# Proceedings

# International Symposium on Statistics in Agriculture and Environmental Research

## SATELLITE CONFERENCE

Para Ciat, con
mucho cariño,

María Cristina

Good Luck for
the future of
CIAT!

Aug 29/96

**ORGANIZED BY:**
CIAT Biometry Unit
CIAT, Palmira, June 7-9, 1995

IV Annual Meeting of the International Biometry Network for Central America, the Caribbean, Colombia and Venezuela.
VI Statistics Symposium of Universidad Nacional de Colombia
V Annual Meeting of the Inter American Statistical Institute (IASI)

CIAT
Centro Internacional de Agricultura Tropical
International Center for Tropical Agriculture

COLCIENCIAS

These proceedings were published with the financial support of COLCIENCIAS, an entity that promotes scientific and technological development in Colombia.

# CONTENTS

Pag

**Session 1: STATISTICAL METHODS IN ENVIRONMENTAL RESEARCH**

Pag

## Session 2: STATISTICAL METHODS IN BIOTECHNOLOGY AND MOLECULAR BIOLOGY

## Session 3: STATISTICAL METHODS IN AGRICULTURAL EPIDEMIOLOGY

# Welcome Address

María Cristina Amézquita
Head, Biometry Unit, CIAT
Satellite Conference Organizer

On behalf of the CIAT Biometry Unit, I take great pleasure in welcoming you all to the 1995 Satellite Conference of the "International Symposium on Statistics in Agriculture and Environmental Research," being held here at our Institution. This year's international event brings together the IV Annual Meeting of "The International Biometric Network for Central America, the Caribbean, Colombia, and Venezuela; the VI Annual Meeting of the "Inter-American Statistical Institute" (IASI); and the V Statistics Symposium of the National University of Colombia. The plenary meeting of the University's Symposium, organized by its Statistics Department, will be held next week in Santa Marta on Colombia's beautiful Atlantic coast. We are very pleased that many of you will also attend this meeting.

We would like to thank our sponsors for their significant financial support that made today's event possible. We therefore thank CIAT, the Colombian Institute for the Development of Science and Technology (COLCIENCIAS), the Instituto Colombiano de Estudios Técnicos en el Exterior (ICETEX), the Centro Nacional de Investigaciones en Caña de Azúcar (CENICAÑA), the Fundación para la Educación Superior (FES) and its Environment Division in Cali, and the Environment Division of the Mayor's Office for Cali. We especially appreciate the full financial support given to Caribbean participants to attend both meetings of this Conference by the Barbados Office of the British Overseas Development Agency (ODA).

The three scientific topics selected for the Conference were those suggested at the III Annual Meeting of the Biometry Network held at Caracas, Venezuela, in June 1994:

- Statistical methods in environmental research
- Statistical methods in biotechnology and molecular biology
- Statistical methods in agricultural epidemiology

These topics are highly relevant to the present research agenda of agricultural and environmental research institutions worldwide and in Latin America particularly. Given the present emphasis on natural resources conservation and monitoring; conservation and enhancement of biodiversity; and understanding, monitoring, and controlling epidemiological problems in agriculture, the role and application of quantitative sciences, such as statistics and biometry, need to be discussed. We will have the opportunity to do so during the presentations of Invited Papers and Discussion Sessions.

We are most grateful to all Invited Speakers and Moderators for their willingness to share with us their knowledge, thoughts, and valuable experience in applying statistics to real research problems. Researchers, biometricians, professors, and students will greatly benefit from your experience.

The CIAT Biometry Unit is most gratified in welcoming more than 180 participants to the Conference. We trust that this event will help you further your understanding of the role of statistics and biometry in agriculture and environmental research, and find practical solutions to your research problems. But more importantly, we hope that you will take this opportunity to strengthen your contacts among yourselves, whether you be researchers, biometricians, university professors, or students. Depending on your participation lies the success of our Conference.

Thank you.

# ACKNOWLEDGEMENTS

The organizing committee of the "Satellite Conference" of the "International Symposium on Statistics in Agriculture and the Environment" wishes to thank the International Center for Tropical Agriculture, CIAT, for its constant collaboration during the preparation and realization of the event, with human, financial, and technical resources.

The committee also wishes to express its appreciation for the economic support of COLCIENCIAS, which together with the contributions of ODA, ICETEX, CENICAÑA, FES, and the Environmental Divison of the Mayor's office of Cali, allowed us to assemble a large number of Invited Speakers and conference participants for the study of the Applied Statistical Methods in Agriculture and Environmental Research.

We wish to especially acknowledge the Invited Speakers' interest in sharing their knowledge and the participants' enthusiasm in attending our meeting.

# PREFACIO

La relevancia de las investigaciones para la conservación del Medio Ambiente no es necesario repetirla, pues a nivel mundial es reconocida la prioridad que merece el tema. Dentro de este marco surge la necesidad de definir y evaluar la calidad, el estado, el nivel y los cambios(recuperación o deterioro) de las condiciones del ambiente, tanto al inicio(etapa de diagnóstico), como en la fase generadora de alternativas de manejo y solución(etapa de investigación), para llegar a la fase decisiva(etapa de evaluación de impacto de soluciones), donde se definirán los ajustes necesarios a las alternativas mencionadas.

Para realizar este proceso se requiere tomar buenas muestras, definir indicadores y hacer evaluaciones que tengan alta precisión y confiabilidad. Lo anterior constituye el trabajo de un profesional especializado de la Estadística.

En algunos países desarrollados, especialmente en Canadá y Estados Unidos se ha venido trabajando desde 1989 en el tema de Estadística formal aplicada a estudios del Ambiente y ya se han definido las bases de una nueva rama de esta ciencia: "Environmetrics", cuya traducción al español sería: "Métodos estadísticos en Investigaciones Ambientales". Involucra investigaciones y aplicaciones estadísticas en el área de conservación y recuperación de los recursos de sub-suelo, suelo, agua, flora y fauna nativas y aire. Cada día justifica más su existencia y gana adeptos gracias a su importancia.

El Diseño Experimental y la selección de las muestras en investigación resulta un problema estadístico bien diferente del ya conocido en el área de mercadeo, de encuestas de opinión pública y aún de experimentación científica bajo condiciones controladas, por carecer de la infraestructura de información básica: no es lo mismo encuestar personas que obtener información ambiental donde frecuentemente la definición de la unidad de muestreo es ya de por si, un problema técnico de definición. Se requiere el conocimiento de fuentes de variación, variabilidad espacial, variabilidad temporal y de métodos de muestreo adaptados o diseñados especialmente para estos casos en los cuales las técnicas tradicionales resultan insuficientes.

La Biotecnología constituye una ciencia de apoyo importante para la generación de variedades mejoradas pues permite identificar genes deseables que actúan como huellas dactilares del material biológico en estudio. Al hacerlo, se avanza rápidamente en el entendimiento y obtención de las características morfológicas, fisiológicas y de comportamiento deseables de la especie y en la selección rápida de materiales biológicos promisorios para actuar como padres de cruzamientos que se comportarán mejor, en ambientes específicos. Por la cantidad de variables que estos estudios generan, por el tipo de estas variables, por los supuestos de trabajo, por la diversidad de problemas que estudia y otras muchas consideraciones de tipo técnico, económico y práctico, el análisis de estos problemas requiere atención y uso de técnicas estadisticas especializadas.

La formación de profesionales en estas áreas no está aún formalizada en las Universidades ni en muchas de las Instituciones Nacionales o Regionales de Investigación Agricola en Colombia y otros países de América Latina. Por lo tanto se hace necesaria su difusión, por medio de Simposios que reúnan a aquellos que enfrentan los problemas y desean recibir orientación, formación y asesoría, con quienes ya han superado las fases iniciales y se dedican a avanzar en este amplísimo campo de la ciencia.

**María Cristina Amézquita**
**Unidad de Biometría**

## INVITED SPEAKERS

**Dr. Larry Nelson**
Profesor Emeritus, North Carolina
State University,
Raleigh, NC.

**Dr. Richard Coe**
Biometrician, ICRAF Headquarters
Nairobi, Kenya

**Dr. Gilberto Gallopin**
Land Managenment
CIAT
Palmira, Colombia

**Dr. David Marx**
Departamen of Biometry
University of Nebraska, USA

**Dr. Pedro Ferreira**
CATIE, Turrialba
Costa Rica

**Dr. Alberto Palma**
Servicio de Análisis Económicos y
Estadísticos
CENICAÑA, Cali, Colombia

**Dr. Carlos Moreno**
Servicio de Análisis Económicos y
Estadísticos
CENICAÑA, Cali, Colombia

**Dra. Janet Riley**
Rothamstead Exp. Station
ODA Biometrician, England

**Dr. Emilio Carbonel**
Instituto Valenciano de Investigaciones
Agrárias
Moncada, España

**Dr. Barry Moser**
Department of Statistics
Lousiana State University, USA

**Dr. Raúl Macchiavelli**
Deparment of Statistics
Lousiana State University, USA

**Dr. Orlando Martínez**
Profesor Titular
Facultad de Agronomía
Universidad Nacional de Colombia
Santafé de Bogotá

**Dra. María Cristina Amézquita**
Head, Unit Biometry
CIAT, Palmira, Colombia

**Dra. Myriam Cristina Duque**
Unidad de Biometría
CIAT, Palmira, Colombia

**Dr. James Nienhuis**
Assistant Professor
University of Wisconsin, USA

**Dr. Paul van der Laan**
Eindhoven University of Technlogy
Deparment of mathematics and computing
Science
Einhoeven, The Netherlands

**Dr. Francisco Morales**
Virology Research Unit
CIAT, Palmira, Colombia

**Dr. Aurelio Pedroza**
Sub-director de Investigación, Unidad
Regional Universitaria de Zonas Aridas
Universidad Autónoma de Chapingo
México

## MODERATORS

**Dr. Fernando Chaparro**
Director General
COLCIENCIAS
Bogotá, Colombia

**Dr. Cristian Samper**
Director División de Medio Ambiente
FES
Cali, Colombia

**Dr. Raúl Vera**
Leader, Tropical Lowlands Program,
CIAT
Palmira, Colombia

**Dr. James Cock**
Director General
CENICAÑA
Cali, Colombia

**Dr. Lincoln Smith**
Plant Epidemiologist
Cassava Program
CIAT
Palmira, Colombia

**Dra. Janet Riley**
Rothamstead Exp. Station
ODA Biometrician, England

**Dr. Rafael Flores**
Coordinator
Science and Technology area
INCAP
Guatemala

## PARTICIPANTS

### BRAZIL

Lucio Vivaldi
EMBRAPA, CPAC
Brasilia

### COLOMBIA, BOGOTA

Barreto, Nancy
CORPOICA
A.A. 240142 Las Palmas
Teléfono : 2672710/2833260
Fax : 2828947

Collazos, Hernán
UNISUR

Gutierrez de G, Astrid
CORPOICA
A.A 240142 Las Palmas
Teléfono : 267-2710/283-326
Fax : 2828947

Jurado, Roberto
CORPOICA
A.A 240142 Las Palmas
Teléfono : 267-2710/283-326
Fax : 2828947

Manrique, Carlos
CORPOICA
A.A 240142 Las Palmas
Teléfono : 267-2710/283-326
Fax : 2828947

Martinez C, Jorge
Universidad Nacional de Colombia
Of. 211, Edificio 405, Ciudad Universitaria
Teléfono : (57)-1-3681431

Martinez, Edgar
CORPOICA
A.A 240142 Las Palmas
Teléfono : 267-2710/283-326
Fax : 2828947

Pacheco, Pedro N.
Universidad Nacional de Colombia
Of. 211, Edificio 405, Ciudad Universitaria
Teléfono : (57)-1-3681431

### COLOMBIA, CALI

Coppens, Geo
IPGRI
A.A. 6713
Teléfono : (57-2)4450000
Fax : (57-2)445-0273

Luna, Carlos A.
CENICAÑA
Calle 58 N° 3N-15
Teléfono : 6648025
Fax : 6641936

Vivas D, Liliana
CENICAÑA
Calle 58 N° 3N-15
Teléfono : 6648025
Fax : 6641936

### COLOMBIA, CHINCHINA

Montoya, Esther C.
CENICAFE
A.A. 2427
Fax : (968)504723

Orozco, Lucelly
CENICAFE
A.A. 2427
Fax : (968)504723

## COLOMBIA, IBAGUE

Aldana, María E.
Universidad del Tolima

Melo, Omar
Universidad del Tolima

## COLOMBIA, LA PAILA

Marulanda, Ana María
Ingenio Riopaila S.A.
Teléfono : 922-205191
Fax : 922-205036/205352

## COLOMBIA, MANIZALES

Cruz, Gabriel
Universidad de Caldas
A.A. 275 Calle 65 N° 26-10
Teléfono : (968) 861250

Restrepo, Luis H.
Universidad de Caldas
A.A. 275 Calle 65 N° 26-10
Teléfono : (968) 861250

## COLOMBIA, MEDELLIN

Camargo, Mauricio
Universidad de Antioquia
A.A. 1226 Calle 67 N° 53-108
Fax : 94-341-4526

## COLOMBIA, MONTERIA

Alvarez, Andres
CORPOICA
A.A. 602/603 Km 13 via Cerete
Teléfono : 860211-15
Fax : 860219

Iguaran C, Hugo
Universidad de Cordoba
Km 3 Via Cerete
Teléfono : Conm 3381

Navas, Alejandro
CORPOICA
A.A. 602/603 Km 13 via Cerete
Teléfono : 860211-15
Fax : 860219

## COLOMBIA, PALMIRA

Dominguez, Argemiro
CORPOICA
Teléfono : 92-758161/6
Fax : 92-733687

Garcés T, Deudania
Ingenio Manuelita
Via Cerrito - Palmira
Teléfono : conm 756851
Fax : (922) 732486

Huertas, Carlos
CORPOICA
Teléfono : 758161

Rodriguez, Nubia
CORPOICA
Teléfono : 92-758161/6
Fax : 92-733687

Sánchez, Ines
CORPOICA
Teléfono : 758161

Trochez, Adolfo
CORPOICA
Teléfono : 758161

## COLOMBIA, PASTO

Cepeda, Belisario
Universidad de Nariño Post. Ecología
A.A. 1175/1176 Calle 16 N° 30-07
Teléfono : (927)235850

Solarte C, Maria E.
Universidad de Nariño. Post. Ecología
A.A. 1175/1176 Calle 16 N° 30-07
Teléfono : (927)235204

## COLOMBIA, PEREIRA

Castaño Juan C.
CARDER
Calle 25 N° 7-48 U. Adtiva El Lago
Teléfono : 963-355500
Fax : (963)-355501

## COLOMBIA, POPAYAN

Amaya, Maria del S.
SENA, Calle 4 N° 2-80
Teléfono : 240732/243543
Fax : 237678

## COLOMBIA, SANTA MARTA

Giraldo H, Ramón
INVEMAR
Punta de Betin

Salazar, Juan G.
INVEMAR
Punta de Betin

## COSTA RICA , HEREDIA

Blanco, Fabio
Universidad Nacional
Esc. Ciencias Agrarias, Heredia 3000
Fax : 506-2610035

Camacho, Jorge
Universidad Nacional
Esc. Ciencias Agrarias, Heredia 3000
Fax : 506-2607509

## EL SALVADOR

Napoleón Mejía
CENDA
Programa de Producción Animal

## GUATEMALA

Jorge Matute
INCAP
2, Calle 20-92

Rafaél Flores
INCAP
A. Postal 1188

**JAMAICA**

Reid, Heather
CARDI
University Campus, Mona, Kingston 7
Teléfono : 809-9271231/9274
Fax : 809-9272099

**MEXICO**

Jorge Franco
Colegio de Postgraduados
C.E.C Km. 35.5, carretera a Veracruz

Trujillo, Patricia
Univ. Autónoma de México

**PANAMA**

De León, Edilberto
Universidad de Panama

**SAINT LUCIA**

Fletcher P, Lystra
CARDI
L'Anse Road, PO Box 971 Castries St. Lucia
Teléfono : 807-45-24160/243
Fax : 807-45-26934

Floyd, Sian
CARDI
L'Anse Road, PO Box 971 Castries St. Lucia
Teléfono : 807-45-24160/243
Fax : 807-45-26934

**TRINIDAD**

Lauckner, Bruce
CARDI
University of the West Indies, St. Augustine
Teléfono : 809-645-1205/7
Fax : 809-645-1208

**VENEZUELA**

Laura Pla
Universidad Francisco Miranda
Coro 4101, Apdo 7430 Venezuela
Teléfono : 58-68-78637/5255
Fax : 58-68-78140

Orlandoni, Giampaolo
Univ. de los Andes
Inst. de Estadística Aplicada. Mérida 5101

Ramoni, Josefina
Univ. de los Andes
Inst. de Investigaciones Económicas

Torres, Elizabeth
Univ. de los Andes
Inst. de Estadística Aplicada

Cobo M, Margarita
Universidad Central de Venezuela
Ap 2458, CP 2102, Maracay, Venezuela
Fax : 58-43-456323

**COLOMBIA, CALI**

CIAT
A.A 6713
Teléfono: (57-2)4450000
Fax: (57-2)445-0273

**Director General**
Havener, Robert

**Programa de Forrajes Tropicales**
Franco, Luis H.
Ramirez, Gerardo
Rao, Idupulapati
Ricaurte, Jaumer
Sotelo, Guillermo

**Unidad de Manejo de Información y Servicio de Redes**
Rodriguez, Marco
Rojas, Fernando

**Programa de Manejo de Tierras**
Becerra, Jorge H.
Bell, William
Byne, Sara
Clavijo, Luz A.
Cox, Julie
Crawford, Euan
Hurtado, Martha L.
Jones, Peter
Madrid, Otoniel
Rincon, Mauricio
Thomas, Nick
Urbano, Paloma

**Programa de Arroz**
Berrio, Luis E.
Borrero, Jaime
González, Daniel
Guimaraes, Elcio
Martínez, César
Mojica, Daniel
Reyes, Patricia
Triana, Mónica

**Programa de Fríjol**
Beebe, Stephen
Fory, Luisa F.
González, Alma V.
Pedraza, Fabio
Posso G., Cármen E.
Ramirez E., Hector F.

