cock et al., 2011). This process of identifying the practices used by good farmers can readily be applied to Andean fruits now that techniques have been developed to identify those growers who manage their crops well in a given environment.

The addition of proxies was helpful to take into account management and social factors that were not captured by the data collection process and that are likely to be related to a given site. When modelling Andean blackberry productivity, geographical areas were included in an attempt to estimate the effect of variables not measured, two geographic areas were identified as locations with higher yields. These are Nariño-La Union-Chical Alto and Nariño-La Union-Cusillo Bajo, which suggests that farmers in these areas manage their crops more effectively compared to elsewhere, even though the data was not precise enough to determine which management factors are related to higher yields.

In lulo, assuming that homogenous environmental conditions covered the environmental variation that explains yield, location was included as a proxy for socio-economic conditions, whilst farm was incorporated as proxy for crop management. HECs and farm, respectively, explained more than 61% and 19% of total variation, suggesting that environmental conditions and farmers'

management skills are strongly associated with yield. In the study of Andean blackberry, location was linked to yield variation.

In the case of lulo, an iterative methodology that combined ANNs with traditional statistical analysis gave the best methodology to interpret the variation in farmers' production experiences. This agrees with one of the conclusions drawn from the survey on applications of ANNs in agriculture (see chapter 3), in which hybrid approaches performed better than traditional techniques.

In this thesis, it has been demonstrated that ANNs techniques could be integrated into traditional analysis. Although, we had hoped it would be possible to use fully-automated analytical methods, lack of sufficient knowledge on the most relevant parameters indicated the need for an iterative approach. ANNs were an effective tool for managing the highly variable, noisy and qualitative data collected by farmers and linking these farmers' data to publicly-available environmental databases. The mixed model was also effective to integrate qualitative data with multiple categorical variables of the same class.

6.3 LIMITATIONS OF THE RESEARCH

Based on the assessment of the whole approach of developing SSCP for under-researched tropical fruit species; several considerations should be made.

One critical factor is the quality of the data collected. As might be expected, the farmers' data contained errors, such as: values of plant distances out of the range, or yields in different units such as boxes. To the extent possible, these were corrected, for example converting boxes, which are usually a standard size, to kilograms. A solution might be the use of Information and Communication Technologies (ICTs) such as mobile phones where recording forms, units, and type are standardized and where there is control in real-time. Several management practices, such as fertilizer and pesticide applications, which are likely to affect yield, were not recorded by the farmers. Since they were so receptive to the RASTA methodology, further training might be useful to obtain this important information.

It would be worthwhile to create a website that could be accessed by farmers to enter their data directly, with interactive data checking. This would be an efficient way to obtain on-farm data, but it would not be practical until access to the internet becomes more widely available in rural areas than it is now.

Regarding the modelling tools implemented in this research, several datasets were tested in order to construct data-driven models that were then implemented to model Andean blackberry and lulo yield. Among these databases, and in addition to the sugarcane database from Colombia, the *Centre de Coopération Internationale en Recherche Agronomique pour le Développement* (CIRAD) was also interested to apply artificial neural network-based models to a database of sugarcane from Réunion island. Different subsets of databases from "data-rich-systems" of sugarcane from Colombia and Réunion, were thus analysed. In the case of the datasets processed from Réunion island, ANNs tools used were not capable of generating correct responses when new input data was presented. Experts on modelling real world data at HEIG-VD, suggested that when the distribution of the output variable is to narrow, low generalization capabilities may occur. According to the latter, that distribution of data makes the Back-propagation algorithm to converge to a local minimum of the MSE function which sets the output of the network to the average value of the target. They suggest conducting more experiments focused on the use of other algorithms developed to minimize error, or also to transform the output by means of mathematical functions so as to obtain a close-to-linear distribution of values form the minimum to the maximum values. However, these types of experiments are out of the scope of this thesis, which deals explicitly with the development of a SSCP for under-researched tropical fruit species in Colombia.

Another research direction that should be considered is to keep testing more data-driven approaches or combination of ANNs models with other statistical techniques in order to obtain better results, and explore their applicability to SSCP development. When modelling lulo productivity for instance, the robust regression applied with a stepwise procedure and the sensitivity analyses of the MLP were in general agreement. Thus, for future studies, it is recommended that before to select the best approach to model a given phenomenon, the Occam's razor principle, should be taken into consideration. This principle states that "the simplest model is more likely to be correct-especially when we are working with an unusual phenomenon".

In this regard, for future work directed to increasing productivity of under researched and neglected crops we recommend that more attention be paid to (a) obtaining and compiling information on a wider range of farmers' production experiences followed by an evaluation of the techniques needed to analyse this wider information base (b) improve the methodologies that define clusters of environmental conditions, (c) explore methods for identifying factors that lead to high productivities within each cluster, and (d) formation of farmer discussion groups that can test the hypothesis that more successful growers will share information with less successful farmers that will assist them in improving productivity.

Regarding the determination of homogenous environmental areas, in this research we made the assumption that environmental variables are constant over the time. Not only do the weather patterns change from year to year, but there are specific phenomena like the *el niño*-southern oscillation, which can alter regional climatic similarities. Therefore, it is recommended that in future; we would take into account this temporal context when areas with homogenous environmental conditions are determined. In addition, in future, it will be necessary to study in depth the effects of year and location, especially when validation techniques similar to those employed in this research are used. According to the study conducted by Landschoot et al. (2012) these validation techniques might generate unrealistic predictions.

Although it is undoubtedly much easier to collect and analyse data for crops, such as sugarcane, coffee, and palm, with well-organized, strong supply chains, than those with weak supply chains as is the case of Andean blackberry and lulo, we have shown that it is possible to collect and analyse pertinent information from farmers' production experiences. These experiences can provide insights on how productivity can be improved under specific environmental conditions.

6.4 FUTURE PERSPECTIVES

The strategy to develop SSCP for under-researched crops presented in this thesis, used yield, crop data, and other information collected by the smallholder growers themselves. These data were combined with publicly-available environmental data to estimate the environmental adaptation of the crops.

There is an interest in Colombia to continue with the idea of SSCP for tropical fruit species. Currently, there is a project called "Site-Specific Agriculture based on Farmers Experiences" (SSAFE) which has included the methodology presented in this thesis. This time, the project, in addition to farmers' providing yield, crop, and soil data, will collect more detailed information on management and farmers' knowledge. The initiative is initially focused on four high-value but still under-researched crops, where in contrast to Andean blackberry and lulo, there is a more organized supply chain. The crops selected are: plantain (*Musa balbisiana)*, mango (*Mangifera indica)*, avocado (*Persea americana)* and citrus (*Citrus* spp.)

In Colombia, it is believed that there is significant potential in SSCP for contributing to income generation for growers in the tropics. It is also believed that an interaction with farmers is mandatory in order to accomplish this task, especially by disseminating the results provided by this approach. We propose therefore that the cultivation of tropical fruit species in tropical countries may benefit from the application of SSCP in order to increase both quality and productivity.

BIBLIOGRAPHY

Acosta, O., Perez, A., Vaillant, F., 2009. Chemical characterization, antioxidant properties, and volatile constituents of naranjilla (*Solanum quitoense* Lam.) cultivated in Costa Rica. ALAN 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.

Aitkenhead, M.J., Dalgetty, I.A., Mullins, C.E., McDonald, A.J.S., Strachan, N.J.C., 2003. Weed and crop discrimination using image analysis and artificial intelligence methods. Computers and Electronics in Agriculture 39, 15 -171.

Altieri, M.A., 2002. The science of natural resource management for poor farmers in marginal environments. Agriculture, Ecosystems and Environment 171, 1–24.

Alvarez, D.M., Estrada, M., Cock, J.H., 2004. RASTA (Rapid Soil and Terrain Assessment). Facultad De Ciencias Agropecuarias. Universidad Nacional De Colombia, Palmira, p. 91.

Arca, B., Benincasa, F., Vincenzi, M., 2001. Evaluation of neural network techniques for estimating evapotranspira. Engineering Application of Neural Networks Conference, Cagliari, pp. 62-69.

Arellano, O., 2004. An Improved Methodology for Land-Cover Classification Using Artificial Neural Networks and a Decision Tree Classifier. Department of Geography. University of Cincinnati.

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.

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 modeling. In: Proceedings of the 6th European Conference on Ecological Modelling (ECEM'07). The 6th European Conference on Ecological Modelling ECEM'07, Trieste, Italy. 27-30 November, pp. 55-56.

Barreto, M., 2012. Bio-inspired computational techniques applied to the clustering and visualization ofspatio-temporal geospatial data. Université de Lausanne, Faculté des Hautes Etudes Commerciales (HEC), Département des Systemes d'Information (ISI). Switzerland.

Basso, B., Ritchie, J.T., Pierce, F.J., Braga, R.P. and Jones, J.W., 2001. Spatial validation of crop models for precision agriculture. Agricultural Systems 68, 97-112.

Bell, T.L., 1987. Space-Time Stochastic Model of Rainfall for Satellite Remote-Sensing Studies. Journal of Geophysical Research-Atmospheres 92, 9631-9643.

Benor, D. and Harrison.,J.Q., 1977. Agricultural Extension: The Training and Visit System World Bank.

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

Bernal, J., Diaz, C., Tamayo, A., Córdoba, O., Tamayo, P., Londoño M., 2005. Tecnología para el cultivo de aguacate. Manual Técnico 5. Corpoica. Centro de Investigación La Selva. Rionegro, Antioquia.

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

Białobrzewski, I., 2008. Neural modeling of relative air humidity. Ecological Modelling 60, 1-7.

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

Bioversity International., 2005a. Information Sheet on *Rubus glaucus* in New World Fruits Database.URL: http://www.bioversityinternational.org/databases/new_world_fruits_database/search.html. Accessed on August 10th 2009.

Bioversity International., 2005b. Information Sheet on *Solanum quitoense* in New World Fruits Database. URL: http://www.bioversityinternational.org/databases/new_world_fruits_database/search.html. Accessed on August 16th 2009.

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

Boaventura, J., 2003. Greenhouse Climate Models: An Overview. EFITA 2003, Debrecen, Hungary.

Bocco, M., Ovando, G., Sayago, S., 2006. Development and evaluation of neural network models to estimate daily solar radiation at Córdoba, Argentina. Pesquisa Agropecuária Brasileira 41, 179-184.

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.

Bongiovanni, R. and Lowenberg-Deboer, J., 2004. Precision Agriculture and Sustainability. Precision Agriculture 5, 359–387.

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

Broner, I., Comstock, C.R., 1997. Combining expert systems and neural networks for learning site-specific conditions. Computers and Electronics in Agriculture 19, 37–53.

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

Burks, T.F., Shearer, S.A., Heath, J.R., Donohue, K.D., 2005. Evaluation of Neural-network Classifiers for Weed Species Discrimination. Biosystems Engineering 91, 293-304.

Carbonell, G.J., Amaya, E.A., Ortiz, B.V., Torres, J.S., Quintero, R. and Isaacs, C., 2001. Zonificación agroecológica para el cultivo de caña de azúcar en el valle del río Cauca. (Agro-ecological zoning for the sugarcane crop in the Cauca River Valley.) Tercera aproximación. Technical Series CENICAÑA no 29. Cali, Colombia.

Cassman, K., 1999. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. Proceedings of the National Academy of Sciences of the United States of America 96, 5952-5959.

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

CENICAÑA, 2006. Informe anual 2005. Centro de Investigación de la Caña de Azúcar de Colombia (CENICAÑA). Servicio de análisis económico y estadístico, Cali, p. 96.

Chambers, R. and Ghildyal, B., 1985. Agricultural Research for Resource Poor Farmers -The Farmer First and Last Model. Agricultural Administration 20, 1-30.

Chambers, R., Pacey, A. and Thrupp, L., 1989. Farmer First: Farmer Innovation and Agricultural Research. Intermediate Technology Publications. London.

Chon, T.S., Park, Y.S., Moon, K.Y., Cha, E.Y., 1996. Patternizing communities by using an artificial neural network. Ecological Modelling 90, 69-78.

Chung Lu, H., Hsieh, J.H., Chang, T.S., 2006. Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network. Atmospheric Research 81, 124–139.

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

Cock, J. and Luna, C.A.,1996. Analysis of Large Commercial databases for decision making. In:Sugar 2000 Symposium. (Eds.).CSIRO, Brisbane, Australia, pp. 24-25.

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.

Cock, J., Oberthür, T., Isaacs, C., Läderach, P., Palma, A., Carbonell, J., Watts, G., Amaya, A., Collet, L., Lema, G. and Anderson, E., 2011. Crop Management Based on Field Observations: case studies in sugarcane and coffee. Agricultural Systems 104, 755-769.

Conroy, C., Sutherland. and Martin A., 1999. Conducting farmer participatory research: what, when and how to be published in Decision Tools for Development. Ian Grant and Chris Sear (Eds). NRI Chatham.

Danielson, R.E. and Sutherland, P.L., 1986. Porosity. In: Klute.A. (Ed.), Methods of soil analisys.Part 1. Physical and mineralogical methods. American Society of Agronomy/ Soil Science society of America, Madison, pp. 443-462.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 1, 95-104.

Deadman, P., Gimblett, H.R., 1997. An Application of Neural Net Based Techniques and GIS for Vegetation Management and Restoration., AI Applications.

Diamantopoulou, M.J., 2005. Artificial neural networks as an alternative tool in pine bark volume estimation. Computers and Electronics in Agriculture 48, 235-244.

Dietterich, T.J., 2000. "Ensemble Methods in Machine Learning". In: Multiple Classifier Systems First International Workshop (MCS 2000), Cagliari, Italy, pp. 1-15.