**Programa de Laderas**
Castaño, Jairo
Feijoo, Alexander
Rubiano, Jorge

**Programa de Yuca**
Cadavid, Luis F.
Calle, Fernando
Castillo, Jesús
Claroz, José Luis
Florchinger, Felicitas
Iglesias, Carlos
Jaramillo, Gustavo
Marín, Norbey
Maya C, María M.
Mejía de T, Sara
Morante, Nelson
Olson, Paul D.
Riis, Lisbeth
Roa R, Carolina
Smith, Lincoln

**Programa del Trópico Bajo**
Ayarza, Miguel
Friesen, Dennis
Hoyos, Phanor

**Unidad de Biometría**
González M, Eliana R.
Lema, German
Mesa, Eloina
Silva, James
Trujillo, Leonardo

**Unidad de Biotecnología**
Arana, Fernando
Corredor, Mauricio
Fregene, Martín

Gaitán , Eliana
González, Orlando
Gutierrez, Janeth
Mayer, Jorge
Mejía, Alvaro
Palacio, Natalia
Tohme, Joseph

**Unidad de Recursos Genéticos**
Andrade, Mercedes

**COLOMBIA, CALI**

Universidad del Valle
A.A. 25360
Teléfono : 3393041

Aragón, Efrain A.
Behar, Roberto
Benavides, Carlos A.
Biojo V, Laura
Burgos, Norma C.
Cabra, María C.
Calvache, Miguel E.
Carria, Leiber
Delgado, Jorge
Díaz, Robby N.
Fradas, Licenia
Gordillo, Marisol
Jaramillo P, Edwin
Klinger, Rafaél
Leyton, Lina Y.
Munoz O. Mario
Ochoa, María F.
Olaya, Javier
Perez, John J.
Quiroga, Francisco
Riascos, Yilton
Sánchez, Milton J.
Silva, Helver M.
Solorzano, Juan B.
Tovar, José R.
Valdéz, Guillermo
Yepes, Mario

# STATISTICAL METHODS

# IN

# ENVIRONMENTAL RESEARCH

# PRINCIPLES OF DESIGN OF AGROFORESTRY EXPERIMENTS

by Larry A. Nelson[1] and Richard Coe[2]

Focus is directed on experiments in which treatment plots consist of woody perennial species planted in association with crop plants. Agroforestry is a well-established practice in lesser developed countries because of the useful products and services provided by the plants being grown in association. More recently it has been studied in the field by research scientists who wish to determine the mechanisms, refine some of the techniques currently being used by the farmers and establish some new more efficient practices. This important research is necessary and must be carried out in spite of the difficulties which are described in this paper.

Agroforestry is a collective term for land-use systems, practices and technologies based on the integration of woody perennials with crops and/or animals. The tree component in these provides a variety of functions related to production (food, fuelwood, fodder, poles...) and/or service (shade, windbreak, erosion control, boundary marking...) ultimately leading to improved and more sustainable land-use as well as reduced environmental degradation. In order to qualify for agroforestry, there must be significant ecological and economical interactions between the woody and non-woody components.

In traditional experimental design, biometricians have put great emphasis on the experimental design itself, the number of replications, randomization and blocking. These we would call the "Fisherian Principles". They apply to agroforestry experiments like they do to any other types of experiments. There are a host of other important design parameters which are important in designing an experiment. Some of these are having a clear concise, researchable set of objectives; selecting a set of treatments which bears upon the objectives; and measuring those response variables which give a good estimate of the outcome of the experiment in terms of the questions asked. In field

Professor Emeritus, North Carolina State University, Raleigh, NC

Biometrician, ICRAF Headquarters, Nairobi, Kenya

experiments, there are others such as using an appropriate size; shape, and orientation of plot; minimizing competition among plots; managing the applications of fertilizer, pesticides, etc. to each species of plant within the plot. It is these other important design parameters which are critical in the design of agroforestry experiments and upon which we will direct our attention in this paper. It should be pointed out that these experiments are long-term in nature (five to 10 years) because of the slow growth of trees and the time necessary for the manifestation of their effects upon other components of the plot. Consequently, any design defect will persist for years and have serious consequences.

Thus far, the introduction of clever applications of existing experimental designs (e.g., systematic designs) or innovative new designs has not been particularly useful because the problem is not with the experimental designs, but more with other aspects of design mentioned in this paper.

Practical aspects are very important in the design of agroforestry experiments. A competent researcher needs a good biological background (including crop science and forestry), a thorough knowledge of agroforestry concepts (including sustainability), and common sense.

Planning is very important in agroforestry experimentation. Time spent in careful planning of all aspects of the research will usually pay off.

Organization of the discussion will be according to important phases and also critical problems in agroforestry research. These are: Objectives, Treatment Selection, Competition, Complexity, Sustainability, and Precision Problems.

## Objectives

One of the challenges in agroforestry experimentation is setting a clear, concise set of researchable objectives. The reason one does everything else in the experiment is because of the objectives. If these are not chosen well, an experiment which is very good in every other respect, could be answering the wrong questions. Complicated phenomena are usually being studied in agroforestry experiments and this could result in confusion as to what the objectives should be. Agroforestry experiments often result in measurement of multiple responses. The objectives need to tie together these responses into some sort of measurable outcome. Some researchers make the objectives too extensive and often they are too vague. Since this is a relatively new area of research, we don't have much of a historical perspective upon which to base these objectives. Some sets of objectives have not been very creative and thus we are studying the same old limited questions time and time again. The field is wide-open for creative objectives.

Another important point concerning objectives is that there are different types of agroforestry trials depending upon the objectives. In one called system trials, the objective is to compare certain responses from two or more systems. In another, the object is to study the processes taking place (e.g., at the tree-crop interface). One has to decide between the different types of agroforestry designs, choose one and write the objectives accordingly. This will have implications regarding design specifications such as treatments chosen, consequences of interplot competition, measurements taken, etc. Too often, stated objectives have mixed the two types of trials and this has made it nearly impossible to choose a suitable compromise design.

**Treatment selection**

First, the treatment set should correspond with the set of objectives. Every treatment should have a purpose in relation to the objectives. There have been many errors in choice of treatments as they relate to the objectives in past agroforestry experiments.

Treatments are usually much more complicated in agroforestry experiments than in monoculture trials. This results from the inclusion of the woody perennial species in addition to the crop plants in the same plot, but also from the fact that these different components may individually receive different cultural treatments (and at different times) and these cultural treatments may also affect indirectly the other components. There is also the matter of location of the plants in relation to the other components and the edges of the plot. Many errors have been made in the past by placing tree rows near the edges of a plot which subjects them or neighboring plots to competition effects and renders the results nonrepresentative of the "system" which the farmer uses on a larger scale. Even the spacing between tree rows within a plot often doesn't reflect what the farmer does in practice in the field.

Each treatment needs to be considered not only in itself but how it relates to other treatments in the experiment. Since two or more factors are often being studied, it makes sense to use full factorial experiments in many cases so that the various effects (including interaction) may be estimated in a "clean" and efficient fashion. There have been a number of instances in the past where comparisons have been made between "package" treatments which had more than one factor being varied at the same time. This often is a result of a poor choice of treatments. A definite concept of what contrasts are to be made in the analysis helps in the choice of treatments to be included in the experiment.

One needs to assess the treatments in terms of potential competition with other treatments in the experiment. Should plots for two treatments end up to be adjacent in the experiment, will there be strong above or below-ground competition? This competition will be discussed in more detail in the section on Competition, but at this point one should be aware that large competitive effects might be expected between certain treatments and if they they are going to be included in the experiment, measures should be taken to reduce the extent of the competition effects.

The inclusion of one or more controls is usually important in agroforestry experiments. If one is comparing pruning or fertilizer on plots which have woody perennial species, with comparable plots with no pruning or fertilizer, there should also be a control treatment in which there are no woody perennial species. And this crop only control should be a realistic one. It should be given a base level of inputs comparable to what the farmer is using in practice.

## Competition

Competition generally is a yield reducing interaction among plants. The net effect is not always a lowering of total output, however, because we have outputs from more than one component to consider. Competition among plants exists even in monocultures. However, there are several types of competition in agroforestry experiments. There is competition among plants of the woody species, that among plants of the crop species, and competition between woody plants and crop plants. Competition may be for nutrients, water, and/or light. It is difficult to separate these different competition effects since they they may produce similar manifestations.

In agroforestry experiments, both above-ground and below ground competition are relevant. The below ground may be more important in many cases and yet it usually it is more difficult to evaluate.

In agroforestry experiments, the intensity of competition often varies in different locations within the plot. Because the trees are often in rows across the plot, the degree of competition between the trees and the crop plant varies according the perpendicular distance out away from the tree row (Fig. 1). In short, the response varies according to where the measurement is taken within the plot. This implies that there is always a question what to measure and where to measure it in the evaluation phase of experimentation.

Often, there is a strong competition between neighboring plots having different treatments. A good example of this is between a treatment plot which contains trees intermingled with crops and a plot on which a monoculture (crop) is grown (Fig. 2). This is especially true where the trees are grown in an area having limited moisture. Invariably, the monoculture crop plot will be mined by the roots of the neighboring plot containing a treatment with trees. The net result of this is that there is a misrepresentation of the effects of the agroforestry treatment in comparison with the control. This usually causes the agroforestry treatment to appear better in relation to the control than it actually is.

One can experience competition not only between the monoculture control and a plot with trees in it, but also between two plots each having trees within. Intermingling of the roots of adjacent plots would seemingly render the results of the experiment invalid. It is difficult to estimate how many past agroforestry experiments have had serious root competition effects which would cause the data to be invalid. This effect is difficult to see on the surface, and one almost needs to dig trenches to verify that there is a root competition problem and to what distance it extends.

Ways of avoiding the root competition problem are to have large plots which give more room for the roots within the same plot without invading neighboring plots, provide underground barriers between plots to prevent root interchange between plots, and to periodically cut the roots along the borders. The physical barriers and root pruning appear to provide only temporary and incomplete relief from the root invasion problem. Pruning of the tops of trees may also temporarily affect growth and nature of the roots in a plot. Pruning modifies the growth and structure of the root. The root competition effect with neighboring plots does not extend as far in shrubs as it does in trees. Another possibility is to leave a non-experimental buffer strip between two adjacent plots, but of course this goes against the goal to have a compact experimental area and there can be border effects introduced (shade and moisture effects) at the interface of the experimental plot and buffer strip. If such a practice is followed, a cropped guard area should be provided between plots and blocks and around the experiment. Another approach is to make the evaluation while the trees are still young and before their roots have grown profusely.

## Complexity

Complexity in agroforestry experiments arises because of a number of factors or circumstances:

(1) There may be several components in an experimental plot (such as tree rows, crops, hedges, or individual trees and even animals) and these may be treated with different cultural treatments. Each component may respond not only to its own

treatment, but to the treatment applied to other components.

(2) Often one needs to evaluate several woody plant species grown in association with one or more crop species. This results in a number of treatment combinations which need to be evaluated.

(3) There may be occasions where it is difficult to use standard plot sizes for all treatments. This goes against acceptable experimental design procedures.

(4) There may be situations where certain treatments need to be situated on an end or a next to a border. This would violate the principle of complete randomization and again violates acceptable design procedures. Researchers have even been known to place "difficult to manage" treatments outside of the experimental area and adjacent to the experiment.

(5) The results of the experiment must be expressed in terms of yields and values of more than one commodity and one must also take into account more than one commodity when evaluating costs in an attempt to study the economic aspects of agroforestry. There is a question of how to assign relative weights to the different outputs (and inputs) when integrating the various responses (or inputs).

(6) There may be strong competition between adjacent plots. The extent and consequences of this competition are difficult to to ascertain because much of it is underground.

(7) The plot is not homogeneous with respect to response (e. g. distance from the tree row may affect the result). Errors have sometimes been made in extrapolating results to a per hectare basis from the plot because the wrong part of the plot has been used as a basis for extrapolation. In alley-cropping research, some researchers have

extrapolated only from the central rows of the alley in converting to yield per hectare. Another error has been to ignore the area occupied by the hedge when calculating yield per hectare. In either case, a sizeable overestimate of yield for the alley cropping treatment would result and this might favor the alley cropping treatment in relation to the control.

(8) Experiments often need to be conducted on sloping lands to simulate farmer conditions where they are used for erosion and moisture control and these may be difficult from the standpoint of orienting plots and plants within plots. Not only the direction of the soil variability gradient but also the risks of erosion must be taken into consideration when orienting plots.

(9) Agroforestry experiments are often long-term (5-10 years) due to the slow growth of trees and usually the effects which trees have on their surroundings are long term. One has to measure not only short-term effects but also long-term effects.

(10) The objectives of agroforestry experiments often involve multifactors so factorial arrangement of treatments may be called for. Such an arrangement implies a minimum of four treatments (and usually more).

(11) There are many different possible arrangements of the plants within a plot. The appropriate one depends upon the objectives of the experiment and the treatment and of course, practical considerations. Often, decisions about the location of the trees (one row vs. two rows of trees and where the rows are located in relation to the edge of the plot) are critical because they determine whether the treatment plot adequately represents the "system" which a farmer might use on a larger scale. An example of location of trees in relation with crop plants is shown in Fig. 3. Treatment 3 has a different spacing between trees than Treatments 1 and 2.

(12) There are systems in which trees are pruned at some stage. There are different possible ways of dealing with the organic material obtained from the pruning. Often it is applied back to the plot and on some occasions it is completely removed. If it is applied to the plot, there is a question as to where it should be placed within the plot in order to simulate what the farmer practices in the field.

(13) There are socioeconomic as well as physical aspects of agroforestry and these must both be given due consideration in the design, conduct of the experiment and evaluation of the results.

(14) There probably are more constraints on at what sites experiments can be located than in monoculture situations. This raises a question as what population of sites the particular experimental site is representing.

(15) It is almost impossible to conduct agroforestry experiments in series at different locations as is done with monocultures. This forfeits information on interaction of the treatments and the environment and also forfeits another source of replication.

**Sustainability**

Sustainability is part of almost every definition of agroforestry. It is a term which is not well-defined even though it is often used. It has to do with conservation and maintenance of fertility in the soil over time through the use of woody perennials in proximity to crop plants. Part of the consideration of sustainability has to do with the selection of treatments (treatments to be included in the experiment should be sustainable). Much of it has to do with the evaluation phase. What parameters in the soil should be measured to reflect conservation and soil fertility maintenance to give a good evaluation of sustainability? Nair (1993) reported: " At present, there is no quantitative measure of sustainability". However, he listed some measures of sustainability which are in use: Calculate the total factor productivity (TFP) of the system over a defined period of time (which could be the summation of total factor productivities of individual components). Another approach is to develop separate indices for biological and socioeconomic characteristics. On the biological side, we are often interested in the growth of multistemmed shrubs. It is not clear what what products such as leaves, fruits and firewood should be measured. Sampling of plant materials for dry matter and nutrient content needs to be done. There are repeated harvests of the woody plants which do not correspond to crop harvests. There is the question of how to handle leaves which have fallen from the trees. In short, the usual evaluations done in the forestry profession when dealing with pure stands of trees do not apply well here.

Root studies are important, especially when one is studying mechanisms. This is a very difficult area because of the intensive labor involved and the high level of variability of the root data.

Cost-benefit analysis involves quantitative evaluation of the values of inputs and outputs for assessing economic importance. One difficulty here is that there is a combination of various outputs which all should enter into the consideration of economic output. How should these components be weighted? And often, the input-output relationships are measured on experiment stations and there is a question as to how relevant this is to individual farms.

**Precision Problems**

There is a great potential for having poor precision in agroforestry experiments. This may result from a variety of factors. A few are listed below:

(1) Often agroforestry experiments are conducted in the tropics and tropical soils are known to be quite variable.

(2) To simulate farmer conditions, often they are conducted on slopes and this may compound the problem (i. e. tropical sloping soils may be extremely variable).

(3) Many species used in agroforestry experiments have not benefitted from extensive breeding programs and consequently the germplasm may be highly variable. This may contribute to the precision problem.

(4) Large plots are often needed to represent the system being studied. This may minimize the number of replications possible.

(5) In certain cases, there needs to be nonexperimental buffer strips between plots, and this increases the area of each block and raises the possibility of considerable within block variability (Fig. 4).

(6) Often there are a number of cultural treatments being applied to the different components in a plot, so this could result in a compounding of the nonuniformity of application errors of the different cultural treatments to the different components. There is a time dimension to these applications also (raising the possibility of a compounding of errors over time).

(7) The long-term nature of the experiments may also result in some changes in vegetative growth over time for certain species and this could be reflected in more variability.

(8) For some response variables, there can be large measurement errors. (e. g. root count data).

As with any other types of experiments, one should remember that some idea of precision of these kinds of experiments is usually available from past experience either personal or from the literature (assuming that we can settle on appropriate response measures to evaluate the agroforestry treatments). With such information along with the sensitivity required and the levels of Type I and Type II errors which are acceptable, it is possible to determine if there is a reasonable chance of achieving the precision goals required with the experiment being designed. In some cases, it might be possible to change the design parameters to improve the precision (e. g. increase the number of replications), but there are other cases where the experimental realities preclude the achievement of the precision required. In such cases, one might question the feasibility of initiating the anticipated experiment.

## Summary

The design of agroforestry experiments involves standard experimental design principles and designs but many complexities arise because of the difficulties of formulating a clear researchable set of objectives, in translating these objectives into a set of measurements on the plant material and/or the economic parameters, the "system" nature of the treatments, competition between adjacent plots, the long-term nature of the effects, and the large plots required. The unique design implications of some of these features of agroforestry experiments have been pointed out in this paper. In spite of the difficulties, these experiments are important and must be conducted.

The answers to design of agroforestry experiments lie not in the development of new experimental designs or statistical techniques, but in developing a set of clear well-thought out objectives, choosing the treatments carefully to answer the questions under study and then conducting a carefully controlled experiment which minimizes some of the common problems such as competition between adjacent plots. The importance of planning in the entire process cannot be minimized.

## Reference Cited

Nair, P. K. R. 1993. *An Introduction to Agroforestry.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

**Figure 1.**

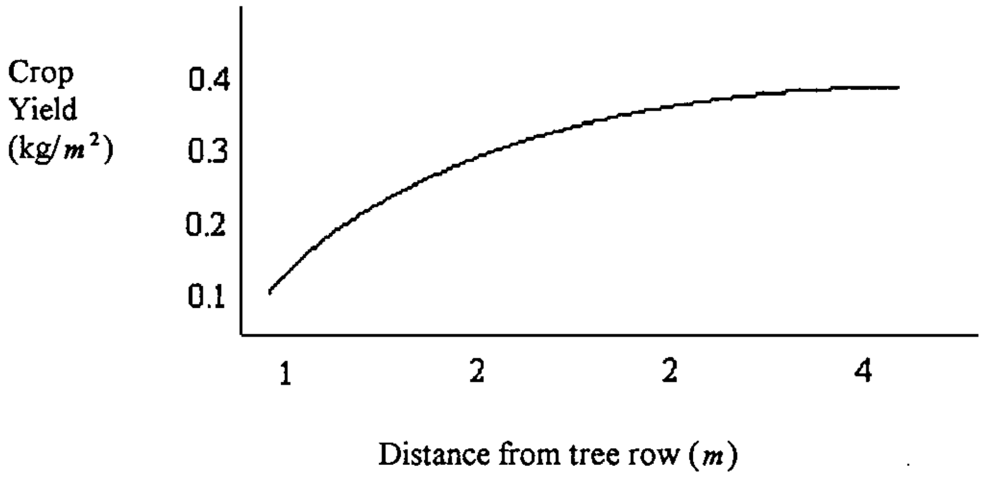


**Figure 2.**

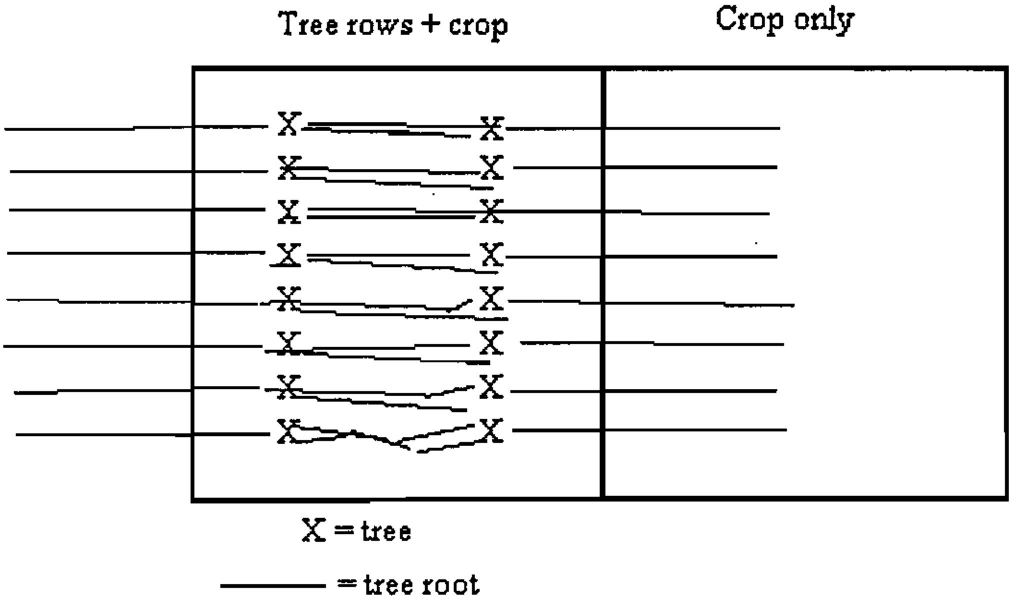
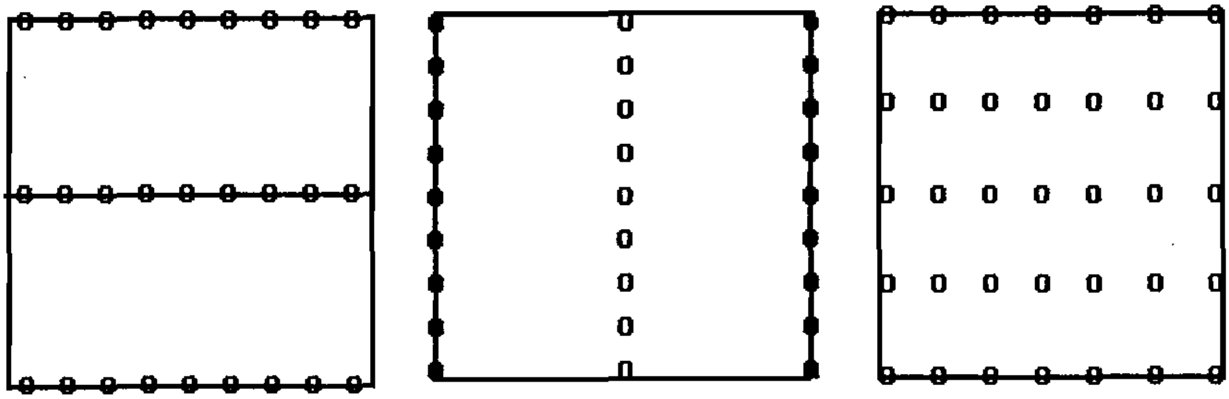


**Figure 3.**



Trt. 1

Trt. 2

Trt. 3

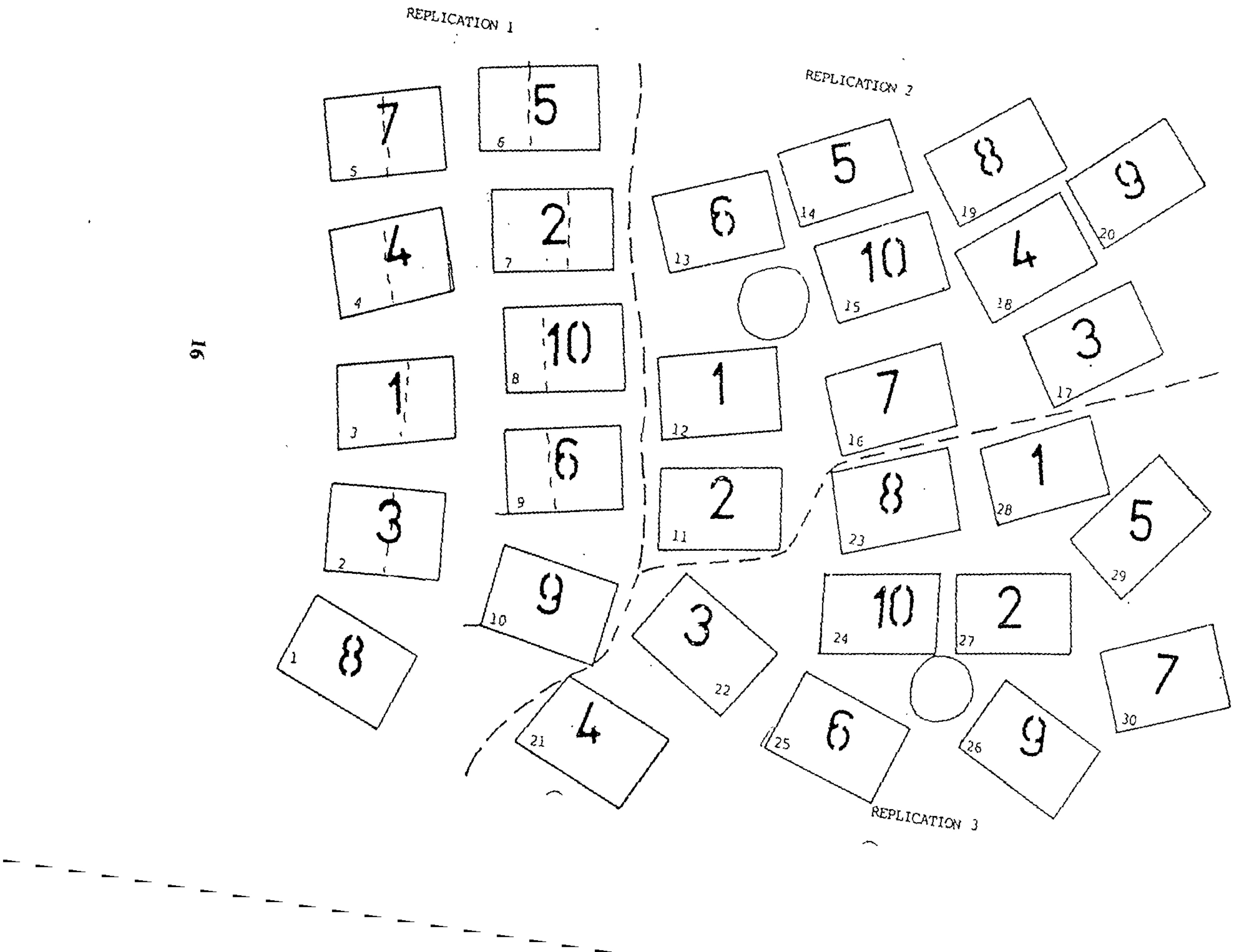

REPLICATION 1
7
5
5
6
4
4
2
7
1
3
10
8
3
2
6
9
8
1
9
10
REPLICATION 2
6
13
5
14
8
19
9
20
10
15
4
18
3
17
1
12
7
16
2
11
8
23
1
28
5
29
3
22
10
24
2
27
7
30
4
21
6
25
9
26
REPLICATION 3

Figura 4.

# CONCEPTUAL CHALLENGES IN NATURAL RESOURCES RESEARCH: BIOMETRICAL ISSUES

Gilberto C. Gallopín[1]

## SUMMARY

By the very definition of "resource", the natural resources research includes a human or economic element. Natural resources may be renewable and non-renewable, exhaustible and non-exhaustible.

Basic compoments of natural resources research are clustered into three major approaches: descriptive, explanatory, and normative. Descriptive research focus on questions such as which are the constitutives elements of the resource system? What is the basic structure of the system? What is it geographical distribution and ahundance? Explanatory research concentrates on the functioning of the natural resources system, its causal structure, and the determinants and consequences of change. Normative aspects of research on Natural Resources concentrate on value ladden issues such as which is the optimal rate of harvest or utilization of the resource, what is its sustainable use, what is the proper way to value the resources.

The role and potential usefulness of Biometry varies widely across the three approaches.

Basic systemic properties of natural resources systems, their relevance as components of the human environment and as the ecological basis for sustainable development, and their implications for biometrical issues are discussed.

[1] Leader, Land Management, CIAT, Apartado Aéreo 6713, Cali, Colombia.

# ANALYSIS OF DESIGNED EXPERIMENTS IN THE PRESENCE OF SPATIAL CORRELATION

David B. Marx and Walter W. Stroup
Department of Biometry
University of Nebraska-Lincoln
Lincoln, NE 68583-0712

Many data sets in agricultural research have spatially correlated observations. Examples include field trials conducted on heterogeneous plots for which blocking is inadequate, soil fertility surveys, ground water resource research, etc. Such data sets may be intended for treatment comparisons or for characterization. In either case, linear models with correlated errors are typically used. The majority of this talk is found in Stroup, Baenziger and Mulitze (1994).

Statistical methods traditionally used in agricultural research have emphasized designs and analyses which assume that variation among experimental units is either (i) homogeneous or (ii) can be controlled by blocking. In many field situations, however, variation is more likely to be characterized by smooth, localized, irregular trends - variation which is neither homogeneous nor necessarily well-controlled by blocking. Figure 1 illustrates the distinction among these three general classes of variability. Imagine that the rows columns represent, say, field plots in a rectangular arrangement and the y-variable is some response of interest.

In the third case, with irregular local trends, the pattern of variability can often be characterized by a linear model with spatial correlated errors. Mixed linear model methods (Henderson, 1975; Harville, 1976, 1977; McLean, Sanders, and Stroup, 1991) are therefore a useful tools to analyze such data. A derivation of the mixed linear model with geostatistical implications is given in appendix A.

The example of particular interest here is cultivar evaluations. Most cultivar evaluation trials use blocked designs, such as Latin squares (LS), randomized complete blocks (RCB), or incomplete block designs (IBD) and are analyzed using classical analysis of variance (ANOVA). However, standard ANOVA for blocked designs does not account for spatial variability. Recent advances in spatial statistics suggest that superior alternatives may exist. In our example we will compare RCB-ANOVA with two nearest neighbor adjustment (NNA) methods and a generalized least squares approach to removing spatial variability. Yield data from a breeding nursery involving diverse, adapted, and unadapted germplasm, grown in Alliance, Nebraska during 1988-1989, is used in the comparisons.

Because spatial homogeneity within blocks is extremely uncommon in field experiments with more than 8-12 plots per block, the RCB assumption is rarely satisfied in agronomic trials involving large sets of treatments or cultivars. Plot to plot variation within a block is affected by competition between cultivars within the trial (Jensen and Federer, 1964; Kempton, 1982; Kempton and Lockwood, 1984), by soil variation and fertility, by previous land use (Pearce, 1978, 1980), and by various weather related conditions (such as snow cover; Fowler, 1979). It seems that in such experiments, some method of accounting for within block variation would be beneficial.

Nearest neighbor adjustment (NNA; Wilkinson et al., 1983; Besag and Kempton, 1986) is a relatively simple method of accounting for within block variation caused by all of the above conditions. Recently, a random field approach (Zimmerman and Harville, 1991), a mixed linear model with spatially correlated errors, has been proposed as an alternative to NNA methods.

The conventional analysis of variance procedure for the RCB, i.e. partitioning sources of variation according to the model

$$y_{ij} = \mu + r_i + \tau_j + \epsilon_{ij},$$

*where* $y_{ij}$ *is* the observation on i$^{th}$ block, j$^{th}$ treatment,

$\mu$ is the overall mean,

$r_i$ is the i$^{th}$ block effect,

$\tau_j$ is the j$^{th}$ treatment effect, and

$\epsilon_{ij}$ is random error.

The $\epsilon_{ij}$'s are assumed independent with constant variance $\sigma^2$, which implies no spatial trend.

A second method of analysis was a procedure using adjacent residuals to correct for within block variability among plots (Papadakis, 1937). This procedure will hereafter be referred to as NNA-PAP. Plots were coded by their latitude and longitude. A residual at the i$^{th}$ latitude and j$^{th}$ longitude was defined as:

$$e_{ij} = y_{ij} - (\text{cultivar mean})_{ij},$$

where $e_{ij}$ denotes residual and $y_{ij}$ observed yield for the ij$^{th}$ plot. Longitudinal nearest neighbor adjustment will hereafter be referred to as the east-west (EW) adjustment for the ij$^{th}$ plot. Note that a given field may be oriented in any direction, so the "EW" is strictly for notational convenience. The EW adjustment was computed as:

$$EW_{ij} = 1/2(e_{i,j-1} + e_{i,j+1})$$

where $e_{i,j-1}$ is the residual in the neighboring plot to the "west" of the ij$^{th}$ plot and $e_{i,j+1}$ is the residual to the east. If a plot was on the border and had no east or west neighbor, $EW_{ij}$ was calculated averaging the remaining terms. Border plots are the subject of much controversy and there is no conclusive "right way" to handle them.

Latitudinal nearest neighbor adjustment (north-south, NS, adjustment) was computed as:

$$NS_{ij} = 1/2(e_{i-1,j} + e_{\ i+1,j})$$

where $e_{i-1,j}$ is the residual in the neighboring plot to the north and $e_{i+1,j}$ is the residual to the south. Border plots were handled the same way as the EW adjustment.

These terms were then used for analysis of covariance using the model

$$y_{ij} = \mu + \tau_i + \rho_1 EW_{ij} + \rho_2 NS_{ij} + e_{ij}$$

where $\rho_1$ and $\rho_2$ are the regression coefficients associated with the EW and NS NNA covariates, respectively. Tests of significance of spatial trend (i.e. of $\rho_1$ and $\rho_2$) and of cultivar means adjusted for spatial trend were obtained. Three adjustment procedures were computed using these terms: EW, adjust for longitudinal trends only; NS, adjust for latitudinal trends only; and EWNS, adjust for trends in both directions by using both EW and NS covariates. There are other "patterns" of covariates which can be used, but only these were calculated in our example.