Dimopoulos, I., Chronopoulos, J., Chronopoulou-Sereli, A., Lek, S., 1999. Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city. Ecological Modelling 120, 157–165.

Edgerton, D., 2004 The linear model' did not exist: Reflections on the history and historiography of science and research in industry in the twentieth century. In: Karl Grandin and Nina Wormbs (Eds.). The Science-Industry Nexus: History, Policy, Implications. New York, Watson.

Edwin, J. and Masters, W.A., 2005. Genetic Improvement and Cocoa Yields in Ghana. Experimental Agriculture 41,491-503.

Efron, B., 1983. the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association 78, 316-331.

Erazo, E.O., 2011. Efecto de los factores limitantes de la productividad del cultivo de caña de azúcar a nivel intrasuerte, Universidad Nacional De Colombia, Palmira.

Estrada, l.E., 1992. Genetic Potential Of Lulo (*Solanum quitoense* Lam.) And Factors That Limit Its Expression. Acta Horticulturae 310,171-182.

Evans, L.T. and Fischer, R.A., 1999. Yield Potential: Its Definition, Measurement, and Significance. Crop Science., 39:1544-1551.

Evenson, R., 1981. Developing a state extension program - TOPCROP in Victoria.

FAO-Unesco.1974. Soil Map of the World 1:5 000 000. Vol.1. Legend. Unesco. Paris. pp. 59.

Faraway, J.J., 2002. Practical Regression and Anova using R.Available from the R Project. http://cran.r-project.org/ 213 p.

Farkas, I., Reményi, P., Biro, A., 2000. Modelling aspects of grain drying with a neural network. Computers and Electronics in Agriculture 29, 99–113.

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. and Magnitskiy, S., 2008. Growth of lulo (*Solanum quitoense* Lam.) plants affected by salinity and substrate. Revista Brasileira de Fruticultura 30, 402-408.

Foody, G.M., 1999. Applications of the self-organising feature map neural network in community data analysis. Ecological Modelling 120, 97–107.

Foody, G.M., Cutler, M.E.J., 2006. Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks. Ecological Modelling 195, 37-42.

Framingham heart study., 2006. Framingham Heart Study. A project of the national heart, lung, and blood institute and Boston university. URL: www.nhlbi.nih.gov/about/framingham. Accessed on July 7th 2006.

Francl, L.J., 2004. Squeezing the turnip with artificial neural nets. Phytopathology 94, 1007-1012.

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, Manizales.

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

Gauch, H.G. and Zobel, R.W., 1997. Identifying mega-environments and targeting genotypes. Crop Science 37, 311-326.

García, J. ,2003. Evaluación del crecimiento de dos ecotipos de lulo amazónico (*Solanum sessiliflorum* Dunal) bajo tres ambientes en el piedemonte amazónico del Caquetá. Universidad Nacional De Colombia, Bogotá, p. 63.

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.

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

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.

Glezakos, T.J., Moschopoulou, G., Tsiligiridis, T.A., Kintziosb, S., Yialouris, C.P., 2010. Plant virus identification based on neural networks with evolutionary preprocessing. Computers and Electronics in Agriculture 70, 263-275.

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.

Granitto, P.M., Navone, H.D., Verdes, P.H., Ceccatto, H.A., 2000. Automatic Identification Of Weed Seeds By Color Image Processing. VI Argentine Congress on Computer Science Ushuaia, Argentina.

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.

Grijalba, C., Calderón, L., Pérez, M., 2010. Rendimiento y calidad de la fruta en mora de castilla (*Rubus glaucus* Benth), con y sin espinas, cultivada en campo abierto en Cajicá – Cundinamarca-Colombia. Revista de la Facultad de Ciencias Básicas, Bogotá 6, 24-41.

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.

Gupta, R., Narayana, B., Reddy, P.K., Rangarao, G.V., Gowda, C., Reddy, Y., Murthy, G.R., 2003. Understanding *Helicoverpa armigera* Pest Population Dynamics related to Chickpea Crop Using Neural Networks. Third International Conference on Data Mining. IEEE Computer Society Press, Florida, USA.

Guyer, D.E., Yang, X., 2000. Use of genetic artificial neural networks and spectral imaging for defect detection on cherries. Computers and Electronics in Agriculture 29, 179–194.

Hall, A., 2005. Capacity development for agricultural biotechnology in developing countries: an innovation systems view of what it is and how to develop it. Journal of International Development 17 ,611-630.

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

Higgins, A., Prestwidge, D., Stirling, D., Yostc, J., 2010. Forecasting maturity of green peas: An application of neural networks. Computers and Electronics in Agriculture 70, 151-156.

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.

Hijmans, R.J., Guarino, L., Jarvis, A., O'Brien, R., Mathur, p., Bussink, C., Cruz, M., Barrantes, I. and Rojas, E., 2005. DIVA-GIS Version 5.2.Manual.

Hilbert, D.W., Ostendorf, B., 2001. The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. Ecological Modelling 146, 311-327.

Hilera, J.R., Martínez, V.J., 1995. Redes neuronales artificiales: Fundamentos, modelos y aplicaciones, Rama, Madrid.

Himberg, J., 1998. Enhancing the SOM-based Data Visualization by Linking Different Data Projections. Proceedings of 1st International Symposium IDEAL'98, Intelligent Data Engineering and Learning–Perspectives on Financial Engineering and Data Mining. pp. 427-434.

Hoogenboom, G., Georgiev, G., Gresham, D., 2000. Development of weather based products for agricultural and environmental applications. Preprints of the 24th Conf. On Agricultural and Forest Meteorology,. American Meteorological Society., Boston, Mass, pp. 66-67.

Hsieh, W., 2009. Machine Learning Methods in the Environmental Sciences.Cambridge University Press.Cambridge.,UK.

Huang, K-Y., 2007. Application of artificial neural network for detecting Phalaenopsis seedling diseases using color and texture features. Computers and Electronics in Agriculture 57, 3-11.

Huffman, G.J., Adler, R.F., Rudolf, B., Schneider, U., Keehn, P.R., 1995. Global precipitation estimates based on a technique for combining satellite-based estimates, rain gauge analysis, and NWP model precipitation information. Journal of Climate 8, 1284-1295.

Isaacs, C., 1999. SEGUITEC. Seguimiento de tecnología. Carta trimestral. CENICAÑA. 21:25-29.

Isaacs, C., Carrillo, V.E., Caicedo, M., Paz, H.G. and Palma, Z., 2000. Los clientes de la nueva tecnología. Censo y tipificación de productores de caña de azúcar en la industria azucarera colombiana, (The new technology customers. Census and typification of sugarcane growers in the Colombian Sugarcane Industry). Technical Series CENICAÑA no 27. Cali, Colombia.

Isaacs, C., Carrillo, C., Anderson, A., Carbonell, C.J., Ortiz, U., 2004. Desarrollo de un sistema interactivo de información en web con el enfoque de agricultura específica por sitio. CENICAÑA Serie Tecnica 34.

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.

Jarvis, A., Reuter, H.I., Nelson, A. and Guevara, E., 2006. Hole-filled SRTM for the globe Version 3. CGIAR-CSI SRTM 90m Database: http://srtm.csi.cgiar.org.

Jiménez, D., Satizábal, H.F. and Pérez-Uribe, A., 2007. Modelling Sugar Cane Yield Using Artificial Neural Networks The 6th European Conference on Ecological Modelling, Trieste, Italy. 27-30 November, pp. 244-245.

Jiménez, D., Pérez-Uribe, A., Satizábal, 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-Verlag, 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.

Jiménez, D., Cock, J., Jarvis, A., Garcia, J., Satizábal, H.F., Damme, P.V., Pérez-Uribe, A. and Barreto-Sanz, M.A., 2011. Interpretation of commercial production information: A case study of lulo (*Solanum quitoense*), an under-researched Andean fruit. Agricultural Systems 104, 258-270.

Jones , P. and Gladkov , A., 2003. A Computer Tool for Predicting the Distribution of Plants and Other Organisms in the Wild. Version 1.02. Centro Internacional de Agricultura Tropical (CIAT): Cali, Colombia.

Jones, P., Diaz, W. and Cock, J., 2005. Homologue: A Computer System for Identifying Similar. Environments throughout the Tropical World. Version Beta a. Centro Internacional de Agricultura Tropical (CIAT): Cali, Colombia.

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.

Kavdır, I., 2004. Discrimination of sunflower, weed and soil by artificial neural networks. Computers and Electronics in Agriculture 44, 153–160.

Kehagias, A., Panagiotou, H., Maslaris, N., Petridis, V., Petrou, L., Spais, V., 1998. Predictive Modular Neural Networks Methods for Prediction of Sugar Beet Crop Yield. IFAC Conference on Control Applications and Ergonomics in Agriculture, Athens, Greece.

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: P. C. Robert, R.H.R., and W. E.Larson. Precision agriculture: Proceedings of the Fourth International Conference (Ed.), Madison, WI, USA, pp. 573–579.

Kim, M., Gilley, J.E., 2008. Artificial Neural Network estimation of soil erosion and nutrient concentrations in runoff from land application areas. Computers and Electronics in Agriculture 64, 268-275.

Koekoek, E.J.W., Booltink, H., 1999. Neural network models to predict soil water retention. European Journal of Soil Science 50, 489-495.

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

Kondo, N., Ahmad, U., Monta, M., Murase, H., 2000. Machine vision based quality evaluation of Iyokan orange fruit using neural networks. Computers and Electronics in Agriculture 29, 135–147.

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

Kummerow, C., Barnes, W., Kozu, T., Shiue, J., Simpson, J., 1998. The Tropical Rainfall Measuring Mission (TRMM) sensor package. Journal of Atmospheric and Oceanic Technology 15, 809-817.

Lacy, J., 2011. Cropcheck: Farmer benchmarking participatory model to improve productivity. Agricultural Systems 104, 562-571.

Läderach, P., 2009. Management of intrinsic quality characteristics for high-value specialty coffees of heterogeneous hillside landscapes. A Framework Developed and Tested in Coffee Growing Regions Across Latin America. Verlag Dr Müller, Saarbrücken, Germany.:1–157.

Läderach, P., Oberthür, T., Cook, S., Estrada Iza, M., Pohlan, J.A., Fisher, M. and Rosales Lechuga, R., 2011. Systematic agronomic farm management for improved coffee quality. Field Crops Research 120, 321-329.

Landschoot, S., Waegeman, W., Audenaert, K., Vandepitte, J., Haesaert, G. and De Baets, B., 2012. Toward a Reliable Evaluation of Forecasting Systems for Plant Diseases: A Case Study Using Fusarium Head Blight of Wheat. Plant Disease 96, 889-896.

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.

La Patria., 2011. Tener variedades resistentes a la roya no es suficiente si no las siembran. La Patria, Manizales. http://www.lapatria.com/story/tener-variedades-resistentes-la-roya-no-es-suficiente-si-no-las-siembran.

Lawes, R.A. and Lawn, R.J., 2005. Applications of industry information in sugarcane production systems. In: Sugarcane physiology: Integrating from cell to crop to advance sugarcane production.

N. Geoff Inman-Bamber and Graham D. Bonnett and Peter J. Thorburn and D. Mark Smith (Eds.). Field Crops Research 92, 353-363.

Levine, E.R., Kimes, D.S., Sigillito, V.G., 1996. Classifying soil structure using neural networks. Ecological Modelling 92, 101-108.

Li, B., 2002. Spatial Interpolation Of Weather Variables Using Artificial Neural Networks Artificial Intelligence. University of Georgia, Athens.

Liu, M. and Samal, A., 2002. A fuzzy clustering approach to delineate agroecozones. Ecological Modelling 149, 215-228.

Ljung, L., 1999. System Identification - Theory For the UserUser, 2nd ed., PTR Prentice-Hall, Inc, Upper Saddle River, N.J., USA.

Lyon, F., 1996. How farmers research and learn: The case of arable farmers of East Anglia,UK. Agriculture and Human Values 13, 39 - 47.

Marsh, S.P. and Pannell D., 2000. Agricultural extension policy in Australia: the good, the bad and the misguided. The Australian Journal of Agricultural and Resource Economics 44, 605-627.

Medina, C., I., Martinez, E., Lobo, M., 2008. Lulo (*Solanum quitoense* Lam) Biomass Partitioning Under Full Sunhine Light At The Low Mountain Rain Forest Of East Antioquia, Colombia. Revista Facultad Nacional de Agronomía, Medellín 61, 4256-4268.

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.

Montaner, J.G., 2004. Successful integration of research and extension combining private and public organizations: lessons from Argentina. 4th International Crop Science Congress. Crop Science Congress, Brisbane, Australia. 26 Sept- 4 Oct.

Montgomery, M.R., Gragnolati, M., Burke, K.A., Paredes, E., 1999. Measuring Living Standards with Proxy Variables. Demography 37, 155-174.

Morimoto, T., Hashimoto, Y., 2000. AI approaches to identification and control of total plant production systems. Control Engineering Practice 8, 555-567.

Moshou, D., Bravo, C., Oberti, R., West, J., Bodria, L., McCartney, A., Ramon, H., 2005. Plant disease detection based on data fusion of hyper-spectral and multi-spectral fluorescence imaging using Kohonen maps. Real-Time Imaging 11, 75–83.

Moshou, D., Vrindts, E., Ketelaere, D.B., Baerdemaeker, D.J., Ramon, H., 2001. A neural network based plant classifier. Computers and Electronics in Agriculture 31, 5–16.

Moshou, D., Ramon, H., Baerdemaeker, D.J., 2002. A Weed Species Spectral Detector Based on Neural Networks. Precision Agriculture 3, 209–223.

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.

Murase, H., 2000. Special issue:artificial intelligence in agriculture. Computers and Electronics in Agriculture 29, 1-2.