Wilkinson et al. (1983) found that the Papadakis procedure can be improved because the residual is computed from entry means, which are inaccurate as a result of within block trends. They suggested recomputing the residuals using the adjusted cultivar means from the first analysis, i.e.

$e_{ij} = y_{ij}$ - (adjusted cultivar mean)$_{ij}$,

then recomputing the analysis of covariance, and repeating this procedure until the differences between adjusted cultivar means in successive iterations are negligible. This iterative procedure was applied to the best of the NNA models above, best being determined by which trends were significant in the analysis of variance after the first application of the NNA procedure.

The third procedure was a NNA procedure developed by Schwarzbach (1984), hereafter referred to as NNA-SB. For every plot, a "nearest neighbor difference" was computed as:

$NND_1 = y_{ij} - 1/2(y_{i,j-1} + y_{i,j+1})$,

where $y_{i,j-1}$ is the observation in the neighboring plot to the west and $y_{i,j+1}$ is the observation to the east. Then, the mean $NND_1$ for each cultivar was computed and denoted $NND_a$. A second "nearest neighbor difference" was computed as:

$NND_2 = \overline{y}_{(ij)} - 1/2(\overline{y}_{(i,j-1)} + \overline{y}_{(i,j+1)})$

where $\overline{y}_{(ij)}$ denotes the cultivar mean for the cultivar grown in the ij$^{th}$ plot. The mean $NND_2$ for each cultivar was also computed and denoted $NND_e$. Cultivar means were then adjusted using the following formula:

$$\text{adjusted mean} = \text{cultivar mean} + 3/4(NND_a + NND_e).$$

As in the NNA-PAP procedure, this procedure was iterated. The coefficient 3/4 is the typical value used to minimize oscillations in adjusted cultivar means on successive iterations, improving the algorithm's efficiency. Note that in NNA-SB, adjustment occurs in only one direction.

A fourth procedure, generalized least squares or random field linear model (Zimmerman and Harville, 1991), was also evaluated. In the random field linear model (RFLM) approach, observations are characterized according to the model (see also appendix A)

$$y_{ij} = \mu_i + e_{ij},$$

where $y_{ij}$ is the jth observation on the ith cultivar, $\mu_i$ is the ith cultivar mean, and $e_{ij}$ is the residual for the ijth observation. In standard RCB analysis, the $e_{ij}$'s are assumed to be independently distributed with a constant variance of $\sigma^2$. This corresponds to the assumption of no spatial trends. In the random field approach, the $e_{ij}$'s are assumed to distributed according to some spatial correlation model descriptive of local trends. The idea is that when local trends exist, neighboring plots tend to be more alike than those farther apart.

Spatial correlation models originally developed for geostatistics (Journel and Huijbregts, 1978) appear to be particularly useful. In geostatistics, the correlation structure is described by the semivariance $\Gamma(h)$. Consider two residuals, say $e_{ij}$ and $e_{i'j'}$, a distance of h units apart, then the semivariance is defined as

$$\Gamma(h) = (1/2)\, \mathrm{Var}[e_{ij} - e_{i'j'}].$$

Several semivariance models exist. One of particular use to agronomic field trials is the spherical model

$$\Gamma(h) = K + C(0)\{1.5(h/r) - 0.5(h^3/r^3)\}, \text{ if } h < 0$$

$$\Gamma(h) = 1 \text{ if } h \geq r$$

where a is the "range" - the distance at which correlation between observations is effectively zero, K is the "nugget" effect, and C(0) is the "sill" or covariance at distance zero. In agronomic data, the "nugget" is typically negligible and was dropped from the analysis discussed in this paper. A typical semivariance will be zero at distance zero and increase with distance until it reaches a plateau - called the sill - at distance r. The sill is interpreted as the variance of the yield.

The covariance of two observations h units apart is related to the semivariance as

$$C(h) = C(0) - \Gamma(h).$$

Thus, using the spherical model,

$$\mathrm{Var}(e_{ij}) = \sigma^2$$

$$\mathrm{Cov}(e_{ij}, e_{i'j'}) \quad = \sigma^2\{1-1.5(h/r)+0.5(h^3/r^3)\} \quad \text{if } h<a,$$

$$= 0 \qquad \text{otherwise.}$$

The unknown parameters describing variance and spatial correlation, $\sigma^2$ and r, were estimated using restricted maximum likelihood procedures (Harville, 1977) and the cultivar means, $\mu_i$, were estimated using generalized least squares.

The RCB and NNA-PAP analyses were implemented using PC-SAS (1988). The NNA-SB procedure was implemented on a PC using Schwarzbach's Analysis of Field Trials (ANOFT) software (version 33, U.K.Z.U.Z. Branch of Variety Testing, 656 06 Brno, Hroznova 2, CSFR, Czechoslovakia, 1989) and Agrobase (Mulitze, 1992). The RFLM was implemented using SAS-IML and results were recomputed using SAS-MIXED (1992). See appendix B for SAS-MIXED specifications of the RFLM procedure.

One purpose of NNA procedures is to adjust means on basis of spatial trends to more accurately estimate cultivar value. Because adjustments could affect ranking of means, it is important to compare NNA adjusted means with RCB unadjusted means. The ten highest yielding lines in a trial at Alliance as estimated by RCB, NNA-PAP, and NNA-SB with their ranks are given in Table 1, to illustrate the effect of NNA on estimated cultivar means where spatial trends are large. Where spatial trends were present, large differences between the relative ranks of RCB and NNA means could be found. Only four of the top ten RCB cultivars were included in the top ten cultivars of NNA procedures. Most striking is that the NNA estimated highest yielding line ('Buckskin') was ranked 28 by the RCB. Adjusted means were 687 kg/ha (40% for NNA-PAP) to 807 kg/ha (47% for NNA-SB) higher than the RCB mean. Hence only a fifty percent selection differential would have resulted in retaining this line in the program. Similarly, NE86527 was ranked 44 (of 56) by RCB, but 8 in both NNA procedures.

With the large difference in ranks between the RCB and NNA procedures, it is reasonable to ask if this may be a statistical fluke or if a trial with a high RCB CV (27.6%) is worth using. Breeders often discard trials with CVs greater than 15% as being unreliable.

Past experience would indicate that the rankings for the NNA analyses are probably correct. Alliance is a difficult site for accurate yield measurements due to sporadic winterkilling and wind erosion. It is notoriously hard to block (Fowler, 1979), leading to very high CVs. The cultivar mean yields were low, hence compressed with small differences between cultivars. Large adjustments between analytical procedures would cause large changes in rank, as seen in Table 1. As for specific changes in ranking, Buckskin is a high yielding wheat cultivar widely grown in western Nebraska and eastern Wyoming because it has good wintersurvival and rapid spring regrowth which reduces wind erosion. In two of the four replications, Buckskin was grown in an area that had winterkilling and wind erosion, thus reducing its mean yield in the RCB analysis. Also 'Brule' and 'Redland', a selection from Brule, historically do not differ for yield, yet in the RCB analysis they ranked four and 25, respectively. In the NNA-PAP analysis, they ranked six and seven - a more reasonable result. Finally, Alliance, despite its high CV, is an important testing site because approximately one third of Nebraska's wheat crop is grown in western Nebraska under similar conditions.

Another view of the Alliance trial is provided by the random field analysis applied to these data. Analysis of the residuals using the methods of Zimmerman and Harville (1991) resulted in a pattern of spatial correlation, accurately described by a spherical semivariance model with an estimated range of 18.1 m and a sill of 61.6 (kg ha$^{-1}$)$^2$. These values measure the variance and spatial correlation among plots. Since the plots were 4.3 m long (trimmed to 2.4 m), the range indicates that the data on plots up to 18.1 m (four experimental units) apart were spatially correlated - evidence of strong local field trends. Table 1 includes the estimated cultivar means using the spherical model (RFLM). The general rankings of the cultivars using the spherical correlation model and NNA analyses were similar. RCB means and ranks were poorly correlated with any of the spatially adjusted means and ranks. Spatially correlated data typically provide an inflated estimate of experimental error and inaccurate estimates of treatment means using methods which assume independence (RCB). In view of the magnitude of spatial correlation in these data, the very high CV obtained using the RCB model is not surprising.

Several improvements in analysis have been reported. The most promising appears to be the random-field linear model approach of Zimmerman and Harville (1991). Their work indicates that random field methods are even more accurate than the NNA procedures. The results obtained for the Alliance trial are consistent with Zimmerman and Harville. A lack of computing software has severely limited the usefulness of such methods, but as new software is introduced this problem is not likely to persist. For the present, the NNA procedures remain the most easily implemented procedures using currently existing software. Their value is clearly demonstrated by the trials reported in this paper.

## TABLE 1

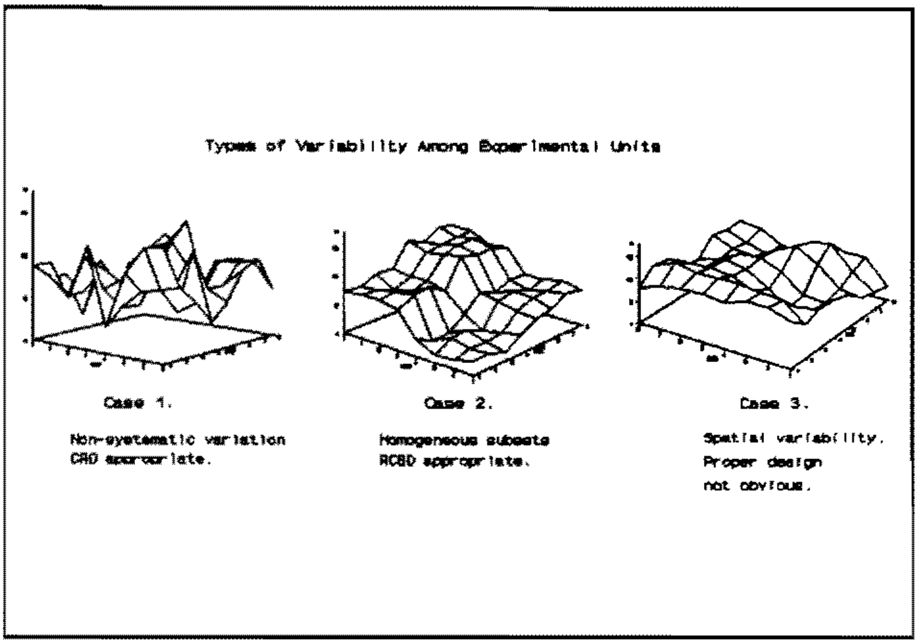
| | | MEANS | | | | RANKS | | | |
|---|---|---|---|---|---|---|---|---|---|
| Location | Cultivar | (1) | (2) | (3) | (4) | (1) | (2) | (3) | 4 |
| Alliance | NE86503 | 2200 | 1860 | 1890 | 1890 | 1 | 12 | 12 | 7 |
| | NE87619 | 2100 | 2030 | 1980 | 2030 | 2 | 4 | 5 | 3 |
| | NE86501 | 2080 | 1760 | 1730 | 1780 | 3 | 24 | 28 | 21 |
| | Redland | 2050 | 1920 | 1980 | 1960 | 4 | 6 | 4 | 5 |
| | Centurk78 | 2040 | 1860 | 1830 | 1870 | 5 | 14 | 21 | 8 |
| | Rahwide | 2030 | 2040 | 2050 | 1990 | 6 | 2 | 2 | 4 |
| | Siouxland | 2030 | 1740 | 1810 | 1680 | 7 | 27 | 22 | 31 |
| | NE86606 | 2000 | 1870 | 1890 | 1850 | 8 | 11 | 13 | 12 |
| | Arapahoe | 1980 | 1820 | 1910 | 1800 | 9 | 17 | 9 | 16 |
| | NE87613 | 1980 | 1880 | 1900 | 1830 | 10 | 9 | 11 | 13 |
| | | | | | | | | | |
| | Buckskin | 1720 | 2400 | 2530 | 2330 | 28 | 1 | 1 | 1 |
| | NE85556 | 1780 | 2040 | 2010 | 2060 | 23 | 3 | 3 | 2 |
| | Karl | 1620 | 1930 | 1960 | 1860 | 37 | 5 | 7 | 10 |
| | Brule | 1750 | 1890 | 1880 | 1940 | 25 | 7 | 14 | 6 |
| | NE86527 | 1480 | 1880 | 1930 | 1810 | 44 | 8 | 8 | 15 |
| | NE86507 | 1600 | 1870 | 1900 | 1850 | 39 | 10 | 10 | 11 |
| | | | | | | | | | |
| | Scout 66 | 1850 | 1860 | 1960 | 1820 | 17 | 14 | 6 | 14 |
| | NE87409 | 1440 | 1860 | 1840 | 1870 | 50 | 13 | 19 | 9 |

(1) : RCB

(2): NNA-PAP

(3): NNA-SB

(4): RFLM



Types of Variability Among Experimental Units

Case 1.

Non-systematic variation
CRD appropriate.

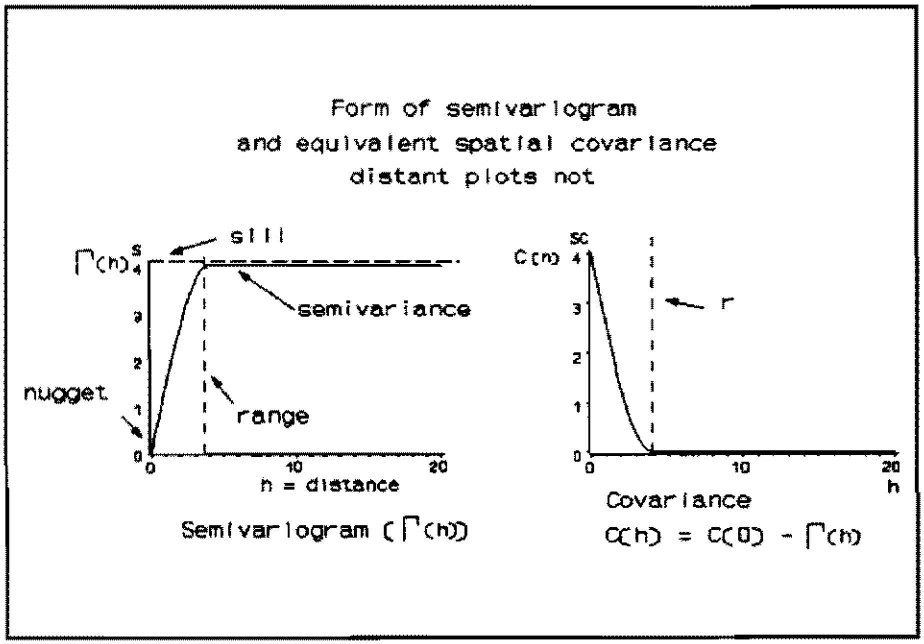
Case 2.

Homogeneous subsets
RCBD appropriate.

Case 3.

Spatial variability.
Proper design
not obvious.

**Figure 1**



Form of semivariogram
and equivalent spatial covariance
distant plots not

Semivariogram (Γ(h))

Covariance
C(h) = C(0) - Γ(h)

**Figure 2**

## References

Besag, J. and R. Kempton. 1986. Statistical analysis of field experiments using neighboring plots. Biometrics 42:231-251.

Cressie, N. 1991. Statistics for Spatial Data. John Wiley: New York

Fowler, D.B. 1979. Selection for winterhardiness in wheat. II. Variation within field trials. Crop Sci. 19:773-775.

Gusmao, L. 1986. Inadequacy of blocking in cultivar yield trials. Theor. Appl. Genet. 72:98-104.

Harville, D.A. 1976. Extension of the Gauss-Markov Theorem to include the estimation of random effects. Ann. Statist. 4:384-395.

Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Statist. Assoc. 72:320-340.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:19-28.

Jensen, N.F., and W. T. Federer. 1964. Adjacent row competition in wheat. Crop Sci. 4:641-645.

Journel, A.G., and Huijbregts, C.J. 1978. Mining Geostatistics. Academic Press, London, England.

Kempton, R.A. 1982. Adjustment for competition between varieties in plant breeding trials. J. Agric. Sci. Camb. 98:599-611.

Kempton, R.A., and G. Lockwood. 1984. Inter-plot competition in variety trials of field beans (Vicia faba L.) J. Agric. Sci. Camb. 103:293-302.

Louw, J.H. 1990. A selection index to cope with genotype-environment interaction with an application to wheat breeding. Plant Breeding 104:346-352.

Martin, R.J. 1982. Some aspects of experimental design and analysis when errors are correlated. Biometrika 69: 597-612.

McLean, R.A., W.L. Sanders, and W.W. Stroup. 1991. A unified approach to mixed linear models. Amer. Statistician. 45:54-63.

Mulitze, D.K. 1992. Agrobase IV Reference Manual. Agronomix Software, Inc. Portage la Prairie, Manitoba, CANADA.

Papadakis, J.S. 1937. Methode statistique pour des experiences sur champ. Bulletin de l'Institut d'Amelioration des plantes. Thessalonike 23.

Pearce, S.C. 1975. Row-and-column designs. Applied Statistics 24: 60-74.

Pearce, S.C. 1978. The control of environmental variation in some West Indian maize experiments. Trop. Agric. (Trinidad) 55:97-106.

Pearce, S. C. 1980. Randomized blocks and some alternatives: A study in tropical conditions. Trop. Agric. (Trinidad) 57:1-10.

Pearce, S. C., G. M. Clarke, G. V. Dyke, and R. E. Kempson. 1988. Manual of Crop Experimentation. Oxford University Press. New York, NY.

SAS Institute Inc. 1988. SAS/STAT User's Guide Release 6.03 edition. SAS Campus Drive, Cary, NC.

SAS Institute, Inc. 1992. SAS Technical Report P-229. SAS/STAT Software: Changes and Enhancements, Release 6.07. SAS Campus Drive, Cary, NC.

Schwarzbach, E. 1984. A new approach in the evaluation of field trials. The determination of the most likely genetic ranking of varieties. Proc. EUCARPIA Cer. Sect. Meet., Vortr. Pflanzenz. 6:249-259.

Stroup, W.W. and Mulitze, D.K. 1991. Nearest neighbor adjusted best linear unbiased prediction. American Statistician, 45:194-200.

Stroup, W.W., P.S. Baenziger and D.K. Mulitze. 1994. Removing Spatial Variation from Wheat Yield Trials: A Comparison of Methods. Crop Science, Vol 34, No. 1, 62-66.

Wilkinson, G. N., S. R. Eckert, T. W. Hancock, and O. Mayo. 1983. Nearest neighbor (NN) analysis of field experiments (with discussion). J. Royal Statistical Soc., Series B 45:152-212.

Zimmerman, D.L. and D.A. Harville, 1991. A random field approach to the analysis of field-plot experiments and other spatial experiments. Biometrics 47: 223-239.

## Appendix A

The general form of the mixed linear model is as follows.

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \mathbf{e}, [1]$$

where $\mathbf{y}$ is a vector of observations;
X is a matrix of constants (describing regression or design structure) for the fixed effects;
$\boldsymbol{\beta}$ is a vector of fixed effects parameters;
Z is a matrix of constants for the random effects;
$\mathbf{u}$ is a vector of random effects parameters; and
$\mathbf{e}$ is a vector of residuals.

For the random effects, $\mathbf{u}$ and $\mathbf{e}$, $E(\mathbf{u}) = E(\mathbf{e}) = \mathbf{0}$, $Var(\mathbf{u}) = G$, $Var(\mathbf{e}) = R$, and $Cov(\mathbf{u},\mathbf{e}') = 0$. In "traditional" models, e.g. standard analysis of variance models for completely random and randomized block designs, $R = I\sigma^2$ is assumed, but mixed model theory places no requirements on R or G - both can be general.

Inference with mixed linear model has three basic building blocks. The first is the *mixed model equation*, used to estimate $\boldsymbol{\beta}$ and $\mathbf{u}$, given as follows:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}\mathbf{y} \\ Z'R^{-1}\mathbf{y} \end{bmatrix} [2]$$

In most cases, the variance and covariance components of G and R are unknown, and estimates must be used. Estimates are typically obtained using restricted maximum likelihood (REML), although other methods can be used.

The second building block of inference is the *predictable function*, $K'\boldsymbol{\beta} + M'\mathbf{u}$, which is *predictable* is $K'\boldsymbol{\beta}$ is *estimable*. Adjusted marginal treatment means (a.k.a. "Least Squares means"), treatment differences, and contrasts are typical predictable functions of interest to researchers.

The third building block of interest is the "standard error," or, more accurately, the square root of the prediction error of the estimated predictable function. The standard error is given by the formula $\sqrt{L'CL}$, where $L' = [K'\ M']$ and C is the generalized inverse of

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}$$

When L is a vector, the ratio estimate/(standard error) is a t-statistic and be used as such. In the more general case, where L is a matrix

$$\boldsymbol{\Theta}'L(L'CL)^{-1}L'\boldsymbol{\Theta}/\mathrm{rank}(L), \text{ where } \boldsymbol{\Theta}' = [\boldsymbol{\beta}'\ \mathbf{u}']$$

is an approximate F-statistic.

Several spatial correlation models are potentially useful for agricultural data. Typically, spatial correlation refers to variability among experimental units in a single location (e.g. among plots in a field) and is thus modeled through the residual vector, **e**, that is, the covariance matrix R. Zimmerman & Harville (1991) discuss several alternative structures for R. Many of these were originally developed for applications in geostatistics (Journel & Huijbregts, 1978). This basic idea is as follows.

If a typical field trial with spatial variability, responses of plots close together are likely to be highly correlated, whereas plots farther apart are less correlated. At some critical distance, responses of plots farther apart are essentially uncorrelated. In geostatistics, the *semivariogram* is used to characterize spatial variability. The *semivariance* defined as

$$\Gamma(h) = \tfrac{1}{2}\mathrm{Var}(\text{difference between pairs of observations } h \text{ units apart})$$

The semivariogram is a plot of $\Gamma(h)$ versus h. A typical semivariogram is given in Figure 2, below. The key features of the semivariogram are the *range*, defined as the critical distance above which observations are uncorrelated, the *sill*, the semivariance of uncorrelated observations (equal to the error variance, it turns out), and the *nugget*, defined as the semivariance at distance zero. The nugget describes abrupt changes and was originally developed to model data from searches for diamonds, where probes a very short distance apart could find nothing or a very high concentration of diamonds. Such abrupt variation is uncommon in agricultural field trials, so the nugget is frequently assumed to be zero.

The semivariance is related to the R matrix in the mixed model as follows:

$$\mathrm{Cov}(2 \text{ observations } h \text{ units apart}) = C(h) = C(0) - \Gamma(h).$$

Thus, C(0) corresponds to the diagonal elements of R and the C(h) are the off-diagonal elements of R. Typical semivariance models for the mixed model are

*Spherical*

$$C(h) = C(0)[1 - (3h/2r) + (h^3/2r^3)], \text{ if } h < r$$

$$= 0, \text{otherwise}$$

*Exponential*

$$C(h) = C(0)[\exp(-h/r)]$$

*Gaussian*

$$C(h) = C(0)[\exp(-h^2/r^2)]$$

*Linear*

$$C(h) = C(0)[1-hr], \text{if } h < 2/r$$
$$=0, \text{otherwise}$$

Note that for all of the above models, two parameters, C(0) and r, corresponding to the error variance and range, respectively, must be estimated.

## Appendix B

SAS, PROC MIXED (SAS Institute, 1992) permits specification of spatial correlation models such as those described in this introduction. **ß** and **u** are estimated using the mixed model equations [2], and $\sigma^2$ and r are estimated using REML. In the following sections, we will present examples of basic applications of mixed models with spatial correlation. We will present the basic SAS-MIXED programming requirements, highlights of the output of particular interest, and our experiences with problems and pitfalls users should anticipate.

### BASIC *PROC MIXED* PROGRAMMING

Consider the simplest mixed model with spatially correlated errors,

$$y_{ij} = \mu + f_i + r_j + e_{ij},$$

where $\mu$ and the $f_i$'s are fixed, the vector **u** of $r_j$'s is distributed $N(\mathbf{0}, I\sigma_R^2)$, and the vector **e** of $e_{ij}$'s is distributed $N(\mathbf{0},R)$, where R is some spatial covariance matrix. To analyze data using this model using PROC MIXED, the following input statements are minimally required:

```
DATA a;
INPUT fix_eff rand_eff row col y;
```

Note that our convention here will be that words in capital letters are mandatory ***verbatim*** in the program and words in lower case are mandatory but the specific word is user's choice. The variables "fix_eff" and "rand_eff" name the fixed and random factors in the model, "row" and "col" locate the observation in space, and "y" is the observed response.

The following is the basic SAS program. Variables in *italics* are not mandatory, but we have found them to be useful options.

```
PROC MIXED SCORING=n;
CLASS fix_eff rand_eff;
MODEL y=fix_eff;
PARMS ($\sigma_R^2$) (nugget) (range) (sill);
RANDOM rand_eff;
REPEATED / SUBJECT=rand_eff LOCAL TYPE=SP(SPH) (row col);
```

The *SCORING* option forces the REML algorithm to use a scoring procedure at least *n* times per iteration. We have found this option to be very helpful in obtaining convergence to reasonable solutions for the range and sill. The values in parenthesis in the PARMS statement are initial numeric values for the variance and covariance parameters. Typically, the nugget is assumed to be zero, so the *nugget* option often will not be used in the PARMS statement. $\sigma_R^2$ will only be specified is one wishes to include the random effect in the model (which, note, is included in a separate RANDOM statement, not in the MODEL statement *a la* SAS-GLM). Our experience is that it is essential to specify initial estimates of the range and sill; PROC MIXED's default initial values frequently lead to grossly unreasonable estimates of the sill and range.

The REPEATED statement specifies the structure of the covariance matrix, R, of the **e** vector. If there is a random effect in the model, use SUBJECT=rand_eff; otherwise use SUBJECT=INTERCEPT. *LOCAL* is used if the nugget is assumed to be non-zero. The TYPE statement specifies the covariance (or semivariance) model to be estimated. The example above is for the spherical semivariance model. Other options include EXP (exponential), GAU (gaussian), and LIN (linear). Consult the SAS manual for other options.

# MEASURING STABILITY AND DEGRADABILITY OF AGROFORESTRY SYSTEMS

P. Ferreira and D. Kass
CATIE, Turrialba, Costa Rica

## ABSTRACT

Tropical agriculture usually deals with severe limitations on management alternatives, arising from a combination of factors such as low-resource farmers, fragile lands, high pest susceptibility, etc. Under these conditions, agroecosystem properties other than productivity, such as stability, sustainability, degradability, etc. are of crucial importance.

Although a conceptual framework has been elaborated on these topics, the problem of evaluating agroecological system properties and comparing systems is still an almost open field. In addition to evaluating system properties, an appraisal of the relative importance of factors which affect those properties should be given.

## 1.Introduction

The management of an agroecoystem is a multiple goal task, seeking the optimization of productivity, economic benefit, stability, and the conservation of resources which are basic for the sustainability of production (Watts, 1973). Therefore, the search for optimum management alternatives requires not only the usual quantification of productivity and economic benefit, but the measurement of production stability and a close examination of long term resource trends.

The concern about production stability and resource conservation is especially important to tropical agriculture where the combination of fragile lands, high pest susceptibility and low resource farmers sets a severe limit on possible management alternatives.

This paper discusses the problem of measuring and analyzing the stability and degradability of fragile land agroecosystems from a statistical point of view. The methodology proposed is an extension of the one given in Ferreira which has been already applied to the study of alley cropping stability by (see Sánchez 1989 and Sánchez et al. 1990).

The definitions of stability - constancy of production from harvest to harvest in the face of small disturbing forces arising from the normal fluctuation and cycles in the surrounding environment-, and of sustainability - the ability to mantain a specified level of production over the long term - are taken from the literature (Conway, 1987 and Marten, 1988, respectively).

Degradability is defined as the rate of production decay per unit time. When an agricultural technology system is implemented, a trend might be observed in production, due to various reasons such as reduction in soil organic matter, exportation of soil nutrients, etc. Systems are classified as degrading, upgrading or constant in tendency according to this behaviour. Long term degrading systems are non-sustainable; however, some short term degrading systems may become sustainable in the long run. Notice that, under constant practices, no degradability is expected in the long term.

This paper discusses the statistical evaluation of stability and degradability of agricultural technology systems (ATS) using data from an alley cropping experiment. More general approaches of the subject have been presented elsewhere (see Barnett et al., 1994, and de Camino and Müller, 1993).

## 2. Quantitative Analysis of System Properties.

The evaluation of system properties is frequently adifficult task because system performance is highly dependent on environmental conditions. Stability, sustainability and degradability are in this sense situational because they depend on the magnitude or duration of the disturbance which induces fluctuation in production. This problem severely limits the generalization of agroecosystem assessments from one set of environmental conditions to another (Marten, 1988).

As a consecuence, the evaluation of the impact of specific components of environmental fluctuation, such as precipitation or radiation, on system properties, seems to be a crucial issue.

Mead et al. (1986) classify the approaches to the description and analysis of stability under three main headings, (i) variation, (ii) environmental dependence, and (iii) risk. The authors state that measures of variation are usually given in terms of coefficients of variation, the major criticism being that the structure of the data samples has been usually ignored.

As an approach to the assessment of environmental dependence, the Finlay - Wilkinson's (1963) method of regression on an environmental index, given by the mean of the treatments on each environment, is mentioned. The major drawback of this approach appears to be the dependence of the environmental index on the set of treatments being considered and the lack of stochastic independence between regressions.

In this paper, total instabilities (TI) are measured through variations from harvest to harvest around production means. On the other hand, adjusted instabilities (AI) are measured as variations from harvest to harvest when certain environmental or trend factors are kept constant. The reduction observed when comparing TI and AI measures the influence of those factors and provides a decomposition of TI into components due to those environmental or trend factors. Some authors consider that the identification of sources of instability may be even more important than obtaining a quantitative measure of stability
(Marten and Rambo, 1986).

The approach bears some resemblance with the Analysis of Variance methodology proposed by Eberhardt and Russell (1966). However, the present paper does not consider indexes based on mean yields per environment. Instead, environmental factors are included into models and their contribution to instability is assessed.

In addition, degradabilities are defined as production trends through time. Observed trends in raw data, which are affected by the specific behaviour of environmental variables during the observational period, are distinguished from adjusted trends, which estimate degradabilities under average values and average fluctuations of environmental variables. Although these last estimates are site-specific, they have the appealing property of not being period-specific.

For example, a decreasing trend in precipitation may lead to a decreasing trend in production, leading to an erroneous evaluation of ATS behaviour. Adjustment by precipitation should show a corrected trend and an appropriate estimate of degradability.

As an example of application, bean yield data from a five years maize-beans relay system with several treatments, set up on a farmer's field, including Gliricidia sepium (Jacq.) both as a mulch and in an alley cropping system, are analyzed. The purpose of this example is to illustrate a methodology and not to make an exhaustive analysis of the experiment.

## 3. An Example With Real Data

Bean yield from an alley cropping experiment (see Kass and Araya, 1987), set up on farmer's field in Jilgueral, Costa Rica, were chosen to illustrate the proposed methodology. The experiment was established in 1983 and measured for five years. Beans were planted in September and maize in April of each year. Precipitation in the area, as given by the October precipitation in mm measured at the Jilgueral Weather Station, National Institute of Meteorology is presented
in Figure 1.

The experiment was laid out in a randomized block design with four replicates. The treatments were formed as a combination of the following:

GM - Gliricidia mulch (Gliricidia sepium (Jacq.) Walp.)
GAC - Gliricidia Alley Cropping (6666 trees/ha)
N - as NH4 N03, 30 kg/Ha

NLM - Non-leguminous mulch (Hypharrenia rufa, 1 Kg/m$^2$)

H - Herbicide (Alaclor 1.8 l/ha. + Pendimethaline 0.75 l/ha.)

C - Control, with no nitrogen and no herbicide and are given in the following table:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| GM | GAC | N | NLM+N | H+N | H | C | NLM |

Degradabilities and stabilities were computed using five years data for all the treatments.

To understand the general behaviour of the experiment some preliminary analyses are necessary. A general analysis of variance is presented in Table 1 showing significance for all the sources of variation with the exception of blocks.

**Table 1. Analysis of variance of bean yields (kg/ha)**

| Source | d.f. | SS(1) | C.V. |
|---|---|---|---|
| Blocks | 3 | 594.59 | 39.96% |
| Treatments | 7 | 2185.19** | |
| Error(a)=BlxTr | 21 | 1484.01 | |
| Years | 4 | 11873.84** | |
| Treat. x Years | 27 | 2491.79** | |
| Error (b) | 93 | 3333.39 | |
| C.Total | 155 | 21962.81 | |

(1) SS x 10-3

Means of bean yields are given in Table 2.

Although the TreatmentsYears interaction is significant, an overall means comparison is presented in this table as a general appraisal of treatment performances.

**Table 2. Means of bean yields (kg/ha) over the 1983-87 period(*). The treatments order is given by their overall performance.**

| **Year** | **Treatment** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **GM** | **GAC** | **N** | **NLM+N** | **H+N** | **H** | **C** | **NLM** |
| 1983 | 1428.50 | 1250.04 | 1142.68 | 636.16 | 1114.13 | 985.36 | 657.00 | 592.76 |
| 1984 | 383.99 | 278.86 | 375.39 | 349.81 | 326.59 | 165.77 | 172.00 | 198.66 |
| 1985 | 964.21 | 671.42 | 756.89 | 980.19 | 734.76 | 583.63 | 397.00 | 440.00 |
| 1986 | 350.25 | 559.11 | 165.21 | 292.75 | 147.62 | 334.82 | 117.00 | 99.57 |
| 1987 | 315.07 | 421.18 | 316.23 | 298.81 | 167.88 | 133.57 | 217.29 | 203.46 |
| Mean | 688.41 | 636.12 | 551.28 | 511.54 | 498.20 | 440.62 | 312.06 | 306.89 |
| a | ab | ab | ab | ab | ab | b | b | a |

(*)Overall means with the same letter do not differ significantly according to a Tukey's test (5%).

From Table 2 we observe that GM gave the highest overall yield followed by GAC. However, it is interesting to observe that during the last two years, the best yields were given by the GAC treatment. It is also noticed comparing the C, H and NLM treatments with the N, H+N and NLM+N, respectively, that the addition of N has almost always shown an increase in yields throughout the whole experimental period.

Degradabilities were computed as slopes of linear regressions of yield over years. Relative or adjusted degradabilities have also been computed as slopes of yields over years but are related to models that include the variable precipitation.