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.

Nakano, K., 1997. Application of neural networks to the color grading of apples. Computers and Electronics in Agriculture 18, 105-116.

National Research Council., 1989. Lost crops of the Incas: little known plants of the Andes with promise for worldwide cultivation., National Academy Press, Washington, D.C., USA.

National Research Council., 1997. Precision Agriculture in the 21st Century: Geospatial and Information Technologies in Crop Management. Committee on Assessing Crop Yield: Site-Specific Farming, Information Systems, and Research Opportunities. National Academy Press, Washington D.C. pp. 73.

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.

Noble, P.A., Tribou, E.H., 2007. Neuroet: An easy-to-use artificial neural network for ecological and biological modeling. Ecological Modelling 203, 87-98.

Noguchi, N., Reid, J.F., Zhang, Q., Tian, L.F., 1998. Vision Intelligence For Precision Farming Using Fuzzy Logic Optimized Genetic Algorithm And Artificial Neural Network. ASAE Paper 983034 St. Joseph, MI. UILU-ENG-98-7020.

Noguchi, N., Terao, T., 1997. Path planning of an agricultural mobile robot by neural network and genetic algorithm. Computers and Electronics in Agriculture 18, 187-204.

O'Brien, R., 2004. Spatial decision support for selecting tropical crops and forages in uncertain environments, Perth.

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.

Oide, M., Ninomiya, S., 2000. Discrimination of soybean leaflet shape by neural networks with image input. Computers and Electronics in Agriculture 29, 59-72.

Operational research society., 2006. http://www.orsoc.org.uk/orshop/(0tic0pjmqgos3ajjs1zaww55)/orhomepage2.aspx. Accessed on July 7th 2006.

Orduz, J.O. and Baquero.J., 2003. Aspectos básicos para el cultivo de los cítricos en el piedemonte llanero. Revista Achagua 7, 7-20.

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

Overton, M., 2006. Agricultural Revolution in England.

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

Park, S.J., Hwang, C.S., Vlek, P.L.G., 2005. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. Agricultural Systems 85, 59-81.

Park, Y.S., Cérégino, R., Compin, A. and Lek, S., 2003. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modelling, 160, 265-280.

Paruelo, J.M., Tomasel, F., 1997. Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. Ecological Modelling 98, 173-186.

Pasgianos, G.D., Arvanitis, K.G., Polycarpou, P., Sigrimis, N., 2003. A non-linear feedback technique for greenhouse environmental control. Computers and Electronics in Agriculture 40, 153-177.

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.

PAVUC., 2010. Producir valor agregado a partir de productos subutilizados. URL:http://www.pavuc.soton.ac.uk/Default.aspx. Accessed on September 10th 2009.

Peña-Reyes, C., 2002. Coevolutionary Fuzzy Modeling. Section d'informatique. Ecole Polytechnique Fédérale De Lausanne (EPFL), Lausanne, Suisse.

Pérez-Uribe, A., 1998. Artificial Neural Networks:Algorithms and Hardware Implementation. In: Tomassini, D.M.a.M. (Ed.), Bio-Inspired Computing Machines: Toward Novel Computational Architectures, PPUR Press, pp. 289-316.

Philip, N.S., Joseph, K.B., 2003. A neural network tool for analyzing trends in rainfall. Computers & Geosciences 29, 215-223.

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.

Pollock, C.J., 1990. The response of plants to temperature change. The Journal of Agricultural Science 115, 1-5.

Pretty, J., 1991. Farmers' Extension Practice and Technology Adaptation: Agricultural Revolution in 17-19th Century Britain. Agriculture and Human Values 8, 132-148

Proceedings of the 14th Annual Symposium on Precision Agriculture in Australasia., 2010. Centre for Precision Agriculture and SPAA Precision Agriculture Australia.14th Annual Symposium on Precision Agriculture in Australasia,Albury, New South Wales Australia. 2 - 3 September, pp. 5-55.

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

Quintero, R. and Castilla, C., 1992. Agrupación de los suelos del valle geográfico del río Cauca. (Soil groups in the Cauca) Technical Series CENICAÑA no 8. Cali, Colombia.

Rabe-Hesketh, S., Skrondal, A., 2008. Multilevel and Longitudinal Modeling Using Stata, 2nd edition, Stata Press,College Station, Texas. pp.156-160.

Raju, K.S., Kumar, D.N., Ducksteinc, L., 2006. Artificial neural networks and multicriterion analysis for sustainable irrigation planning. Computers & Operations Research 33, 1138–1153.

Ramos, F.A., Delgado, J.L., Bautista, E., Morales, A.L., Duque, C., 2005. Changes in volatiles with the application of progressive freeze-concentration to Andes berry (*Rubus glaucus* Benth. Journal of Food Engineering 69, 291–297.

Razali, N.M., Wah, Y.B. , 2011. Power comparisons of Shapiro-Wilk Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical modeling and Analytics 2, 21-33.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge.

Rist, G., 1997. The History of Development: from Western Origins to Global Faith. New Edition. Zed Books. London.

Robinson, G.K., 1991. That BLUP is a Good Thing: The Estimation of Random Effects. Statistical Science 6, 15-32.

Rosenheim, J.A., Parsa, S., Forbes, A.A., Krimmel, W.A., Law Y.H., Segoli, M., Segoli M, Sivakoff, F.S., Zaviezo, T. and Gross, K., 2011. Ecoinformatics for integrated pest management: expanding the applied insect ecologist's tool-kit. Journal of Economic Entomology 104, 331-342.

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

Russell, D.B. and Ison, R.L., 2000. The Research Development Relationship in Rural Communities: An opportunity for contextual science. In: Agricultural and Extension and Rural Development: Breaking out of traditions (Eds): Cambridge University Press. pp 10-29.

Sandoval, H., 2011. Cafeteros se resisten a sembrar variedad Colombia, pese a los estragos de la roya. La Republica Colombia, Bogotá. http://www.larepublica.com.co/archivos/ECONOMIA/2011-01-12/cafeteros-se-resisten-a-sembrar-variedad-colombia-pese-a-los-estragos-de-la-roya_118953.php.

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

Sarle, W.S., 1994. Neural Networks and Statistical Models. Proceedings of the Nineteenth Annual SAS Users. Group International Conference. SAS Institute.Cary, NC, USA, 1538-1550.

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. Computational and Ambient Intelligence, pp. 1147-1154.

Satizábal, H.F., Pérez-Uribe, A., 2007. Relevance Metrics to Reduce Input Dimensions. ICANN 07 International Conference on Artificial Neural Networks, Porto, Portugal. 9 – 13 September pp. 39-48.

Satizábal H.F., 2010. Using Biological Inspiration to Perform Incremental Modelling Tasks. Thesis. Université de Lausanne, Faculté des Hautes Etudes Commerciales (HEC), Département des Systemes d'Information (ISI). Switzerland.

Satizábal, H., Barreto-Sanz, M., Jiménez, D., Pérez-Uribe, A., Cock, J., Bolay, J.-C., Schmid, M., Tejada, G. and Hazboun, E., 2012. Enhancing Decision-Making Processes of Small Farmers in Tropical Crops by Means of Machine Learning Models. Technologies and Innovations for Development. Springer Paris, pp. 265-277.

Schank, R., 2011. Experimentation. Edge.org. John Brockman. http://www.edge.org/q2011/q11_2.html

Shamseldin, A.Y., 1997. Application of a neural network technique to rainfall-runoff modelling. Journal of Hydrology 199, 272-294.

Shearer, J.R., Burks, T.F., Fulton, J.P., Higgins, S.F., 2000. Yield Prediction Using A Neural Network Classifier Trained Using Soil Landscape Features and Soil Fertility Data Annual International Meeting, Midwest Express Center. ASAE Paper No. 001084., Milwaukee, Wisconsin.

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

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

Schulz, L.J., Storer, C.E., Murray-Prior, R. and Walmsley, T., 2001. Maintaining links with stakeholders in partnership extension models: Lessons learnt from TOPCROP West Australia http://regional.org.au/au/apen/2001/r/SchultzL.htm.

Seginer, I., 1997. Some artificial neural network applications to greenhouse environmental control. Computers and Electronics in Agriculture 18, 167-186.

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

Spaans, E. and L. Estrada., 2004. Sense and nonsense of satellite navigaton for precision agriculture in the tropics. European Journal of Navigation 2, 71-76.

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.

Tafur, R., 2006. Propuesta frutícola para Colombia y su impacto en la actividad económica nacional, regional y departamental. In: Fischer, G.M., D; Piedrahita, W;Magnitskiy, S. (Ed.), Memorias primer congresocolombiano de horticultura., Unibiblos, Bogotá, pp. 47-66.

Thomas, D., Strauss, J., Henriques., M., 1990. "Child Survival, Height for Age and Household Characteristics in Brazil." Journal of Development Economics. 33, 197-234.

Thompson, J. and Scoones, I., 1994. Challenging The Populist Perspective: Rural People's knowledge. Agricultural Research, And Extension Practice Agriculture and Human Values 11, 58-76.

Tien, B.T., van Straten, G., 1998. A NeuroFuzzy Approach to Identify Lettuce Growth and Greenhouse Climate. Artificial Intelligence Review 12, 71-93.

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

Torres, J.S., 1998. A simple visual aid for sugarcane irrigation scheduling. Agricultural Water Management 38, 77-83.

Torres, J.S., Cruz, V., R., and Villegas, T., 2004. Avances técnicos para la programación y el manejo del riego en caña de azúcar (TechnicalAdvances in irrigationprogamming in sugarcane). Technical Series CENICAÑA, Cali, Colombia 33, 39-44.

Tourenq, C., Aulagnier, S., Mesleard, F., Durieux, L., Johnson, A., Gonzalez, G., Lek, S., 1999. Use of artificial neural networks for predicting rice crop damage by greater flamingos in the Camargue, France. Ecological Modelling 120, 349-358.

Tuma, R.S., 2007. Statisticians set sights on observational studies. Journal of the National Cancer Institute 99, 664-668.

Uno, Y., Prasher, S.O., Lacroix, R., Goel, P.K., Karimi, Y., Viau, A., Patel, R.M., 2005. Artificial neural networks to predict corn yield from Compact Airborne Spectrographic Imager data. Computers and Electronics in Agriculture 47, 149-161.

Van Asten, P.J.A., Kaaria, S., Fermont, A.M. and Delve, R.J., 2009. Challenges and lessons when using farmer knowledge in agricultural research and development projects in Africa. Experimental Agriculture 45, 1-14.

Vellido, A., Lisboa, P. and Meehan, K., 1999. Segmentation of the on-line shopping market using neural networks. Expert Systems with Applications 17,303-314.

Veronez, M.R., Thum, A.B., Luz, A.S., da Silva, D.R., 2006. Artificial Neural Networks applied in the determination of Soil Surface Temperature – SST. 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences., Lisboa.

Vesanto, J., 1999. “SOM-based data visualization methods”. Intelligent Data Analysis 3, 111-126.

Vesanto, J. and Alhoniemi, E., 2000. Clustering of the Self-Organizing Map. IEEE Transactions on neural networks 11, 568-600.

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

Welch, J.R., Vincent, J.R., Auffhammer, M., Moya, P. and Dobermann, A., 2010 Rice yields in tropical/subtropical Asia exhibit large but opposing sensitivities to minimum and maximum temperatures. PNAS:www.pnas.org/cgi/doi/10.1073/pnas.1001222107.

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.

Yan, W., Rajcan, I., 2003. Prediction of Cultivar Performance Based on Single- versus Multiple-Year Tests in Soybean. Crop Science 43, 549-555.

Yang, C.-C., Prasher, S.O., Landry, J.A., Ramaswamy, H.S., 2003. Development of an Image Processing System and a Fuzzy Algorithm for Site-Specific Herbicide Applications. Precision Agriculture 4, 5–18.

Yao, X., Liu, Y., 1998. Making use of population information in evolutionary artificial neural networks,'. IEEE Transactionson Systems. Man Cybernetics 28, 417-425.

Zaidi, M.A., Murase, H., Honami, N., 1999. Neural Network Model for the Evaluation of Lettuce Plant Growth. Journal of Agricultural Engineering Research 74, 237-242.

Zee, F., Bubenheim, D., 1997. Plant Growth Model Using Artificial Neural Networks.

Zhai, Y., Thomasson, J.A., Boggess III, J.E., Sui, R., 2006. Soil texture classification with artificial neural networks operating on remote sensing data. Computers and Electronics in Agriculture 54, 53-68.

Zhao, Z., Chow, T., L., Rees, H.W., Yang, Q., Xing, Z., Rui Meng, F., 2009. Predict soil texture distributions using an artificial neural network model. Computers and Electronics in Agriculture 65, 36-48.

APPENDIX

APPENDIX A1.

RASTA

(Rapid Soil and Terrain Assessment)

A Practical Guide to the Characterization of Soils and Terrains

CIAT - Corporación BIOTEC- Universidad Nacional

James H. Cock
Diana M. Álvarez
Marcela Estrada

This *Guide* is divided into two parts: a) Basic Soil Properties and field observations, and b) Inferred Soil Properties

To start characterizing the soils on your farm

Divide your farm into study sites

Closely observe your plot or farm and divide it into units which are relatively uniform. This can be done by using your knowledge and experience on the farm and by using various indicators, for example, **soil colour** (if you find soils of different colours), **soil type** (if your soil is sandy, clayey, good, bad, poor, fertile, soft, or hard), **slope and topography** (if your land is flat or sloping), **current use** (if your land has a different use from plot to plot), **crop development** (if you observe differences in plant growth), **natural vegetation,** and any other differences that you have observed.