**Table 3. Unadjusted (upper figures) and adjusted (lower figures) degradabilities (kg/ha/year) measured by the slope of a linear regression of yield on years.**

| Treatment | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| GM | GAC | N | NLM+N | H+N | H | C | NLM | StdErr |
| -226.06** | 31.46 | -186.31** | -73.18* | -207.15** | -153.45** | -93.44** | -87.77** | 29.93 |
| 112.09 | -78.61 | 88.67** | -85.63* | -41.07 | -35.96 | -23.96 | 41.95 | -67.25 |

The adjustment through precipitation records shows that most of the treatments have non-significant degradabilities. In addition, a significant positive tendency is shown by the GAC and the NLM+N treatment. Furthermore, a significant negative tendency is observed for the H+N treatment. A similar but non-significant behaviour is shown by the N treatment.

The importance of the adjustment is clear from Table 3, in which significant negative or degrading tendencies are shown by the upper figures. These significances are due to a decrease in precipitation through the experimental period (Figure 1).

Sums of squares, with 4 degrees of freedom, measuring total variation about the overall treatment means, or total instabilities (TI), are partitioned into three components, corresponding to reduction sum of squares: the first one due to precipitation, the second one due to a linear trend over years adjusted for precipitation (LT/P), and the third one due to non-linear components of the trend or unexplained variation, (U), with 1, 1 and 2 degrees of freedom, respectively (Table 4a).

The same table is presented as percentages of total instabilities (Table 4b) to enable a better understanding of treatment behaviour.

**Table 4. Partition of total instability (TI) of yields into components due to precipitation (P), linear trend adjusted for precipitation (LT/P) and unexplained (U/LT,P), as given by (a) sum of squares and (b) percentages of total variation.**

**(a) Sum of squares(1)**

| | GM | GAC | N | NLM+N | H+N | H | C | NLM |
|---|---|---|---|---|---|---|---|---|
| P | 2998.3** | 145.7* | 1743.7** | 1140.7** | 2179.6** | 1431.5** | 460.0** | 465.3** |
| LT/P | 92.1 | 175.6* | 125.9 | 160.1* | 149.3* | 34.4 | 26.3 | 11.7 |
| U/P,LT | 790.4** | 25.9 | 639.5** | 117.1 | 458.3** | 527.1** | 285.1 | 182.3* |
| TI | 3880.8** | 347.2* | 2509.1** | 1417.8** | 2787.2** | 1992.9** | 771.4** | 659.3** |

(1)Sum of squares x 10-3

**(b) Percentages of total instabilities**

| | GM | GAC | N | NLM+N | H+N | H | C | NLM |
|---|---|---|---|---|---|---|---|---|
| P | 77.3 | 42.0 | 69.5 | 80.4 | 78.2 | 71.8 | 59.6 | 70.6 |
| LT/P | 2.4 | 50.5 | 5.0 | 11.3 | 5.4 | 1.7 | 3.4 | 1.8 |
| U/P,LT | 20.3 | 7.5 | 25.5 | 8.3 | 16.4 | 26.5 | 37.0 | 27.6 |

According to Tables 4(a) and (b), variation in precipitation explains most of the instability of yields. However, the effect of precipitation over the GAC treatment is not very strong, explaining only 42% of the total instability.

Instability due to a linear trend over time, adjusted for precipitation, are significant only for the GAC, the NLM+N and the H+N treatment. The linear trend of yields through time explains a very small portion of the total instability, usually less than 10%, a conclusion which is related to the non-significant character of adjusted degradabilities given in the second line of Table 3. This reduction, taken as a % of the total instability, is markedly higher for the GAC treatment (51%), which shows a more predictable linear behaviour of yields.

It is not clear whether the partition of total instability (Table 4a) and its discussion in terms of percentages (Table 4b) should be based on sum of squares or mean squares. The presentation given in this paper has the advantage of giving a decomposition which adds to the total instability. Its disadvantage comes from the fact of comparing chi-square variables with different degrees of freedom. The choice between both alternative presentations is similar to the choice between using a multiple correlation coefficient R or the corresponding adjusted R.

Instabilities expressed as coefficients of variation are presented in Table 5. To obtain the total instability CV, the sums of squares in the last line of Tables 4a are first divided by 4 to produce sum of squares of deviations of the treatment means per year from their overall treatment means. As a second step these figures are again divided by 4. Finally the square roots of the resulting figures are divided by the corresponding means in the last line of Table 2.

Note that all the sum of squares are divided by 16. Again it might be argued that instead of 16 we should consider the alternative of using as denominator 4 times the appropriate degrees of freedom of each reduction. This alternative method has some conceptual advantages. In the chosen method of presentation the columns of Table 5 satisfy the inequalities $PLT \leq P \leq TI$. These appealing properties are lost when using the alternative denominators.

**Table 5. Total instabilities (TI) and instabilities adjusted by precipitation (P) and precipitation and linear trend (PLT) as given by % coefficients of variation.**

| Treatment | PLT | P | LT | TI |
|---|---|---|---|---|
| GM | 32.3 | 34.1 | 49.2 | 71.5 |
| GAC | 7.8 | 21.6 | 27.6 | 28.4 |
| N | 36.3 | 39.7 | 48.0 | 71.8 |
| NLM+N | 16.7 | 25.3 | 53.6 | 58.2 |
| H+N | 34.0 | 39.1 | 51.9 | 83.8 |
| H | 41.2 | 42.5 | 58.2 | 80.1 |
| C | 42.8 | 44.7 | 52.0 | 70.4 |
| NLM | 34.8 | 35.9 | 48.3 | 66.1 |

Some changes are observed in the ranking of total instabilities when comparing Table 5 with 4a. The H+N and H treatments are now showing maximum CV instabilities while the GM treatment, which was at the top in Table 4a, ranks in fourth place in Table 5.

As a general comment it might be said that the strong reduction observed when going from column TI to column P is due to the high influence of precipitation in the total instability. When comparing columns P and PLT, small differences are observed except for the GAC and NLM + N treatments both showing a positive significant adjusted degradability in Table 3.

Again it is observed that the adjustment by precipitation produces a much bigger reduction than the adjustment by linear trend.

The GAC treatment shows uniformly smaller figures in the four columns, its total instability being a half or a third part of those of the remaining treatments.

# BIBLIOGRAPHY

CONWAY, G.R. (1987) The properties of agroecosystems. Agricultural Systems 24(2), 95-117.

EBERHART, S.A. and RUSSELL, W.A. (1966) Stability parameters for compariing varieties. Crop Science 6, 36-40.

FINLAY, K. W. and WILKINSON, G.N. (1963) The analysis of adaptation in a plant-breeding programme. Australian J. Agr. Res. 14:742-754.

FRANCIS, CH. A. and KING, J.W. (1988) Cropping systems based on farm-derived, renewable resources. Agricultural Systems 27, 67-75.

KASS, D.L. and ARAYA, J.F. (1987) Alley cropping with Gliricidia sepium (Jacq.) Walp. on farmers field in Costa Rica. In: Gliricidia sepium (Jacq.) Walp. management and improvement. NFTA Spec. Pub. 87-01. p.50-58.

KENDALL, M.G. and STUART, A. (1963). The Advanced Theory of Statistics. Volume 1. Distribution Theory. Hafner Publ. Co., New York.

MARTEN, G.G. (1988) Productivity, stability, sustainability, equitability and autonomy as properties for agroecosystem assessment. Agricultural Systems 26, 291-316.

MARTEN, G.G. and RAMBO, A.T. (1988) Appendix 3: Guidelines for writing comparative case studies of southwest asian rural ecosystems. In: Agroecosystem Research for Rural Development. Edited by K. Rerkasem and A.T. Rambo, Multiple Cropping Centre and SUAN. Chiang Mai, Thailand.

MEAD, R., RILEY, J., DEAR, K. and SINGH, S.P. (1986) Stability comparison of intercropping and monocropping systems. Biometrics 42, 253-266.

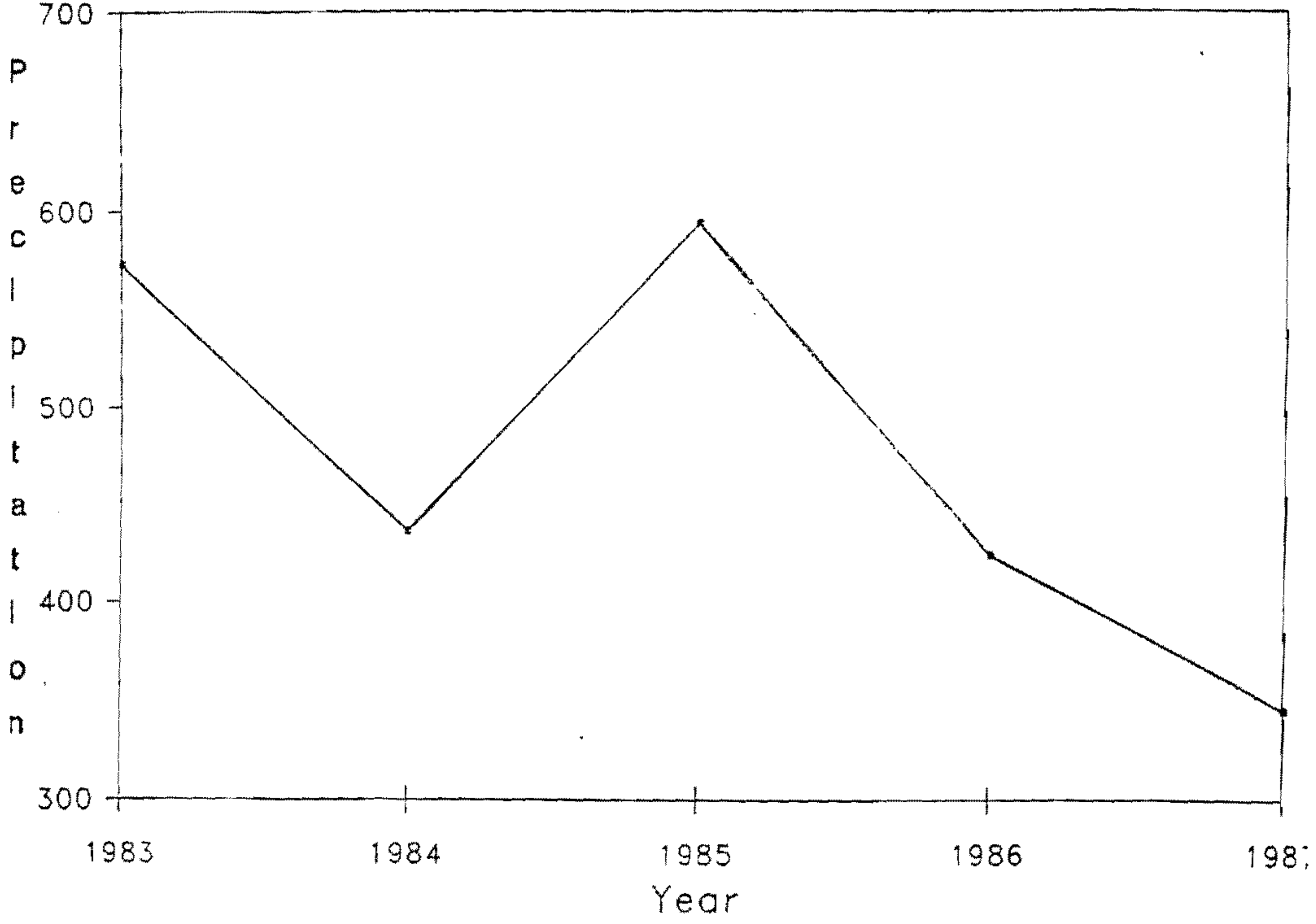
SANCHEZ, J.F. (1989) Análisis de la estabilidad y dinámica de sistemas de producción de cultivos en callejones. M.Sc. Thesis, 173 p. CATIE, Turrialba, Costa Rica.

SANCHEZ, G., SANCHEZ, J. and FERREIRA, P. (1990) Yield stability of maize (Zea Mays L.) and mountain immortelle (Erythrina poeppigiana) Walp. O. Cook in different tree spacings. Agronomy Abstracts, 1990 Annual Meetings, San Antonio, Texas, American Soc. of Agr., p. 62.

SCHEFFE, H. (1959). The Analysis of Variance. John Wiley, New York.

BARNETT, V., LANDAU, S., WELHAM, S.J. (1994) Measuring sustainability. Environmental and Ecological Statistics 1, 21-36.

WATTS, K.E.F. (1973) Principles of Environmental Science. McGraw-Hill, New York.



700
600
500
400
300
Precipitation
1983
1984
1985
1986
198;
Year

# ANALISIS DE INFORMACION SOBRE RESULTADOS DE PRODUCCION COMERCIAL EN CULTIVOS DE CAÑA DE AZUCAR EN COLOMBIA

## RESUMEN

Este documento presenta dos casos de análisis de datos correspondientes a resultados comerciales en un cultivo de caña de azúcar. Análisis que han permitido retroalimentar el proceso de investigación, encontrar variables que son de mayor relevancia para la toma de decisiones en el momento de las comparaciones tanto a nivel comercial como experimental, construir metodologías para diagnosticar el manejo dado al cultivo y generar alternativas que conlleven a incrementos en la eficiencia del subsector azucarero y sobretodo que contribuyan a la sostenibilidad de los recursos naturales; y finalmente resaltar que el análisis de datos es una herramienta fundamental para conseguir mejores decisiones operativas, administrativas y gerenciales.

En el caso 1 se analiza el ingreso neto por tonelada de azúcar, dado unos costos, como una variable que conjuga toneladas de caña y azúcar recuperable por unidad de caña como alternativa más justa para hacer comparaciones entre las diferentes tecnologías. Además se utiliza el ingreso neto por tonelada de azúcar para explorar la mejor forma de incrementarlo.

En el caso 2 se analizan registros sobre manejo de agua, mediante el uso de la técnica de los modelos lineales y se llega a mostrar la posibilidad de incrementar sustancialmente la eficiencia al disminuir costos de producción en la práctica de riego, consiguiendo como consecuencia la conservación del recurso hídrico.

# ANALISIS DE INFORMACION SOBRE RESULTADOS DE PRODUCCION COMERCIAL EN CULTIVOS DE CAÑA DE AZUCAR EN COLOMBIA

**Centro de Investigación de la Caña de Azúcar de Colombia, CENICAÑA**
**Servicio de Análisis Económico y Estadístico**

**Alberto Palma Zamora**
**Carlos Arturo Moreno Gil**
**Liliana Vivas Diaz**
**Carlos Adolfo Luna González**
**Cali, Colombia, Abril 1995**

## INTRODUCCION GENERAL

Con el objeto de empezar estudiar variables que pueden ser utilizadas en las comparaciones de las diferentes tecnologias generadas por el Centro de investigación y adoptadas por la industria y conocer el comportamiento de la producción del sector azucarero Colombiano, se han recopilado datos de los resultados comerciales anuales de toda la industria entre 1960 y 1980; datos de producción mensual y algunas variables de clima para unos pocos ingenios a partir de 1981; y desde enero de 1990 datos completos de producción y datos parciales de manejo y clima. En la actualidad hay más de 160000 Ha. de tierras cultivadas en caña de azúcar a lo largo y ancho del valle geográfico del rio cauca, de los cuales aproximadamente el 75% es cosechada cada año y durante todo el año. El hecho de cosechar caña todos los dias permite que el cultivo esté sometido permanentemente a las variaciones climáticas diarias, aumentando considerablemente los factores que pueden afectar la producción.

El análisis de estos datos, además de proporcionar un conocimiento del cultivo a nivel comercial y retroalimentar la investigación, ha permitido también empezar a resaltar ante los administradores y técnicos de los ingenios la importancia del análisis de datos como herramienta fundamental en la toma de decisiones.

En el manejo de estos datos comerciales se pueden presentar algunos inconvenientes como los siguientes:

- Presencia de datos inconsistentes generados como resultado de cambios en tecnologías u otro tipo de eventos no registrados, en estos casos nos hemos apoyados en técnicos que tienen un buen conocimiento de la industria.
- Limitaciones por el rango de las variables que no permiten detectar relaciones, cuando este es muy pequeño, así estas relaciones sean fuertes (por ejemplo la edad de cosecha).
- Confusión de efectos al atribuirle a una variable un efecto, cuando puede ser debido a una interacción con otra.

Respecto a las características de estos estudios podemos decir que son análisis con datos históricos del sector azucarero (estudio retrospectivo), analizando variables a través del tiempo (estudios longitudinales), y también para un momento específico (estudios transversales). Con estos datos también se han realizado análisis descriptivos, comparativos (de efecto a causa) y se han cuantificados efectos y descritos comportamientos dependiendo de cambios registrados en las variables de explicación (estudios observacionales). El análisis de estos datos como pseudoexperimentos se caracteriza por que no existe una asignación aleatoria a los diferentes niveles de un factor o combinaciones de niveles de varios factores a las unidades de estudio.

Como se mencionó anteriormente estos análisis han permitido retroalimentar el proceso de investigación, al detectar factores donde es necesario realizar experimentación; establecer pautas para manejar el cultivo de una manera más adecuada mientras se liberan nuevas metodologías resultantes de la investigación, generar alternativas para mejorar los niveles de eficiencia y realizar comparaciones más razonables al momento de la toma de decisiones respecto a variedades, número de cortes, edad de cosecha, etc., utilizando variables que integran tanto resultados de producción como recursos involucrados en la actividad.

# CASO 1
# ANALISIS DESCRIPTIVO DEL INGRESO NETO COMO UNA VARIABLE DE DECISION EN EL CULTIVO DE CAÑA DE AZUCAR

## 1.1. INTRODUCCION

El objetivo de este documento es mostrar cómo a partir del análisis de datos comerciales se llegó a una metodología que logra conjugar en el ingreso neto por tonelada de azúcar dos variables importantes en la toma de decisiones en un cultivo de caña de azúcar como son el tonelaje de caña y el contenido de azúcar, permitiendo hacer comparaciones más reales entre variedades, cortes, zonas u otras tecnologías y fuentes de variación que pueden ser de interés.