Dig the soil profile hole

In each selected unit, prepare a square “soil profile” hole according to the specifications indicated below.

However, if the area is very large, prepare several profile holes. Avoid making them near roadsides and plot edges, or in high parts.

The holes should have the following measurements: 60 cm long × 60 cm wide × 70 cm deep. **Note:** If you are going to plant perennial crops such as fruit trees, then dig the holes 150 cm deep.

Choose and evaluate the soil profile

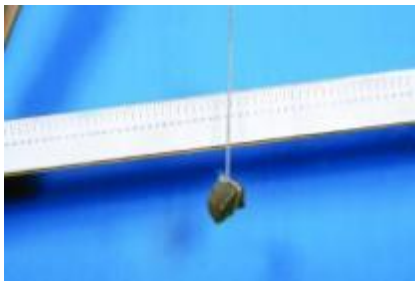
When you have finished making the soil profile hole, check its four faces, and select the one that is the most clearly seen (usually the one that is not in the direct sun). Take up your *Guide* and begin evaluating the soil profile. Do not forget that all results and the site's data should be written in the *Soil Records*

Part 1: Basic soil properties

Slope

Slope refers to the land's inclination. It is related to the retention and movement of water, erosion, use of machinery, soil conservation, and the adoption of field practices such as irrigation and drainage.

Step 1
After constructing a simple "A" level, find the direction of the slope and open the two arms of the level to their fullest extent (1 meter). With one arm straight up the slope and the other straight down, allow the plumb line to come to a standstill.



Step 2
Observe the value where the plumb line crosses the graduated ruler and note the result in *Soil Record Sheet No. 1*. *Answer Sheet No. 1*

Land form

The characteristics of the landscape influence soil properties and enable approximations of the most appropriate land use and management.

Surrounding terrain

Observe the landscape around you and compare it with the photos below. Identify the terrain around you and note your observations in *Answer Sheet No. 1*.

Flat or plains: the land is completely flat; no surrounding mountains can be seen

Undulating: the land has gentle slopes

Mountainous: large mountains are found all around you

Undulating and mountainous: the land has both gentle slopes and mountains

Profile's landscape position

Now describe the exact location of the spot where you are characterizing the soil. Compare the profile's position in the landscape with the following photos and select accordingly. Note your observation in *Answer Sheet No. 1*.

Tableland

Summit

Convex slope

Concave slope

Flat slope

Flat land

Flat land with undulations

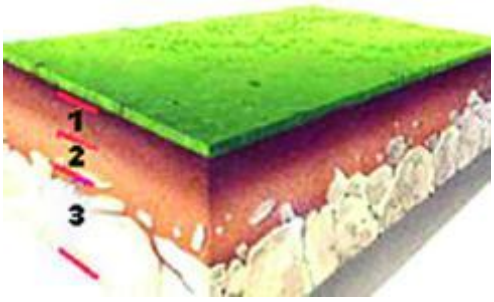
At the foot of a slope

Determining horizons, colour, and texture

Layers or horizons

A *horizon* is a distinct layer that can be seen in a soil profile. Horizons are differentiated from each other by colour, texture, structure, or pedregosity (e.g., stoniness).

You will need a tape measure and sticks to mark the limits of each horizon.

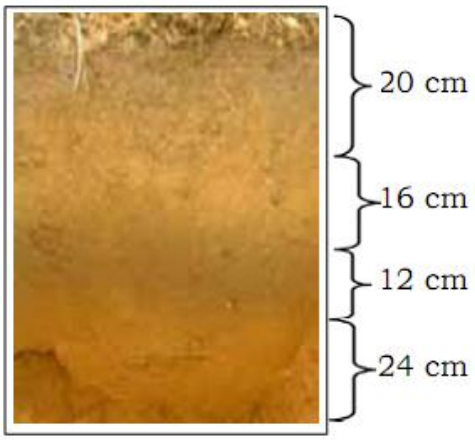


Observe the four faces of the soil profile hole and selecting the one that has the horizons most clearly defined. This face becomes the **SOIL PROFILE**. Then:

- Clean the chosen face with a clean machete or spade.

- Examine the face carefully and, in each place where a change of colour, texture, or structure is evident, mark it with a stick.

- Number each layer and measure its thickness. Note the results in *Answer Sheet No. 1*.



20 cm 16 cm 12 cm 24 cm

Color

Color indicates the soil's fertility, moisture content, parental material, and drainage conditions. For example, black or dark colours represent high organic matter content; red colours, the presence of iron; whitish colours, the presence of carbonates; and olive, green, or gray colours, bad drainage.

You will need a pencil and the attached colour chart. Then:

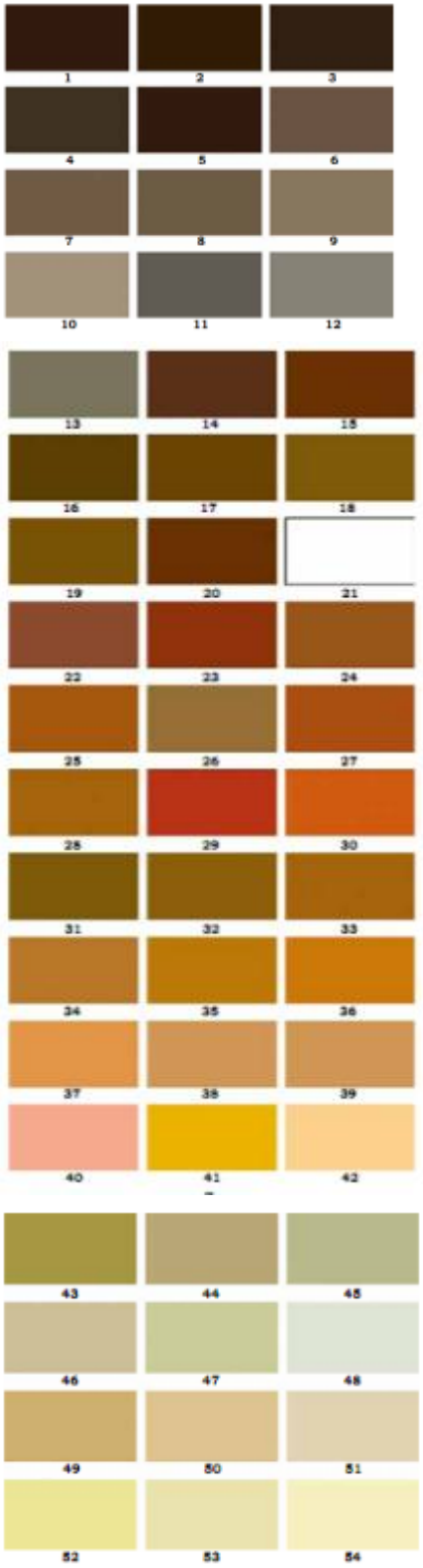
- Observe the soil's humidity: is it moist or dry?

- Collect a lump of soil from each layer and stand with your back to the sun, preventing sunlight from falling directly on the lump.

- Compare the soil with **all** the colours found in the following colour chart and select the most similar. If you find spots in the horizon, note the colour that predominates.

- Write down the colour beside the respective horizon in *Answer Sheet No. 1*.

Color chart



Texture

Feeling the soil

Soil is characterized by diverse particles of matter, the most important of which are classified by size into, for example, sands, loams, and clays. *Soil texture* derives from the proportions in which each particle type is found in the soil. These proportions are expressed in percentages (%).

Texture influences the retention and storage of oxygen and water; and soil fertility, porosity, and drainage, among other factors.

Our fingers can easily distinguish between soils with different particle sizes: we can detect the rough feel of sand, the silky texture of a loam and the stickiness of clays.

Preparing a sample
Have at hand water, the attached textural key, and soil.

Follow the steps indicated below, until you obtain the soil's texture. Note your observations in *Answer Sheet No. 1*, using the letters that appear in parentheses, for example, for **CLAY LOAM (CL)**, write "CL".

Step 1
Collect an easy-to-manage handful of soil.

Step 2
Add a little water so that the soil can be easily kneaded. Avoid forming mud, which can be difficult to manage. If you have added too much water, then add a little more soil, and continue to knead.

Step 3
Knead the soil well until it forms a **COMPLETELY HOMOGENOUS MASS, WITH NO LUMPS.** Keep in mind that if the soil has lumps it cannot form rolls or circles.

Textural key

Step 1
Take a lump of soil, which is well moistened, and then place it between your hands and by rubbing them together try to form a soil worm, that is, a roll that is as thick as a pencil. Once you have formed a soil worm try to shape it into a circle like a donut without its breaking. **If it is impossible to make soil worms moisten the soil more and try again.**

- The soil will not form a soil worm or the worm roll breaks on bending **Go to Step 2**

- The soil forms a soil worm and does **not break on** bending,
  Go to Step 3

Step 2
- Try to form soil balls and worms. If the balls are lumpy or the worms crack or split on bending,
  Go to Step 4

- It does not form balls or worms,
  Go to Step 5

Step 3
Take a small quantity of soil in the palm of your hand and add water. Rub the mixture with your index finger. If you feel that the soil is:

- Smooth and muddy, with some sand grains,
  Go to Step 13

- Rough, with many sand grains,
  Go to Step 14

- Soapy and very smooth; no visible sand grains,
  Go to Step 15

Step 4
Take a small quantity of soil in the palm of your hand and add water. Rub the mixture with your index finger. If you feel that the soil is:

- Soapy and very smooth, with no grains of sand,
  Go to Step 6

- Soft and you see some sand grains,
  Go to Step 7

- Rough and you see many sand grains,
  Go to Step 10

Step 5
You see separate, loose particles that form unstable pyramids. The soil is not sticky and does not stain fingers. Separate grains of sand can be seen. The soil is therefore:

SAND (S)

Step 6

The soil feels silky, like talcum powder, and smooth, is easy to knead, appears opaque, stains the fingers, is not sticky, and, when kneaded, is buttery. The soil is therefore:

SILT (Si)

If your soil does not fit either description, **return to Step 4** and try again.

Step 7
And as you rub the soil between your fingers (as if you were snapping fingers), it feels:

- Smooth, silky, buttery, and very sticky,
  Go to Step 8

- Soft, even if you observe and feel grains of sand.
  Go to Step 9

Step 8
And the soil is easy to knead, heavily stains your fingers, is sticky, and, when adding water and rubbing between your hands, you feel some grains of sand and on looking carefully you can see sand grains, then you have:

SILT LOAM (SiIl)

If your soil does not fit this description, **return to Step 7** and try again.

Step 9
And the soil is also easy to knead, stains the fingers, is somewhat sticky, and, on adding water to a quantity of soil in the palm of your hand and rubbing it, you see and feel grains of sand, then you have:

LOAM (L)

If your soil does not fit this description, **return to Step 7** and try again.

Step 10
Very carefully try to form small rolls or ribbons between the thumb and index finger and make observations. **REMEMBER** to clean your fingers before starting!

- The soil forms very short ribbons that break very easily and are a little sticky,
  Go to Step 11

- The soil does not form ribbons and is not sticky,
  Go to Step 12

Step 11
And sand grains are also visible. The soil is easy to knead, stains your hands, feels both rough and talcum-powdery, looks opaque, curls when you scratch it with your fingernail, and lumps crumble easily when moist. Your soil is a:

SANDY LOAM (SL)

If your soil does not fit this description, **return to Step 10** and try again.

Step 12
And your soil is also very sandy and soft. It barely stains your hands, is opaque, and, when you add water and rub the soil with your hand, you feel and see many sand grains. The soil wrinkles when you scratch it with your fingernail, and crumbles easily when moist. The soil is therefore a:

LOAMY SAND (LS)

If your soil does not fit this description, **return to Step 10** and try again.

Step 13
And when you knead the soil, you feel some lumps; the soil heavily stains your hands; curls when you scratch it with your fingernail; and, when you leave it to dry, it feels like talcum powder, then you have a:

CLAY LOAM (CL)

If your soil does not fit this description, **return to Step 3** and try again.

Step 14
And if your soil is not lumpy, but stains your hands; is somewhat sticky; curls when you scratch it with your fingernail; and, when moist, lumps crumble easily or with slight resistance, then you have a:

SANDY CLAY (SC)

If your soil does not fit this description, **return to Step 3** and try again.

Step 15
When you knead the soil, and it feels:

- Smooth and talcum powdery,
  Go to Step 16

- Hard, smooth, and very soapy,
  Go to Step 17

Step 16
And your soil also forms resistant and firm circles, heavily stains your hands, is very sticky, has a shiny surface, forms a smooth and shiny surface when you scratch it with your fingernail, and has a buttery consistency on kneading, then you have a:

SILTY CLAY (SiC)

If your soil does not fit this description, **return to Step 15** and try again.

Step 17
And your soil is also hard to knead, easily forms circles, stains your fingers, is sticky, has a very shiny surface, and forms a smooth shiny surface when you scratch it with your fingernail, then you have a:

CLAY (C)

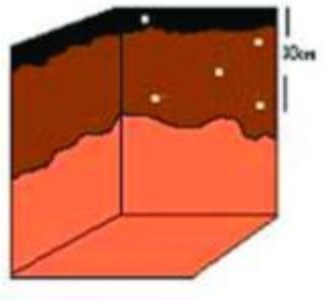
If your soil does not fit this description, **return to Step 15** and try again.

Soil acidity or alkalinity: pH

pH (potential of Hydrogen) ,measures the acidity (1–5), neutrality (5–7), or alkalinity (7 or higher) of soil. It influences the soil's physical, chemical, and biological properties, and hence influences crop growth.

Materials:
Box of indicator paper (e.g., Merck®, graduated from 0 to 14), soup spoon, two disposable glasses, and distilled or bottled water (without gas).