## 1.2. ANALISIS DE LA RELACION ENTRE EL TCH Y EL ARE

### 1.2.1 TONELADAS DE AZUCAR POR HECTAREA

En la actualidad el negocio más importante del sector azucarero es producir y vender azúcar. En el campo, la producción de azúcar depende del tonelaje de caña por hectárea (TCH) y de la relación de azúcar recuperable por unidad de peso de caña (ARE).

Para evaluar nuevas tecnologías en un cultivo de caña de azúcar, el sistema más sencillo y por ende más usado, es la comparación del producto del TCH por el ARE que genera las toneladas de azúcar por hectárea (TAH). Pero esta variable presenta algunos inconvenientes ya que se puede generar el mismo nivel de TAH con diferentes combinaciones de TCH y ARE. Lo cual implica que este método por si sólo no permite seleccionar la tecnología más rentable.

Con información comercial se pudo observar que en promedio el TCH disminuía cuando el ARE aumentaba (**figura 1**). Aunque también hay casos de TCH bajos con ARE bajos y TCH altos con ARE altos. Los ingenios y cultivadores de caña de azúcar se preguntan por los valores de TCH y ARE que produzcan los mejores resultados. Es decir, la combinación óptima.

La variable TAH ha sido utilizada tanto a nivel comercial como experimental para comparar variedades, número de cortes y otras tecnologías, pero en la actualidad existen series dudas de utilizar el TAH con criterio de decisión.

En la **figura 2** se involucran las variables TCH, ARE y TAH en una representación tridimensional y se puede observar todas las posibles combinaciones de TCH y ARE para generar un mismo valor de TAH. Es decir:

$$\text{TAH} = F(\text{ARE}, \text{TCH}) \quad (1)$$

Para simplificar la representación interceptamos las superficie (1) por el plano TAH=K y la curva de intercepción la proyectamos en plano ARE*TCH. Esta curva proyectada tiene por ecuación F(ARE,TCH)=K, y esta sería una curva de nivel o una curva de contorno. Al considerar diferentes valores de TAH se obtiene un conjunto de curvas de nivel (mapa de contorno). Si fijamos un valor de TAH=K, entonces la curva de nivel resultante estaría conformada por todas las combinaciones de ARE y TCH que dan como resultado el valor de K para las TAH.

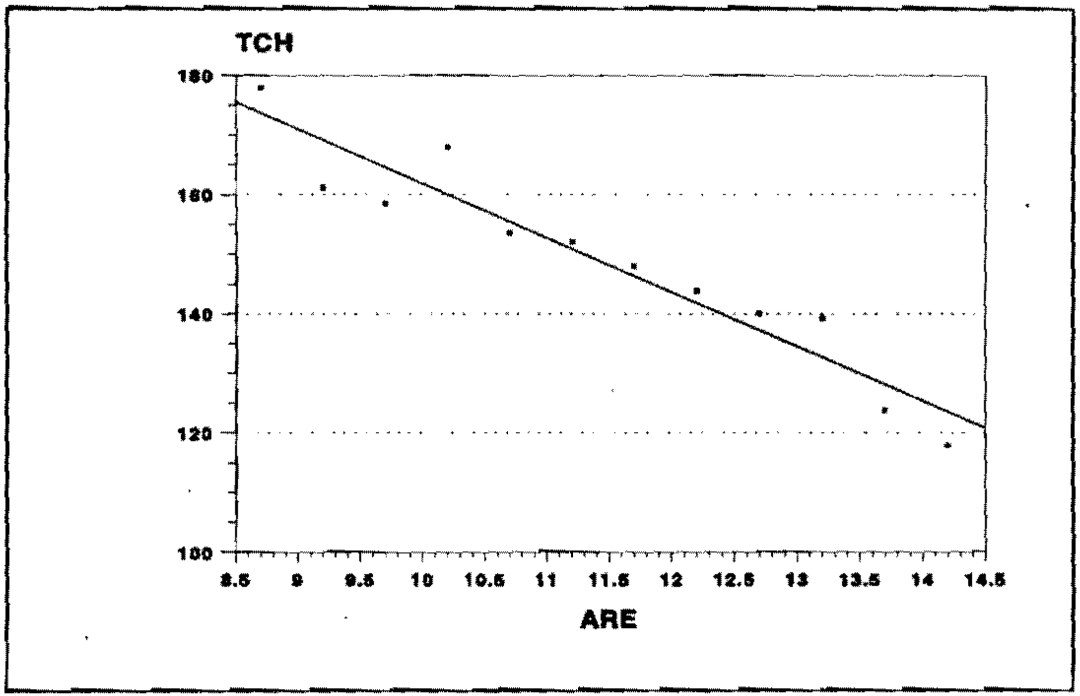


**Figura 1.** Relación entre TCH y ARE.

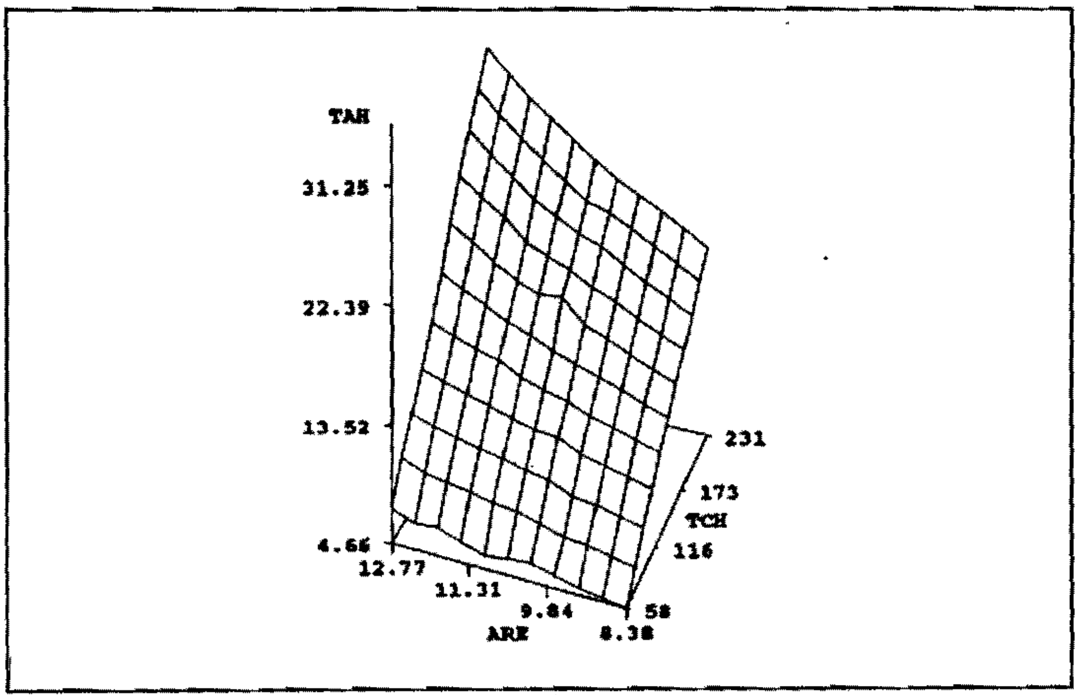


**Figura 2.** TAH como una relación de TCH y ARE

En la **figura 3** se han generado dirferentes curvas de nivel para diferentes valores de TAH.

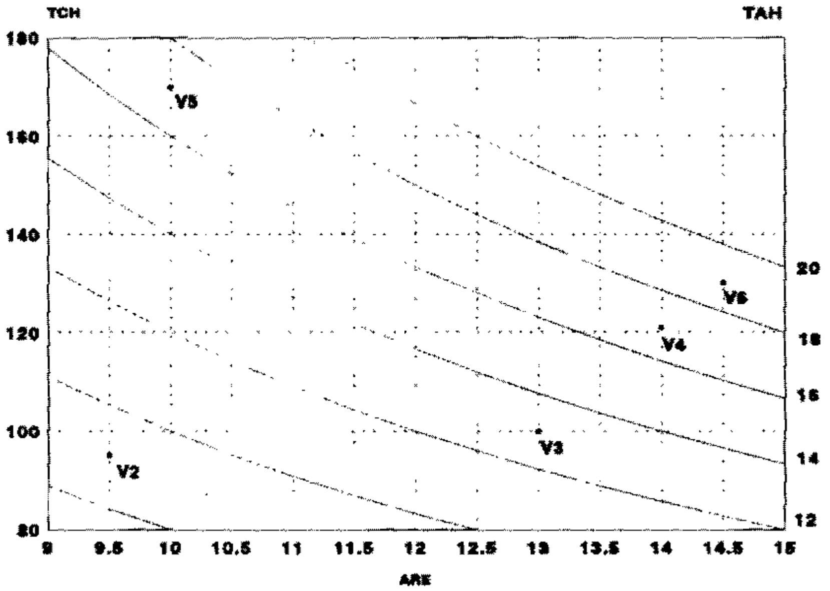


**Figura 3.** Curvas de Isoproductividad por hectárea cosecha.

La **figura 3** nos permite detectar rápidamente las producciones de TAH de diferentes variedades y además cuál es el origen de las TAH, es decir, si es el resultado de altos valores de ARE o de altos valores de TCH o viceversa. Por ejemplo, se puede ver que las variedades V5 y V4 tienen el mismo valor de TAH, pero V5 es el resultado de altos TCH y bajos valores de ARE y V4 es producto de altos valores de ARE y TCH menores. Como se puede observar un resultado bueno o malo de TAH depende de la forma cómo se halla obtenido. Lo cual está indicando que esta variable por si sola, no puede ser utilizada como criterio para decidir si una variedad, zona o tecnología es mejor que otra.

Como observamos en la **figura 3**, las variedades V4 y V5 tienen la misma producción de TAH, pero la variedad V4 incurre en menos costos que la variedad V5, debido a una menor biomasa y esto implica menores costos de corte, alce, transporte, molienda y pago de caña como insumo en la elaboración de azúcar.

### 1.2.2 INGRESO NETO POR TONELADA DE AZUCAR

Se consideró el ingreso neto por tonelada de azúcar, que combina TCH, ARE y los costos asociados para lograr producir la tonelada de azúcar, como la variable que permite hacer comparaciones más reales entre tecnologías.

Esta relación se expresa de la siguiente forma:

Ingreso neto = valor de la producción - costos de producción[3].

De una manera más explícita tenemos:

$$I=V-C1-\frac{C2+\frac{C3}{TCH}}{ARE} \qquad (2)$$

Donde,

| | | |
|---|---|---|
| I | : | Ingreso neto por tonelada de azúcar |
| V | : | Valor de la producción |
| C1 | : | Costos de fabricación |
| C2 | : | Costos de molienda, corte, alce y transporte |
| C3 | : | Costos de campo más costos de la tierra |

Con información de uno de los ingenios, se fijaron rangos de ARE y se determinó el ingreso y el TCH promedio para cada rango, y con esto se pudo observar **(figura 4)** que el ingreso neto por tonelada de azúcar se incrementa hasta alcanzar un valor máximo de $190392, cuando el ARE es de 13.2 y el TCH es de 139.3. Este procedimiento plantea, lo que ocurre con el ingreso neto y el TCH cuando

[3] COCK, J; LUNA, C.A; PALMA, A.E; 1992. "Tonelaje o rendimiento".

se incrementa el ARE. El procedimiento siguiente fue observar lo que ocurría con el ingreso neto por tonelada de azúcar y el ARE si se incrementaba el TCH; el resultado se puede apreciar en la **figura 5**, donde el máximo de ingreso neto es de $184469 y lo alcanza para un valor de 11.2 de ARE y 175 de TCH. Pero este máximo de ingreso neto vía TCH se puede alcanzar también vía ARE, con la combinación: ARE=12.3 y TCH=145 observando la figura 4, que desde el punto de vista práctico es más factible en el mediano plazo, sin necesidad de tratar de obtener el TCH de 175, que es difícil de lograr como promedio y que puede traer repercusiones negativas desde el punto de vista de costos.

Es decir, un objetivo a largo plazo es alcanzar la combinación ARE=13.2 y TCH=139.3; y a mediano plazo el objetivo de ARE=12.3 y TCH=145, logrando de esta manera optimizar los ingresos netos por tonelada de azúcar.

De los comentarios anteriores se puede resaltar en primer lugar, que el ingreso neto por tonelada de azúcar relaciona de una manera más adecuada las variables ARE y TCH dado que considera su efecto conjunto asociado con los costos involucrados en la producción y muestra la combinación más eficiente para obtener azúcar, y un segundo aspecto a considerar, es el hecho de obtener mejores ingresos cuando se incrementa el ARE (conservando unos niveles de TCH), dada la disminución de los costos.

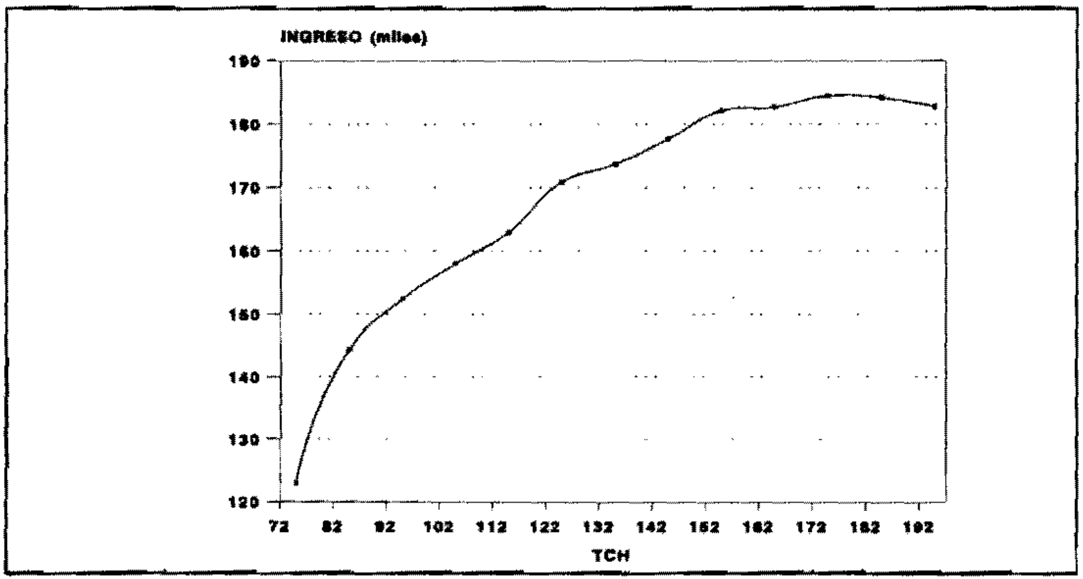


**Figura 4.** Relación entre Ingreso neto por tonelada de azúcar y TCH.

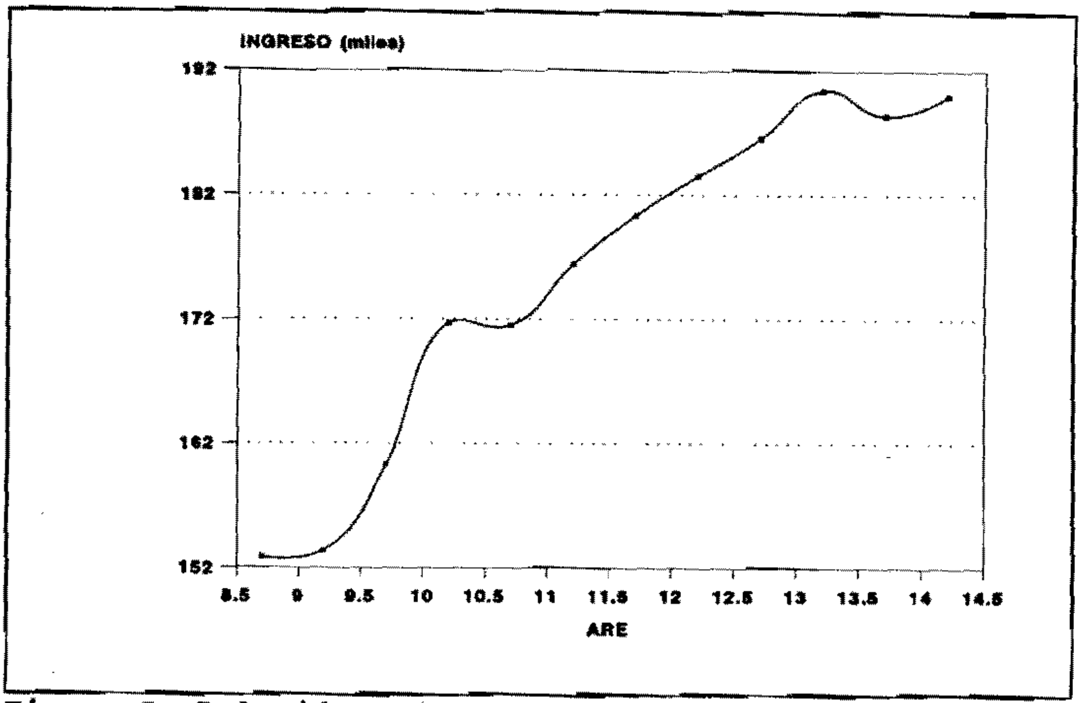


**Figura 5.** Relación entre Ingreso neto por tonelada de azúcar y ARE.

## 1.3. COMPARACIONES

### 1.3.1 COMPARACIONES GENERALES DENTRO DE UN INGENIO

Con información comercial de uno de los ingenios y utilizando la variable I se realizaron comparaciones entre variedades, número de cortes y semestre para evaluar la gestión administrativa de un año y poder hacer ajustes para el periodo siguiente. En este análisis se pueden involucrar otras variables de explicación como zonas, riegos y otros manejos utilizados en el desarrollo del cultivo.