Establish your soil's pH as follows:

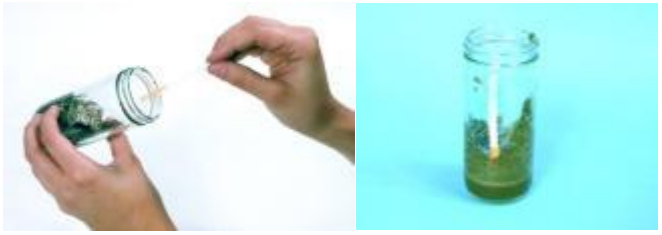


Step 1
Collect several samples from the top 30 cm of the profile and mix in a disposable glass. **Do not touch the soil with dirty or sweaty hands.**

Step 2
Add to one glass, **ONE FLAT SPOONFUL OF SOIL** from the sample previously described.

Step 3
Add **ONE SPOONFUL** of bottled or distilled water. Do not use tap water.

Step 4
Stir for 1 minute, until a homogeneous mixture is formed.



Step 5
Introduce indicator paper into the mixture for 2 minutes or until the paper shows no further color change.

Step 6
If the strip remains very dirty, wash it with a little of the bottled water and do not touch the lower part of the strip with your fingers. Quickly compare the colours of the strip with those of the color chart and note accordingly in *Answer Sheet No. 1*.

Carbonates

The presence of high quantities of carbonates in the soil implies alkaline conditions (e.g., very high pH) and nutritional deficiencies. In dry climates, they can form very hard and dense horizons that prevent root growth, and thus crop growth.

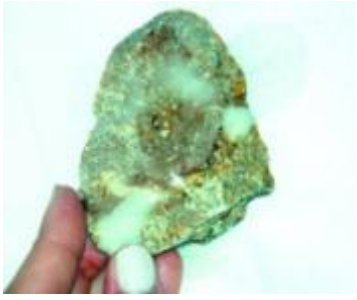
The materials you will need are one dropper (for safety), 10% HCl or muriatic acid (found at any drugstore), and whitish soil lumps.

BE CAREFUL:
THE REAGENT IS DANGEROUS AND MAY CAUSE SKIN BURNS
REMEMBER TO ALWAYS USE IT WITH GREAT CARE
STOPPER IT WELL WHEN YOU FINISH

Follow these steps:

Step 1
If the soil pH is greater than or equal to 7 and the soil profile is from an arid or dry area, search for and identify whitish stains in the profile.

Step 2
With a penknife or machete, scoop out some of the white soil, place to one side, and add some drops of the HCl or muriatic acid.



Step 3
Carefully observe and listen to any effervescence coming from the soil and compare your results with the following table. Note your results in *Answer Sheet No. 1*.

Observation	**Interpretation**
Effervescence is not visible or audible	The soil **does not contain** carbonates

Effervescence is slight, barely visible, but audible	The soil presents **low to very low** contents of carbonates
Effervescence is strong (many bubbles) but **very brief**	The soil presents **moderate levels** of carbonates
Effervescence is strong (many large bubbles) and forms a thick foam that lasts for some time	The soil has a **high** content of carbonates

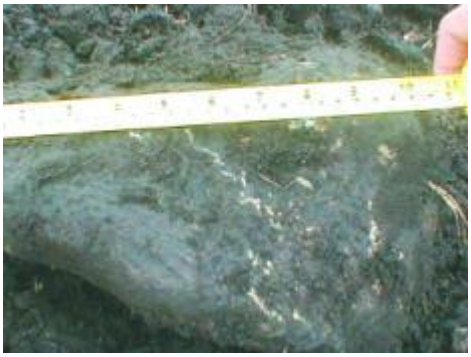
In *Answer Sheet No. 1*, note the depth at which the first carbonates were found in the soil: _____ cm

Pedregosity

Pedregosity (stoniness) refers to the abundance of stones and rocks on the soil surface or in the soil itself. It influences infiltration, evaporation, and availability of water in the soil. It may also prevent plant growth or use of machinery.

If you find pedregosity on your land, check whether you are dealing with stones or rocks, using the attached ruler and determining which predominates:

Step 1
To determine if you are dealing with stones or rocks, merely measure them, using the attached ruler.

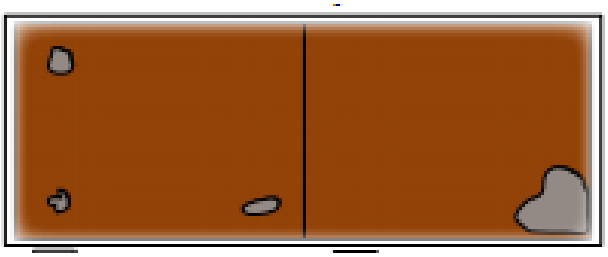


Step 2
If the rocks or stones average less than 8 cm wide, then you have **stones** or **gravel**, but if their widths are more than 8 cm, then you have **rocks.**

Surface stones or rocks

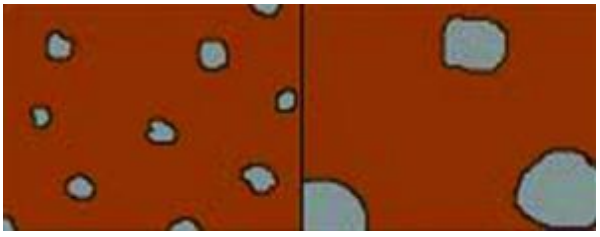
Look at the following drawings to determine the degree of surface pedregosity on your land. Note your observations in *Answer Sheet No. 1*. **Mark the boxes with an "X"**, as according to your situation.

The stones or rocks **do not interfere** with cultivation tasks, or they are not present.



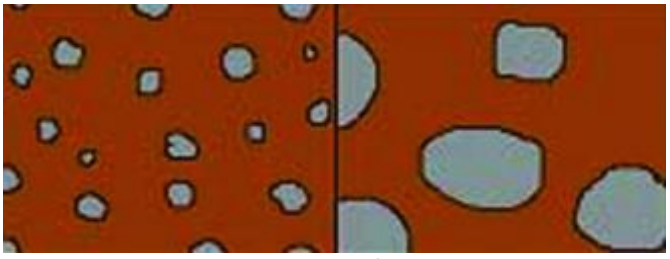
No stone No rocks

The stones or rocks **interfere** with cultivation tasks, but hand tools can be used.



Stony Rocky

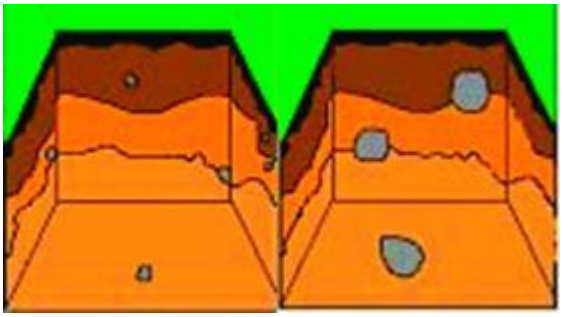
The stones or rocks **do not permit** the use of hand tools or agricultural machinery.



Very stony Very rocky

Stones or rocks in the profile

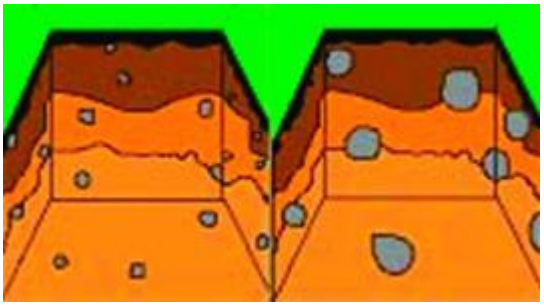
Mark the boxes with an "X", as according to your situation.



No stones **No rocks**

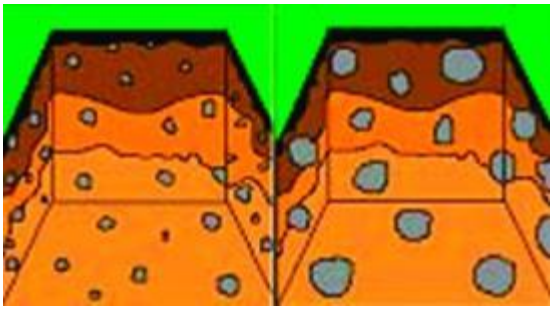
The stones or rocks within the profile **do not interfere** with plant growth or with cultivation, or are absent.

The stones or rocks within the profile **hamper** plant growth and cultivation tasks.



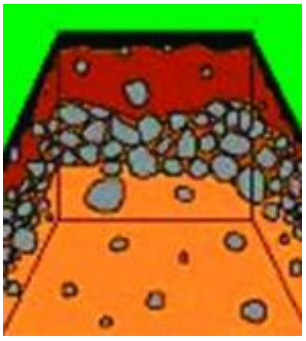
Stony **Rocky**

The stones or rocks within the profile **prevent** plant growth and use of hand tools or machinery.



Very stony **Very rocky**

Depth at which the first rocks or stones are found: _____ cm



Did you find a stony or rocky layer in the profile?

Yes **No**

Depth: _____ cm Thickness: ____ cm

Note your observations in *Answer Sheet No. 1*.

Hardpan layers

Hardpan layers are hard impermeable layers that can prevent root growth, water movement, and soil aeration.

You will need a penknife or knife and a tape measure. Carry out the following steps:

Step 1

Place the tape measure on one face of the soil profile hole and, using the penknife or knife, strike the face with strong sharp blows at different points throughout the profile.

Step 2
Use your thumb to mark the depth to which the knife had penetrated the soil and then pull the knife out without moving your thumb.

Step 3
Measure how many centimetres the knife had penetrated the soil. If the distance was less than 3 cm, then you have a compacted layer. **Note its depth and thickness.**

If you find several hard layers, note the depth and thickness of each one.

Soil mottling

Mottling comprises spots or stains of colours—yellow, red, blue, green, or gray—mixed with the color of the horizon in small or large quantities. They indicate poor drainage and a lack of oxygen for roots.

The height of the *water table* is the depth measured from the soil surface at which groundwater is found. The height varies according to conditions such as amount of precipitation and soil drainage. When the water table is close to the surface, then swampy or waterlogged areas are formed.

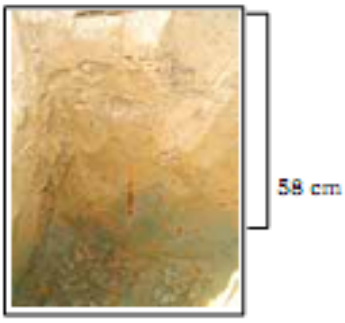
Types of soil mottles:

You will need a tape measure. Observe the profile and answer the following:

Do you see soil mottling, collared brown, red, blue, gray, green, or yellow, in the profile?

Yes **No**

If you answer **Yes**, then measure the depth, between the surface and mottling, as shown in the following photos:



58 cm 25 cm

If you **did not find** mottling in the profile, answer the following questions and verify if the profile could present mottling at 70 cm or deeper. (Use the *Procedure Sheet*.)

- Are there rivers or streams **very close** to the evaluation site?

  Yes **No** **Don't know**

- If you make a deep hole at any time of the year, does water come up?

  Yes **No** **Don't know**

- Are wells or aquifers found close to the plot?

  Yes **No** **Don't know**

  At what depths? _____ m

If you answer **Yes** to some or most of the above questions, the soil may present mottling at 70 cm or deeper because of high water tables. (Write down your response).

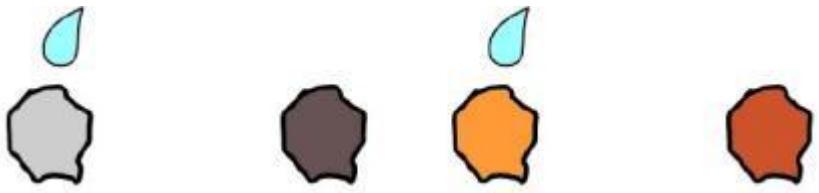
Yes **No**

If you need to know the depth of the water table with greater accuracy to prevent it from interfering with crop growth, prepare a much deeper soil profile hole to 1.5 m.

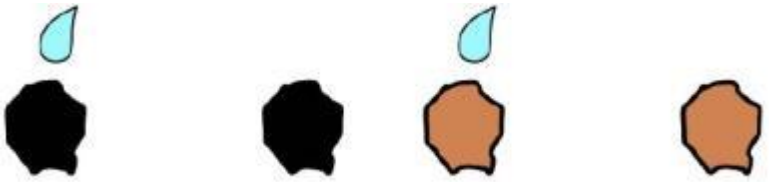
Soil resistance

Soil resistance refers to the force needed to break a lump of soil. This force varies according to moisture content, texture, organic matter content, and soil structure. To test for resistance, follow the procedure below for each layer or horizon of the profile.

Step 1
To determine soil moisture, pick out a lump of soil and add a drop of water:



- If the soil changes color, it is **dry**.



- If the soil does **not** change color and does not wet the hand on picking it up, then it is **moist**.

- If the soil wets the hand on picking it up, then it is **wet**. In this case, let the soil dry until it is either moist or dry and continue with the evaluation.

Step 2
Go to the table on the next page and determine the soil's resistance to breaking. Note your observations in *Answer Sheet No. 1*.

<table>
<tr><td></td><td colspan="6">Resistance to breaking</td></tr>
<tr><td rowspan="3">Dry soil</td><td colspan="2">Soft</td><td colspan="2">Hard</td><td colspan="2">Extremely hard</td></tr>
<tr><td colspan="6">On breaking a soil lump, large fragments remain that, when pressed together, cannot be joined again.</td></tr>
<tr><td>The soil is loose and does not form lumps. When pressed between the thumb and index finger, it breaks very easily into powder or loose grains.</td><td>  </td><td>The soil crumbles between the thumb and index finger only under considerable pressure, or both hands are needed.</td><td>  </td><td>The soil cannot be crushed or broken between the fingers or with both hands, but does break when trodden underfoot or hit with a stone, or hammer.</td><td>  </td></tr>
<tr><td colspan="7"></td></tr>
<tr><td rowspan="3">Moist soil</td><td colspan="2">Friable</td><td colspan="2">Firm</td><td colspan="2">Extremely firm</td></tr>
<tr><td colspan="6">On breaking the lump, the particles join again when pressed together.</td></tr>
<tr><td>The soil is loose or crumbles easily between thumb and index finger.</td><td>  </td><td>The soil crumbles between thumb and index finger, using moderate pressure. Clear resistance is noted.</td><td>  </td><td>The soil crumbles only under very strong pressure and must be broken piece by piece.</td><td>  </td></tr>
<tr><td colspan="7"></td></tr>
<tr><td rowspan="3">Plastic (moist)</td><td colspan="3">Plastic</td><td colspan="3">Highly plastic</td></tr>
<tr><td colspan="6">When you use pressure, the lump does not separate but becomes deformed; it can also be MOLDED.</td></tr>
<tr><td colspan="3">You need moderate pressure to deform a lump of soil.</td><td colspan="3">You need considerable pressure to deform a lump of soil.</td></tr>
</table>

Soil structure

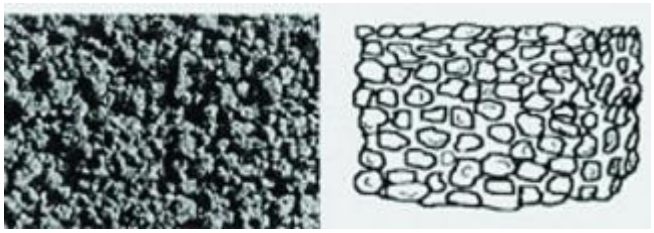
Soil structure is the organization of soil lumps or aggregates (themselves formed of soil minerals and organic matter) and the network of pores between them.

A poor soil structure may negatively affect plants through factors such as excess or deficiency of water, lack of aeration, little microbial activity, prevention of root growth, incidence of diseases, and bad drainage.

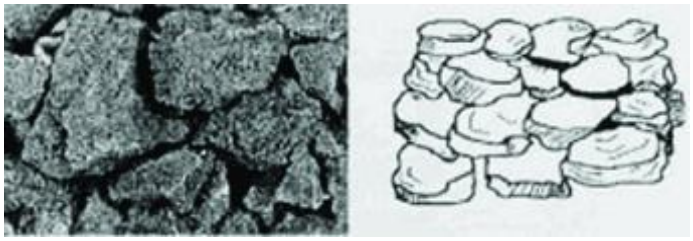
To clearly see the soil structure, the soil profile should be allowed to dry in the sun until natural cracks appear in the soil.

If the soil is moist and its structure is evident, compare it with the following photos to identify it and write down your results in *Answer Sheet No. 1*. If you observe various different structures, note the most prominent.

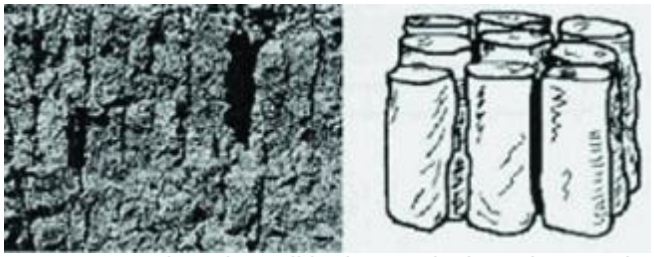
With structure



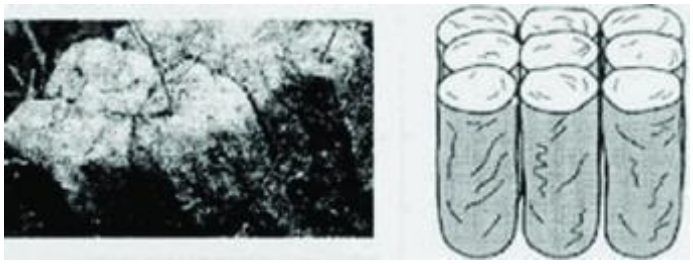
Granular: Often found on the surface as small round grains that, when scooped up, are loose.



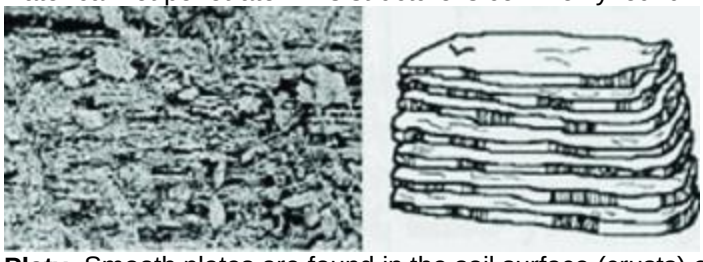
Blocky: Irregular blocks that may have rounded or straight edges.



Prismatic: When the soil is dry, vertical cracks are observed. They are usually found in deeper horizons or layers.



Columnar: The soil forms a compact or hard mass that breaks into columns with rounded edges which water cannot penetrate. This structure is commonly found in deeper layers of sodic soils.



Platy: Smooth plates are found in the soil surface (crusts) or in the profile as overlapping plates of soil.

Photos and drawings from *Soil Characterization Protocol Field Guide* of the GLOBE Program

No structure

Single grained: The soil does not form lumps but feels loose and dusty, like sand.

Massive: The soil does not have a visible structure: no cracks are seen and the soil appears as a solid shapeless mass that is very difficult to break.

You have come to the end of the **first** part of the *Guide*. All the descriptions you made should be noted in *Answer Sheet No. 1* (see *Answer Booklet*).

Use *Answer Sheet No. 1* to respond to the following 14 field observations:

Can you see erosion in the soil?

Yes **No**

Can you see mold or greenish layers on the soil surface?

Yes **No**

Can you see hard or crunchy crusts on the soil surface?

Very marked **Not strongly marked** **Absent**

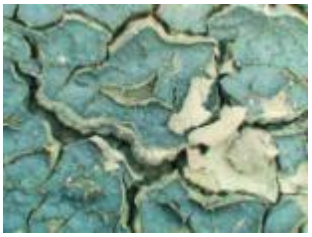
The sampling site is exposed to the sun in the:

Morning and afternoon **Morning** **Afternoon**

Do you see white or peeling crusts?


Very marked **Not strongly marked** **Absent**

Do you see black crusts on the soil surface?



Very marked **Not strongly marked** **Absent**

Are you in a dry or arid region where it rarely rains, or in a humid region near the sea or salt lakes?

Yes **No**

Do you see live roots in the profile?

Yes **No**

Depth of root growth: _____ cm

Do you observe small dry plants, or low production in the crop?

Plants mildly affected **Plants strongly affected** **Normal plants** **No crop**

Do you see a lot of fallen leaves or decaying organic matter on the soil surface?

Yes **No**

Is the soil is very black, very soft, spongy, and, when you walk on it, you sink?

Yes **No**

When you introduce a knife into the first horizon, does it enter easily?

Yes **No**

Are you near rivers, streams, seas, lakes, or wells that maintain groundwater levels close to the surface?

Yes **No**

The soil's plant cover such as grasses, weeds or mosses is:

Very good (abundant) **Good (normal)** **Regular (covers almost half of the plot)**
Spaced out (patches) **No cover**

You have come to the end of the **second** part of the *Guide*. All the descriptions you made should be noted in *Answer Sheet No. 1*

Answer Sheet No. 1:

Hoja de Respuestas No. 1

Características y Observaciones							
1	Pendiente						
2	Terreno circundante						
	Posición del perfil						
3	Capas u Horizontes	Espesor	Color seco	Color húmedo	Textura	Resistencia al Rompimimiento	
4	pH						
5	Carbonatos	No tiene	Bajos a Muy Bajos	Medios	Altos	Profundidad (cm):	
6	Pedregosidad superficial	Sin Piedras	Sin Rocas	Pedregoso	Rocoso	Muy Pedregoso	Muy Rocoso
	Pedregosidad en el perfil	Sin Piedras	Sin Rocas	Pedregoso	Rocoso	Muy Pedregoso	Muy Rocoso
	Horizonte pedregoso o rocoso	Si	No	Profundidad (cm):	Espesor (cm):		
	Profundidad de primeras rocas o piedras (cm)						
7	Capas endurecidas	Si	No	Profundidad (cm):	Espesor (cm):		
8	Moteados	Si	No	Profundidad (cm):			
	Moteados a mas de 70 cm	Si	No				
9	Estructura						

10	Si	No		18	Poco afectadas	Muy afectadas	Plantas normales	No hay cultivo	
11	Si	No		19	Si	No			
12	Muy marcadas	Poco marcadas	No hay	20	Si	No			
13	La mañana y la tarde	La mañana	La tarde	21	Si	No			
14	Muy marcadas	Poco marcadas	No hay	22	Si	No			
15	Muy marcadas	Poco marcadas	No hay	23	Muy Bueno	Bueno	Regular	Espaciado	Sin cobertura
16	Si	No							
17	Si	No	Profundidad (cm):						

Answer Sheet No. 2:

Hoja de Respuestas No. 2

Propiedades Inferidas					
1	**Drenaje Interno**	Lento o Muy lento	Bueno	Excesivo	
	Drenaje Externo	Ninguno	Lento	Moderado	Excesivo
2	**Materia Orgánica**				
	Horizonte	Muy Alta	Alta	Media	Baja
	Suelo Orgánico	Si	No		
3	**Salinidad**	Muy Alta	Muy Alta	Suelo Normal	
	Sodicidad	Muy Alta	Muy Alta	Suelo Normal	
4	**Profundidad Efectiva (cm)** _____				

Lote Número: _____

Cajuela Número: _____

Nombre: _____
(Persona que describe) _____

Uso actual del lote: _____

Dibuje un pequeño mapa de su lote y ubique la cajuela.

Part 2: Inferred soil properties

With ALL the answers for Parts 1 written down in *Answer Sheet No. 1*, you can now infer various important soil properties including:

- Effective depth
- Organic matter
- Drainage
- Salinity and sodicity

Follow the steps described below. To help with your determinations, use the *Procedures Sheet* attached to the *Guide*.

Note: If you have followed the guide carefully and recorded all the observations, the following steps can be made in your house or office.

Effective depth

Effective depth is the depth to which plant roots can reach in a soil without meeting obstacles (whether physical or chemical) such as the water table, hardpan, loose sands, impermeable clays, or presence of salts.

Effective depth is one of the most important properties to take into account when deciding which crop to plant, as optimal root growth and good crop development depend on effective depth. To determine this depth:

Step 1

Go to the following table and write down the depth at which different obstacles were found (use the *Procedures Sheet*).

Step 2
If you do not find any obstacle in your soil profile, draw a line through the corresponding blank space (use *Answer Sheet No. 1*).

If you found:	**Depth**
Hardpan layers in the profile; note their depth (Item 7 in *Answer Sheet No. 1*) (If you find several, note the first one)	
Mottling in the profile; note its depth (Item 8 in *Answer Sheet No. 1*)	
High carbonate contents in the profile; note the depth at which they appear (Item 5 in *Answer Sheet No. 1*)	
A sandy horizon; note its depth (Item 3 in *Answer Sheet No. 1*)	
Stony or rocky layers; note their depths (Item 6 in *Answer Sheet No. 1*)	
Very stony or rocky **profile**; note the depth at which rocks or stones first appear (Item 6 in *Answer Sheet No. 1*)	
Very stony or rocky soil **surface**; note a minimum depth of 0 cm	
Massive or platy structures; note a minimum depth of 0 cm	

Step 3
To determine effective depth, make the following calculations and write your results in *Answer Sheet No. 2*:

- If **none** of the options listed in the table correspond to your soil profile, the effective depth will be the same as or greater than the depth of the profile hole (note your results in *Answer Sheet No. 2*).

- If some or all of the obstacles listed in the table appear, choose the shallowest depth (e.g., nearest to the soil surface or minimum depth):

Minimum depth: _____ cm (depth 1)
Root depth (Item 17 in *Answer Sheet No. 1*): _____ cm (depth 2)

Compare these two results as shown below and note the **effective depth** in *Answer Sheet No. 2*.

- If depth 1 **is greater** than depth 2, then the effective depth will be equal to depth 1.

- If depth 1 **is less** than depth 2, then the effective depth will be equal to depth 2.

- If **no roots are present**, then the effective depth is equal to the minimum depth.

Organic matter

Organic matter is a significant component of soil because it influences the soil's chemical, physical, and biological properties. It improves soil structure and porosity, moisture retention, microbial activity, and soil fertility, among other attributes. To characterise organic matter:

- Go to Step 1, read the question, and select the option that best describes your soil's condition (use *Answer Sheet No. 1*).

- Each option will lead you to another step until the level of organic matter corresponds to your soil (note your results in *Answer Sheet No. 2*).

Step 1
See *Answer Sheet No. 1*, Items 3, 8, and 20. On describing the **horizon**, you found:

- Your soil is very dark, very soft, loose, spongy, and sinks underfoot,
  Go to Step 2

- Textures are light (SL, LS, or S), soil colours are dark (identified in color chart by numbers 1, 2, 3, 4, 5, 6, 14, 15, 16, 17, 18, 19, 20, 22, 24, 28, 31, and 32), and **no** mottling is present,
  Go to Step 3

- Textures are light (SL, LS, or S), soils are of any color, and **some** mottling is present,
  Go to Step 4

- Textures are intermediate (L, SC, Si, or SiL), soils are dark (color numbers 1, 2, 3, 4, 5, 14, 15, 16, 17, 18, 19, 20, 22, 24, 28, 31, and 32), and **no** mottling is present,
  Go to Step 5

- Textures are heavy (CL, C, or SiC), soils are dark (color numbers 1, 2, 3, 4, 5, 14, 15, 16, 17, 18, 19, 20, 22, 24, 28, 31, 32, and 43), and **no** mottling is present,
  Go to Step 6

- Textures are either intermediate or heavy (L, SC, Si, SiL, CL, C, or SiC), soils are of any color, and **some** mottling is present,
  Go to Step 7

Step 2
With *Answer Sheet No. 1* in your hand, answer the following questions. Use the *Procedures Sheet*.

Questions (select your answer, writing "X" in the box)	**Yes**	**No**

Did you see a lot of fallen leaves or decaying organic matter? (Item 19)		
When you introduced the knife into the first horizon, did it enter easily? (Item 21)		
Did you find an extremely acid pH? (Item 4)		
Is your soil's internal drainage either slow or very slow?		

- If you responded "**No**" to any of the four questions, **return to Step 1** and try again.

- If you answered "**Yes**" to **ALL** four questions, then your soil is **ORGANIC.** You can determine the level of organic matter content in your soil by its resistance to breakage (Item 3 of *Answer Sheet No. 1*), structure, and drainage. **Go to Step 3.**

Step 3
The soil feels:

- **Friable** when humid, **soft** when dry, and also very spongy and porous,
  Organic matter content is therefore **high**

- **Firm or friable** when humid and **hard or loose** when dry,
  Organic matter content is therefore **intermediate**

If your soil does not fit either description, **return to Step 1** and try again.

Step 4
See Item 3 of *Answer Sheet No. 1*. Observe the color of your soil and its resistance to breakage:

- Soils with **no mottling**; colours are dark coffee, brown, or yellowish coffee (color numbers 16, 17, 18, 19, 20, 22, 24, 25, 26, 27, 28, 33, 34, 35, and 36). On crumbling, soil feels **friable** when humid or **soft** when dry; and also porous and spongy,
  Organic matter content is therefore **intermediate**

- Soils **may have mottling**; usually dark coffee, yellowish coffee, brown, and sometimes black in color (color numbers 1, 2, 3, 4, 5, 6, 7, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 32, 35, and 36). On crumbling, the soils feel **friable** or **firm** when humid,
  Organic matter content is therefore **low**

- The soil **may have mottling**; usually gray, red, yellow, and orange in colours (color numbers 8, 9, 10, 11, 12, 13, 18, 21, 23, 27, 29, 30, and 34 to 54). The soil feels **loose** when dry, that is, very sandy and rough,
  Organic matter content is therefore **low**

If your soil does not fit any of these descriptions, **return to Step 1** and try again.

Step 5
See Item 3 of *Answer Sheet No. 1*. Soil feels:

- **Friable** when humid; **soft, hard, or extremely hard** when dry; and also spongy and porous,
  Organic matter content is therefore **very high**

- **Firm or plastic** when humid and **hard and extremely hard** when dry,
  Organic matter content is therefore **intermediate**

If your soil does not fit either description, **return to Step 1** and try again.

Step 6

See Item 3 of *Answer Sheet No. 1*. Soil feels:

- **Friable** when humid; **soft, hard, or extremely hard** when dry; and also spongy and porous,
  Organic matter content is therefore **low**

- **Firm** or **plastic** when humid and **hard** or **extremely hard** when dry,
  Organic matter content is therefore **intermediate**

If your soil does not fit either description, **return to Step 1** and try again.

Step 7
See Item 3 of *Answer Sheet No. 1*. Define the color and describe the horizon's or layer's resistance to breakage:

- Soil with **no mottling** and colored dark coffee, brown, or yellowish coffee (color numbers 15, 16, 17, 18, 19, 20, 22, 24, 25, 26, 27, 28, 33, 34, 35). On crumbling, the soil feels **friable**, **firm**, or **plastic** when humid; or **soft, hard, and extremely hard** when dry,
  Organic matter content is therefore **intermediate**

- Soil **may present mottling**, and be of any color and shade from pale to dark. On crumbling lumps, the soil feels **friable, firm,** or **plastic** when humid and **hard or extremely hard** when dry,
  Organic matter content is therefore **low**

If your soil does not fit either description, **return to Step 1** and try again.

Drainage

Drainage is the capacity the soil has to get rid of water by surface runoff and infiltration.

Drainage can affect the growth and development of most crops, modifying factors such as effective depth, soil structure, microbial activity (both good and bad), availability of oxygen and plant nutrients, soil pH, concentration and solubility of certain elements, and decomposition of organic matter. To characterise the soil's internal drainage:

- Go to Step 1 of *Key 1* below and select the option that most fits your soil's condition (use *Answer Sheet No. 1*).

- Each option leads to another step, until the class of drainage of your soil appears.

- Note your answer in *Answer Sheet No. 2*.

Follow the same procedure to describe the soil's external drainage (*Key 2*).

Key 1: Internal drainage

Refers to *infiltration* or the passage of water through soil.

Step 1
Do you find mottling in the profile?

At what depth?

(Item 8 in *Answer Sheet No. 1*).

- Mottling appears at a depth of less than 50 cm,
  Go to Step 2

- **No** mottling appears, or it appears at a depth of more than 50 cm,
  Go to Step 3

Step 2
Carefully read the descriptions presented below and select the option that **MOST** resembles the condition of your soil:

- It is found near rivers, streams, seas, lakes, or wells, or at the bottom of slope that maintains groundwater levels close to the surface and prevents the passage of water through the profile.
  Drainage is therefore **slow to very slow**

- Organic soils or soils with limitations **close to the surface** or **on the surface**. Obstacles appear in the profile such as clayey horizons with or without pedregosity, hardpan or impermeable layers, high sodicity, and massive and nonporous structures that block the passage of water through the profile.
  Drainage is therefore **slow to very slow**

- If **mottling is at a depth of less than 50 cm,** but your soil does not fit either description, then it may have been modified and its internal drainage has improved. If this is the case, **go to Step 3**.

Step 3
Carefully read the descriptions presented below and select the option that **MOST** resembles the condition of your soil:

- The soil may or may not present mottling deeper than 50 cm. The first horizons usually have loamy textures (L, LC, SiL, or CS), and limitations may be **deep** within the profile. Soil structure is good (neither loose nor massive) and porous:
  Drainage is therefore **good**

- The soil does **not** present mottling. It is usually sandy in texture (LS, SL, and S), loose in structure; it may or may not present stones of different grades, has good porosity, and does not present limitations:
  Drainage is therefore **excessive**

If the soil does not fit either description then **return to Step 1** and try again.

Key 2: External drainage

Refers to *surface runoff*.

Step 1
How steep is the land? (Item 1 of *Answer Sheet No. 1*)

- 0% to 2% ..... **Go to Step 2**
- 2.1% to 6% ..... **Go to Step 3**
- 6.1% to 13%..... **Go to Step 4**
- More than 13%..... **Go to Step 5**

Step 2
The water only moves through the soil profile. The water either forms puddles on the surface or infiltrates rapidly through the soil:

Drainage is **absent**

Step 3
The soil usually has good plant cover, comprising grasses, weeds, r mosses. Its internal drainage is good, and the soil surface is not eroded:

Drainage is therefore **slow**

Step 4
- The soil's internal drainage is excessive. The soil surface is not eroded:
  Drainage is therefore **slow**

- The soil's internal drainage is good. Plant cover is good, comprising grasses, weeds, lawns, or mosses. The soil surface is not eroded:
  Drainage is therefore **good**

- The soil's internal drainage is good. Plant cover, comprising grasses, weeds, lawns, or mosses is spaced out, regular, or absent. The soil surface may be eroded:
  Drainage is therefore **moderate**

- The soil's internal drainage is slow to very slow. Plant cover is good or absent. The soil surface may be eroded:
  Drainage is therefore **moderate**

Step 5

- The soil surface is not eroded. Internal drainage is good or excessive. Plant cover is good, comprising grasses, weeds, lawns, or mosses:
  Drainage is therefore **moderate**

- The soil is usually eroded. Plant cover is spaced out. Internal drainage is slow or very slow. Stones or rocks may be present on the surface:
  Drainage is therefore **excessive**

Salinity and sodicity

Salinity and *sodicity* refer to the soil's salt concentration levels, which may limit plant growth. If the soil has high concentrations of salts, it is *saline*. If sodium salts predominate, then the soil is *sodic*.

Excess sodium salts cause large changes in various physical, chemical, and biological properties of the soil, including organic matter; drainage; pH; structure; and availability of nutrients, water, and oxygen.

Salinity

Use *Answer Sheet No. 1* and *Procedures Sheet*. **If your soil and terrain present following conditions**:

- The profile is located in flat or flat and mildly undulating land, where slopes range between 0% and 2% (Items 1 and 2)

- White or peeling crusts can be seen (Item 14)

- The pH is between 7 and 8 (Item 4)

- The location is in a dry or arid region where rain rarely falls, or in a humid region close to the sea or salt lakes (Item 16) …

... and these conditions coincide with two or more of the following conditions:

- The soil has hard or crunchy crusts (Item 12)
- Crop plants are small and parched, or their production is low (Item 18)
- Internal drainage is good or excessive (*Answer Sheet No. 2*, Item 4)
- The soil is extremely hard or firm,

… then your soil is saline.

To evaluate the level of salinity, check effects on susceptible plants cultivated in the area. Compare the following descriptions with your plot's situation and identify that which most resembles your soil's condition.

- **Salinity is low** when crusts and cracks are not readily observed, and effects on developing susceptible plants are **mild**.

- **Salinity is high** when crusts and cracks are highly noticeable and plants are either **severely** affected or do not grow.

Sodicity

Use *Answer Sheet No. 1* and *Procedures Sheet*. **If your soil and terrain present the following conditions**:

- Black crusts are found in the soil (Item 15)
- Internal drainage is slow to very slow (Item 1 of *Answer Sheet No. 2*)
- The pH is more than 8 (Item 4) …

... and these conditions coincide with two or more of the following conditions:

- The soil is soft or friable
- A very hard crust is present on the surface and the top 20–30 cm of soil has a muddy, moist, and very smooth consistency
- Soil structure is prismatic or columnar,

… then your soil is sodic.

To discover the level of sodicity in your soil, check the effects on susceptible plants grown in the area. Select the description that most resembles your soil's condition:

- **Sodicity is low** when black crusts are not readily evident and effects on developing susceptible plants are **mild**.

- **Sodicity is high** when the black crusts are highly noticeable and plants are **severely** affected or do not grow.

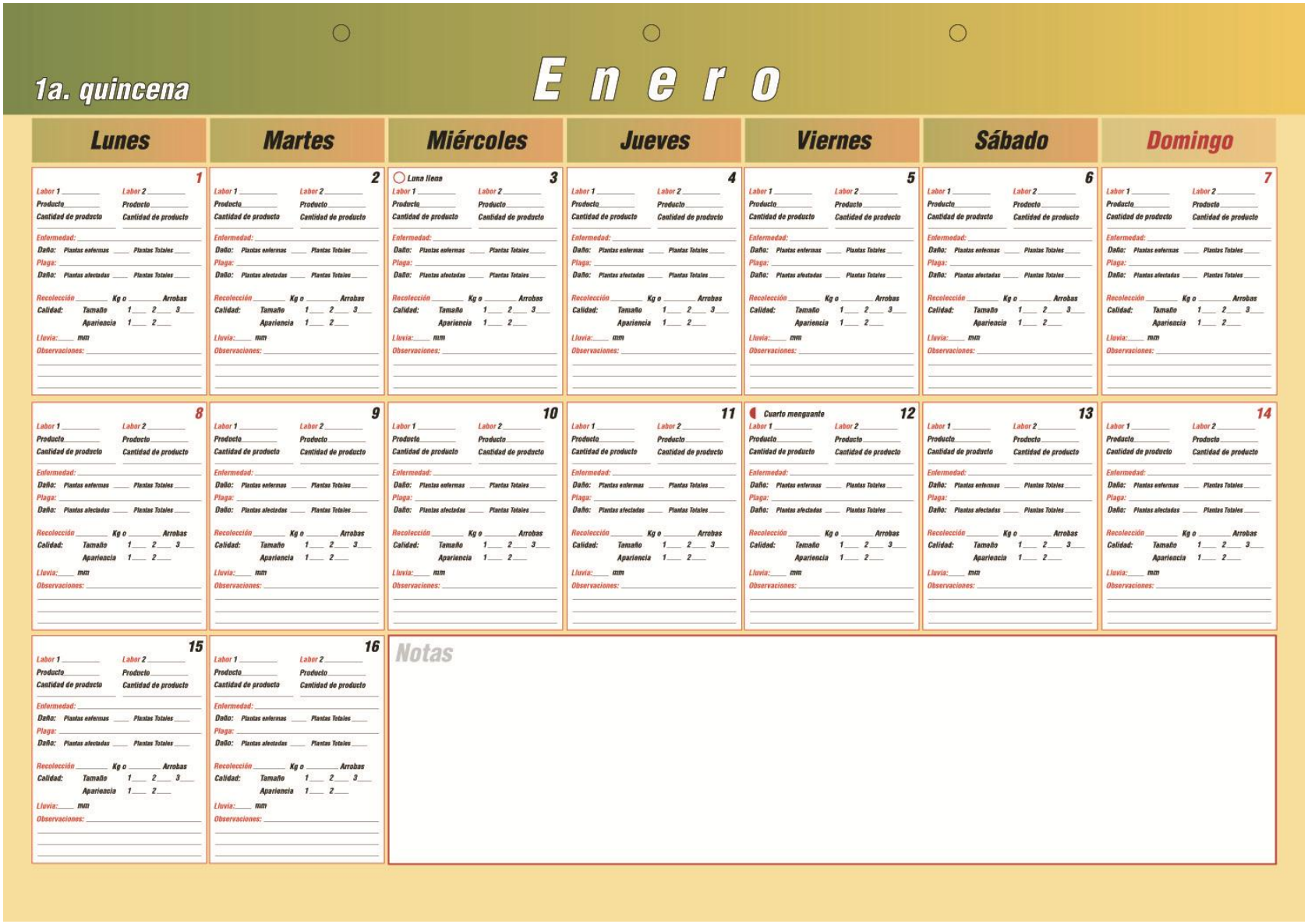
You have come to the end of the *Guide*. All the descriptions you made should be noted in *Answer Sheet No. 1 and 2*

APPENDIX A2.
Guide form based on a calendar used by the farmers to record information on the production of each plot planted to Andean blackberries and lulo.

Guide form for Andean blackberry developed to capture information describing each production site (producer, identification, location, planting date and spacing, number of plants), and variety (thorned or thornless)

Calendar for Andean blackberry developed to capture harvest events. Data was actual weights of fruit harvested per plant each week

Guide form for lulo developed to capture information describing each production site (producer, location, identification, planting date and spacing, number of plants), and variety (thorned or thornless)



Calendar of lulo developed to capture harvest events. Data was actual weights of fruit harvested per plant each week

APPENDIX A3.

MATLAB scripts language, used to train a Kohonen map, cluster prototypes, visualize dependencies, and visualize the bidimensional map and dependencies between clusters. In red, the parameters that should be modified. Expressions between % and bold are expressions used to describe each process.

To train the Kohonen map.

```
sD = som_read_data ('filename.data')

sD = som_normalize(sD,'var')

som_gui(sD)


%%In the initialization training window%%%:

initialize/train/Load/save/Save Map/Save in workspace/Save map as = sM /close
```

To visualize the bidimensional map

```
som_show(sM,'umat','all');
```

To cluster vector prototypes through the K-means algorithm and the Davies-Bouldin index

```
%%% To create a file with the BMUS asociated to the input vectors%%%%%

 Bmus = som_bmus(sM,sD);

%%% Sort BMUs %%%
%%% Bmus = sort(Bmus);
%%% SDd = data denormalized %%%
 sDdn = som_denormalize(sD);
 sDda = sDdn.data;
 sBMUs = [Bmus,sDda]


%%% Clustering with k-means with different values for k. %%%
%%% The Davies-Bouldin index is calculated for each clustering. %%%

figure(7);
 subplot(1,2,1)
 [c,p,err,ind] = kmeans_clusters(sM,number of k values to test);
 plot(1:length(ind),ind,'x-');

%%% choosing the clustering with the minimal Davies-Bouldin index %%%
 [dummy,i] = min(ind);

%%% cl = "cluster number" %%%
```

```
 cl = p{i};
 %cl = p{40};
 figure(7);
 subplot(1,2,2);

 som_cplane(sM,cl);
 ClNum = i;


figure(8)
 som_show(sM,'color',cl);
 som_show_add('label',sM,'Textsize',8,'subplot',1);


figure(9)
 som_cplane(sM,cl);
 hold on
 som_grid(sM,'Label',cellstr(int2str(cl)),...
'Line','none','Marker','none','Labelcolor','k');


%%% To assing a cluster number to each input vector %%%
 for i=1:size(Bmus), ;
 clBM(i) = [cl(Bmus(i),:)] ;
 end ;
 clBMU = clBM'

%%% To save the clusters%%%

cc = clBM';
 save cluster_number.data cc -ASCII -tabs
```

To visualize the dependencies between the clusters shown in the Kohonen map by a "component plane" representation

%%%To show as much component planes as dimensions or variables are available to visualize%%%%

som_show(sM,'umat','all','comp',1: ***the number in the dataset of the last variable or column to see as a component plane*** ,'empty','Labels','norm','d')

%%% To show only one component plane%%%

som_show(sM,'umat','all','comp',***the number of the component plane to visualize*** ,'empty','Labels','norm','d')

CURRICULUM VITAE

DANIEL RICARDO JIMÉNEZ RODAS

Date of Birth: 23.03.1977
Place of Birth: Manizales, Colombia
Nationality: Colombian
Marital Status: Married (Spouse: Maria del Pilar Acosta)

Professional Address
International Centre for Tropical Agriculture (CIAT), Decision and Policy Analysis (DAPA). Recta Cali Palmira km 18, A.A. 6713. Cali, Colombia. E-mail : d.jimenez@cgiar.org

Residential Address:
Cll 14 Oeste # 24D 70. Edificio Montelugano. 601A. Cali. Colombia. Tel: ++57 3218000486. Email: danieljimenez.rodas@gmail.com

UNIVERSITY STUDIES

1996-2002: Agronomist. Universidad de Caldas, Manizales, Colombia.
Thesis: Characterization and study of the genetic diversity of the *Vasconcellea* and *Carica* (Caricaceae) genera in Colombia and Ecuador by means of isoenymatic markers.

2005 – to present
Ph.D. candidate in Applied Biological Sciences (Agronomy)
Thesis at the Faculty of BioScience Engineering: Agricultural Science, Ghent University.

PROFICIENCY AND KNOWLEDGE

Language skills:

	Understanding				**Speaking**		**Writing**	
	Listening		Reading					
Spanish	Proficient	C2	Proficient	C2	Proficient	C2	Proficient	C2
English	Proficient	C2	Proficient	C2	Proficient	C2	Proficient	C1
French	Proficient	C2	Proficient	C2	Proficient	C2	Proficient	C1

*According to Standard Levels (range A1-C2) of Common European Framework of Reference (CEFR) (*http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp

Software Proficiency:

Operating systems: MS Windows XP, MS Windows 7.

Statistical software: STATISITICA, MATLAB, Winstat, XLSTAT, Java SNNS, FENNIX, R.

Geographical information system (GIS) software*:* ARC-GIS, ARC-VIEW, DIVA, Floramap, Homologue, Maxent.

Non-Specific Software: MS Office Professional, Adobe Photoshop, Endnote.

PROFESSIONAL EXPERIENCE

- **2010 to present**

Research Fellow. International Center for Tropical Agriculture (CIAT). In charge of the technical coordination of the Site-Specific Agriculture Based on Farmers Experiences project (SSAFE) for fruits in Colombia. http://www.frutisitio.org/.

- **2005 – 2008**

Researcher at the University of Applied Sciences of Western Switzerland - (HEIG-VD). Project: "Precision agriculture and the construction of crop-field models for tropical fruit species". Responsible for the implementation of data-driven models and in a site-specific program for tropical fruit species in Colombia. Also in charge of coordinating data collection, compilation of data in centralized databases, and data analysis and interpretation..

- **2003 – 2005**

Researcher at Bioversity International. Responsible for the New World Fruit Database http://www.bioversityinternational.org/databases/new_world_fruits_database/search.html. Compiling information on new world fruit trees, and mapping fruit species through Geographic Information System (GIS) to visualize all locations where a species has been found, and more importantly to allow defining potential zones where fruit trees species can be grown.

- **2001 – 2003**

Researcher at Bioversity International. Project: "Use of genetic resources of papayas for breeding and promotion". Responsible for the isoenzymatic and morphologic characterization of genus *Vasconcellea* and *Carica* in Colombia, Costa Rica and Ecuador.

AWARDS:

One of the 40 researches among 400 from all over the world selected to participate in the first biennial seminar for young researchers working in developed and developing countries. France, (2010).

Scholarship holder - *State Secretariat for Education and Research* (SER) Switzerland, (2005 – 2008).

PUBLICATIONS

A. Publications in International Journals with Referee System

Jiménez, D., Cock, J., Jarvis, A., Garcia, J., Satizábal, H.F., Van Damme, P., Pérez-Uribe, A. and Barreto-Sanz, M., (2011). Interpretation of Commercial Production Information: A case study of lulo (*Solanum quitoense*), an under-researched Andean fruit. Agricultural Systems. 104 (3): 258-270. (ISI impact factor: 2.899).

Jiménez, D., Cock, J., Satizábal, F., Barreto, M., Pérez-Uribe, A., Jarvis, A. and 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 (2): 198–208. (ISI impact factor: 1.846).

Scheldeman, X., Willemen, L., Coppens d'Eeckenbrugge, G., Romeijn-Peeters, L., Restrepo, M.T., Romero Motoche, J., **Jiménez, D.**, Lobo, M., Medina, C.I., Reyes, C., Rodriguez, D., Ocampo, J.A., Van Damme, P. and Goetgebeur, P. (2007). Distribution, Diversity and Environmental Adaptation of Highland Papayas (*Vasconcellea* spp.) in tropical and subtropical America. Biodiversity and conservation. 16 (6): 1867-1884. (ISI impact factor: 1.401).

B. Publications in International Journals without Referee System

Restrepo, M.T., **Jiménez, D.,** Coppens d'Eeckenbrugge, G., Vega, J. (2004). Morphological diversity of cultivated mountain papayas (*Vasconcellea* spp.) in Ecuador. Proceedings of the Interamerican Society for Tropical Horticulture, 48: 119-123.

Caetano, C.M., Coppens d'Eeckenbrugge, G., Olaya, C.A. **Jiménez, D**., Vega, J. (2003). Spindle absence in *Vasconcellea cundinamarcensis* (Caricaceae). The Nucleus, 46 (3): 86-89

C. Book Chapters

Satizábal, H., Barreto-Sanz, M., **Jiménez, D.,** Pérez-Uribe, A., Cock, J., Bolay, J.-C., Schmid, M., Tejada, G. and Hazboun, E., 2012. Enhancing Decision-Making Processes of Small Farmers in Tropical Crops by Means of Machine Learning Models. Technologies and Innovations for Development. Springer Paris, 265-277.

Jiménez, D., Pérez-Uribe Andrés., Satizábal, H.F., Barreto S Miguel A., Van Damme, P., Tomassini, M. (2008). A survey of artificial neural network-based. modeling in agroecology. Softcomputing Applications in industry. B. Prasad (Ed.), Springer-Verlag Berlin Heidelberg, Chapter X, 247-269.

Scheldeman, X., Willemen, L., Coppens d'Eeckenbrugge, G., Romeijn-Peeters, L., Restrepo, M.T., Romero Motoche, J., **Jiménez, D.,** Lobo, M., Medina, C.I., Reyes, C., Rodriguez, D., Ocampo, J.A., Van Damme, P. and Goetgebeur, P. (2007). Distribution, diversity and environmental adaptation of highland papaya (*Vasconcellea* spp.) in tropical and subtropical America. In: Hawksworth, D.L. & Bull, A.T. (Eds.). Series Topics in Biodiversity and Conservation. Vol. 6. Plant Conservation and Biodiversity. Springer. Dordrecht. The Netherlands, 293-310.

D. Publications in Congresses and Symposiums

Jiménez, D., Satizábal, H.F., Pérez-Uribe Andrés. (2007). Modelling Sugar Cane Yield Using Artificial Neural Networks. The 6th European Conference on Ecological Modelling, ECEM'07. Trieste, Italy. November 27-30, 244-245.

Barreto S Miguel A., **Jiménez, D.**, Pérez-Uribe Andrés. (2007). Tree-structured Self-Organizing Map component planes as a visualization tool for data exploration in agro-ecological modeling. The 6th European Conference on Ecological Modelling, ECEM'07. Trieste, Italy. November 27-30, 55-56.

Satizábal, H.F., **Jiménez, D.,** and Pérez-Uribe, A. (2007). Consequences of Data Uncertainty and Data Precision in Artificial Neural Network Sugar Cane Yield Prediction. International Work-Conference on Artificial Neural Networks (IWANN 2007) Sandoval, F; Prieto, A; Cabestany, J; Graña, M. (Eds.), San Sebastián, Spain, June 20-22, 2007, LNCS 4507, Springer, 1147-1154.

JOURNAL REVIEWER – DISSERTATION JURY

Reviewer for:

- Agricultural Systems (Elsevier) (http://www.elsevier.com/locate/agsy)

Member of M.Sc Dissertation

- Edwin Erazo Mesa (Universidad Nacional de Colombia – Sede Palmira, 2011) "*Efecto de los factores limitantes de la productividad del cultivo de caña de azúcar a nivel intrasuerte*". *Maestría en Ciencias Agrarias con Énfasis en Suelos*.

SYMPOSIUMS - CONFERENCES – COURSES

Symposium. Forum International Tech For Food. *Les nouvelles technologies au service de l'agriculture et de l'alimentation dans les pays du sud.* Février 25, 2009. Février 26, 2008. Mars 6, 2007. Au Salon International de l'Agriculture. Paris, France.

Conference. The 6th European Conference on Ecological Modelling, ECEM'07. November 27-30, 2007. Istituto Nazionale di Oceanografia e di Geofisica Sperimentale. Trieste, Italy.

Course. Biomodelado con Técnicas de Inteligencia Computacional Aplicado a Sistemas Agroecológicos y Agroindustriales. Noviembre 20-24, 2006. Centro Internacional de Agricultura tropical (CIAT) Cali, Colombia.
Speaker, Presentations:
"Modelado con técnicas de inteligencia computacional aplicado a sistemas agroecológicos".
"las perspectivas de trabajo con sistemas bioinspirados en la agricultura y campos relacionados".

Course. *El uso de sistemas de información geográfica para el análisis de datos de Biodiversidad. Marzo 10-14, 2003. Centro Internacional de Agricultura tropical (CIAT) Cali, Colombia*.