En el cuadro 1 se observa cómo cambia la clasificación cuando se consideran los promedios de ingreso neto respecto a los promedios en TAH y las respectivas combinaciones de TCH y ARE, además el intervalo de confianza para el ingreso neto. Para la variable TAH la variedad 82 y 65 ocupan los primeros lugares, sin embargo para el ingreso neto la variedad 65 no se diferencia mucho de la 45, la variedad 82 tiene un rango muy amplio para el ingreso y las dos últimas variedades del cuadro, se diferencian claramente del resto. Observamos que las TAH de la variedad 45 es menor que las obtenidas por la 65 y la 82, sin embargo debido al valor obtenido en el ARE logra superar a la 82 en el ingreso, acercarse bastante a la 65 y alejarse de la 57 con quien tiene el mismo TAH.

**CUADRO 1.** Promedios por variedad, año de 1993.

| Variedad | TAH | Ingreso Neto | Intervalo de confianza del 95% para el ingreso | TCH | ARE |
|---|---|---|---|---|---|
| 82 | 18.9 | 176706 | 170480 - 182932 | 181.2 | 10.5 |
| 65 | 18.9 | 181241 | 179452 - 183030 | 170.0 | 11.1 |
| 45 | 16.9 | 178827 | 176840 - 180814 | 141.6 | 11.9 |
| 57 | 16.8 | 175103 | 172556 - 177650 | 151.3 | 11.2 |
| 20 | 15.6 | 167389 | 164963 - 169815 | 148.4 | 10.6 |
| 40 | 13.8 | 164379 | 160302 - 168456 | 122.4 | 11.3 |

Al comparar el número de corte se observa en el **cuadro 2**, que las TAH marcan diferencias entre los tres cortes establecidos, por ejemplo entre el 1 y el 2 es de 9.4%, sin embargo para el ingreso neto esta diferencia se disminuyen al 1.6% entre los dos primeros cortes debido al incremento en ARE, a pesar de la disminución en el TCH.

**CUADRO 2.** Promedios para la variable número de corte.

| Corte | TAH | Ingreso Neto | TCH | ARE |
|---|---|---|---|---|
| 1 | 19.2 | 180760 | 177.2 | 10.9 |
| 2 | 17.4 | 177787 | 153.2 | 11.4 |
| 3 | 15.5 | 171635 | 136.1 | 11.4 |

La **figura 6** ilustra la relación variedad*semestre, se observa que el ingreso neto de las variedades depende del semestre. Hay dos casos bien marcados, la variedad 20 y la variedad 40. La primera tiene menor TCH y mayor ARE en el semestre 1, la diferencia en el ingreso es debida al mayor valor del ARE, y la segunda variedad incrementó tanto el TCH como el ARE en el semestre 2.

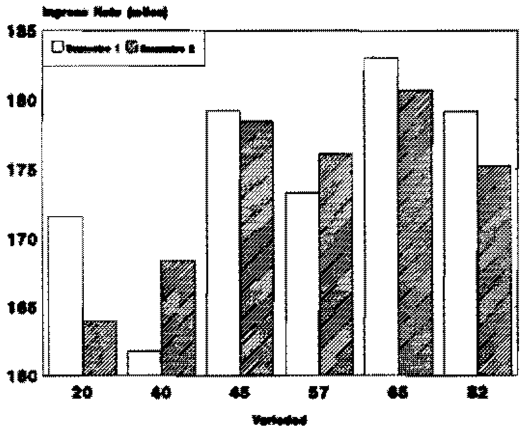


**Figura 6.** Efecto de la relación variedad*semestre en el ingreso neto.

### 1.3.2 COMPARACION CON UN ESTANDAR

En el **cuadro 3** se comparan todas las variedades con la variedad de mayor área sembrada (variedad 45), para saber cuáles difieren o son mejores que este estándar, en las variables TAH e ingreso neto por tonelada de azúcar.

**CUADRO 3.** Comparación con la variedad estándar

| Variedad | Diferencia en TAH | Diferencia en Ingreso Neto | Diferencia en TCH | Diferencia en ARE |
|---|---|---|---|---|
| 82-45 | 2.04 | -2121 | 39.6 | -1.4 |
| 65-45 | 1.96 | 2414 | 28.4 | -0.8 |
| 57-45 | -0.01 | -3724 | 9.7 | -0.7 |
| 20-45 | -1.31 | -11438 | 7.2 | -1.3 |
| 40-45 | -3.11 | -14448 | -19.2 | -0.6 |

Para hacer una representación gráfica de esta comparación se utiliza la relación de ingreso neto por tonelada de azúcar, ecuación (2), y se expresa como[4]:

$$I = I_C + \delta \ (4)$$

Donde,

| | | |
|---|---|---|
| I | = | Ingreso neto promedio por tonelada de azúcar. |
| $I_C$ | = | Ingreso neto promedio por tonelada de azúcar del estándar. |
| δ | = | Delta ingreso neto promedio por tonelada de azúcar. |

Reemplazando (4) en (2) y despejando δ se obtiene:

$$\delta = V - I_C - C1 - \frac{C2 + \frac{C3}{TCH}}{ARE} \ (5)$$

[4] Ibid, pág ...

Esta expresión permite evaluar el impacto que tienen los cambios en parámetros tales como ARE, TCH y costos de producción en el campo, sobre el ingreso neto por tonelada de azúcar producida.

Con la expresión (5) se genera una cuadrícula que incluye una curva estándar de isoingreso neto y una serie de curvas para los delta-ingresos positivos o negativos en intervalos de $10000 (**figura** 7). Sobre esta cuadrícula se pueden ubicar valores de TCH y ARE provenientes de cualquier gestión administrativa de campo o de experimentos y compararlos respecto a un estándar que equivale al delta igual a cero; observando el efecto combinado de TCH y ARE reflejado en delta ingresos netos por tonelada de azúcar.

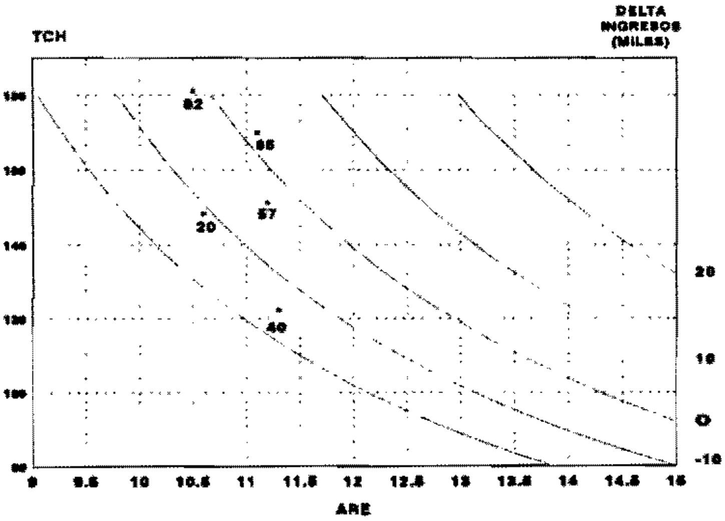


**Figura 7.** Curva de Isoingreso.

## 1.4. EDAD DE COSECHA Y RENTABILIDAD

En las secciones anteriores no se involucra la edad como factor que influye en las decisiones, sin embargo **a priori** se sabe que esta tiene gran influencia sobre la rentabilidad convirtiéndose en un parámetro importante para incluir en el modelo.

Información tanto de datos comerciales como experimentales indican que el TCH y el ARE pueden ser expresados como función de la edad de corte (en meses); TCH=F1(edad) y ARE=F2(edad).

Los costos de campo, C3 en la ecuación (2), incluyendo financiación, dependen de la edad de cosecha. Los costos C3 pueden descomponerse en costos iniciales C31, que incluye preparación de terreno, fertilización, riegos de germinación y control de malezas y C32 que incluye riegos y control de plagas, y los costos finales C33 que incluyen madurantes. Para simplificar el modelo se supone que todos los gastos iniciales ocurren en el momento de la siembra para plantilla y en el momento de la cosecha anterior en el caso de las socas. Para los gastos involucrados durante el ciclo vegetativo se asume que estos ocurren en la mitad del ciclo y los gastos finales (básicamente madurantes) dos meses antes de la cosecha. Así con el interés x por mes:[5]

$$C3 = C31*(1+x)^{EDAD} + C32*(1+x)^{EDAD/2} + C33*(1+x)^{2} \qquad (6)$$

Para el costo de oportunidad de la tierra se utiliza un costo mensual CM y el costo por hectárea CH está dado por:

$$CH = CM*EDAD \qquad (7)$$

Reemplazando (6) y (7) en (5) y expresando TCH y ARE en función de la edad, obtenemos la siguiente expresión:

[5] Cock, J.; Luna C. A.; Palma A.E. 1992. "Edad óptima de corte"

$$\delta = V \cdot I_c - C1 - \frac{C2 + \frac{C31 \cdot (1+X)^{EDAD} + C32 \cdot (1+X)^{EDAD/2} + C33 \cdot (1+X)^{2} + CM \cdot EDAD}{F1(EDAD)}}{F2(EDAD)} \quad (8)$$

Utilizando la ecuación (8) podemos generar un gráfico que nos permite ver el comportamiento de los ingresos netos comparando los resultados obtenidos con una edad de corte establecida, como se puede observar en la **figura 8.**

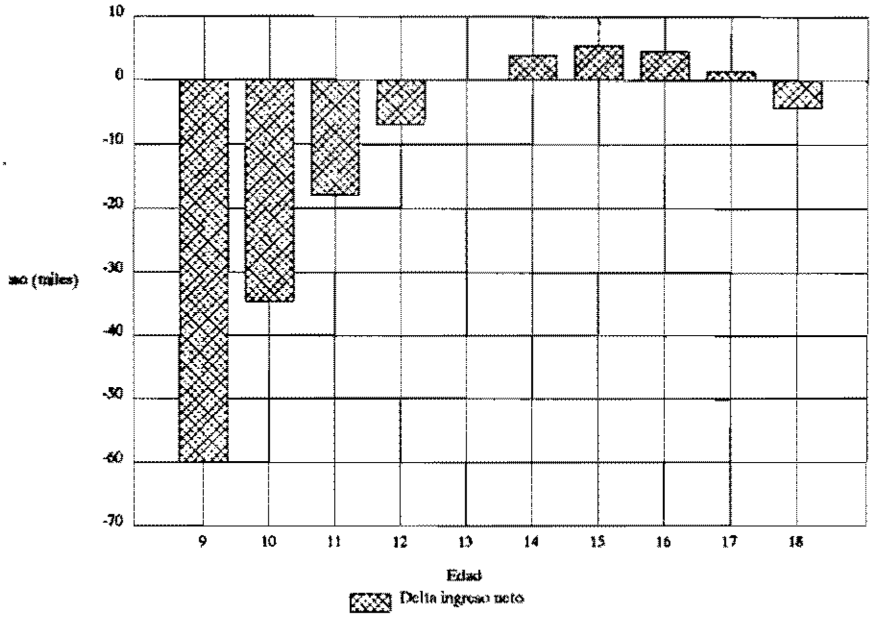


Figura 8. Delta Ingreso neto vs edad de cosecha. Edad de referencia 13 meses.

En la figura 8 se puede observar el resultado de involucrar la edad de cosecha en el modelo. Como ejemplo se considera el efecto de cosechar a los 13 meses de edad; la figura muestra que al cosechar antes de los 13 meses los delta ingresos netos son negativos, pero entre los 13 y los 17 meses estos son positivos obteniéndose un delta ingreso neto óptimo si se cosecha a los 15 meses.

## 1.5. CONCLUSIONES

- Desde el punto de vista metodológico, el de uso de información comercial representativa de los ingenios permite llegar a resultados confiables que orientan la toma de decisiones de cañicultores y retroalimenta el proceso de investigación.

- El análisis del ingreso neto permite detectar que el enfoque del mejoramiento varietal debe ser dirigido hacia la concentración de azúcar recuperable por tonelada de caña (ARE), que optimiza el objetivo del negocio de producir azúcar.

- La variable ingreso neto por tonelada de azúcar es una alternativa que se puede utilizar en el análisis de los resultados experimentales involucrandola en los modelos lineales como variable de respuesta, dada su bondad de conjugar tonelaje, ARE y costos, acercándose de esta manera a las posibilidades comerciales en el momento de las comparaciones de las diferentes tecnologías.

- Esta metodología se puede utilizar en cualquier actividad agropecuaria donde el producto final sea el resultado de dos o más variables de producción.

- Un mayor análisis exploratorio de los resultados de la producción comercial de los ingenios permitirá dar pautas para una mayor eficiencia y también generar alternativas de investigación orientadas hacia el uso más racional de los insumos (agua, fertilización y otras labores) garantizando además la sostenibilidad del cultivo de caña de azúcar.

## 1.6. BIBLIOGRAFIA

COCK, J.; LUNA, C. A.; PALMA, A.E.;1993. "Tonelaje o Rendimiento", CENICAÑA.

COCK, J.; LUNA, C. A.; PALMA, A.E.; 1993. "Edad Optima de Cosecha", CENICAÑA.

MARTINEZ GARZA, A.; 1972. "Diseño y Análisis de Experimentos con Caña de Azúcar", México.

MENDEZ, R.I.; NAMIHIRA D.; MORENO L.; SOSA C.; 1984. "El Protocolo de Investigación: Lineamientos para su Elaboración y Análisis", Trillas, México.

PRAT B., A.; TORT-MARTORELL, X.; GRIMA C., P.; POZUETA F., L.; 1994. "Métodos Estadísticos, Control y Mejora de la Calidad". Barcelona, España.

# 2. CASO 2

# EFECTOS DE LA APLICACION DE AGUA EN EXCESO SOBRE EL ARE Y LA PRODUCCION DE CAÑA. EL CASO DE UN INGENIO AZUCARERO

## 2.1. INTRODUCCION

El siguiente es un caso que presenta cómo el análisis de registros sobre manejo de aguas y fertilizante en uno de los ingenios del sector permite, mediante la utilización de la técnica de los modelos lineales, mostrar la posibilidad de incrementar sustancialmente la eficiencia al disminuir los costos de producción en la práctica de riegos y que trae como consecuencia positiva la preservación del recurso hídrico que desafortunadamente se está agotando.

La teoría de los modelos lineales conjuntamente con la estimación de contrastes de efectos simples de tratamientos, la estimación de promedios por mínimos cuadrados y los métodos de regresión lineal simple, permiten evaluar y caracterizar tendencias de los diferentes factores bajo estudio.

El propósito de este trabajo es cuantificar el efecto de la aplicación de nitrógeno y de agua en exceso sobre el azúcar recuperada (ARE) y la producción de caña de azúcar de las suertes cultivadas a nivel comercial, durante el ciclo de cultivo y valorar en términos económicos la práctica de riego en condiciones de exceso.

Es importante aclarar que la información analizada no atiende a una planeación, ni a un diseño de experimento como tal; por lo contrario, son datos que corresponden a las cantidades reales de insumos aplicados en 873 suertes cosechadas comercialmente en las tierras propias y en participación de un ingenio azucarero del Valle del Cauca durante 1991-1993. Al encontrar tendencias o comportamientos significativos de las variables de producción ante los cambios en la aplicación de nitrógeno y en el agua en exceso, hace pensar que las conclusiones a que se llegan en este análisis son válidas y con seguridad en un estudio planeado sobre respuestas de producción controlando la aplicación de los insumos involucrados conjuntamente, se pueden llegar a encontrar resultados mucho más alarmantes sobre la incidencia de la aplicación de insumos excesivamente.

## 2.2. METODOLOGIA

Con base en las mediciones comerciales del total de agua aplicada en los riegos y la precipitación total se establecen los niveles de exceso de agua que las suertes efectivamente presentaron en el transcurso del ciclo del cultivo[1], definidos como la diferencia entre el total de agua y lo requerido para el desarrollo del cultivo de caña de azúcar en todo el ciclo. Así, el 98,5% de las suertes con las tres principales variedades: CP 57-603, MZC 74-275 y PR 61-632 que equivalen a 691 suertes (11.520 hectáreas) cosechadas en los últimos tres años por este ingenio presentaron aplicación de agua en exceso; excesos que se distribuyen en 8 categorías de 3000 $m^3$ de agua por ciclo, que van desde cero $m^3$ hasta niveles que sobrepasan los 18.000 $m^3$ y es por tal motivo que se considera de fundamental importancia analizar cómo este factor incide sobre las variables de producción y cuantificar su efecto de manera aislada y en forma conjunta con otros factores y con el nitrógeno aplicado.

Se caracterizan las respuestas del tonelaje de caña por hectárea (TCH) y del azúcar recuperable (ARE) debidas a variedades, grupos de exceso de agua, grupos de corte y sus interacciones y a la variable dosis de nitrógeno por hectárea considerada como una covariable, mediante el siguiente modelo de 3 factores.

$$Y_{ijkl} = M + V_i + E_j + VE_{ij} + C_k + VC_{ik} + EC_{jk} + VEC_{ijk} + \beta X_{ijkl} + E_{ijkl}$$

[1] "La precipitación efectiva se define como el 80% del total de precipitación ocurrida en el trascurso del ciclo de cultivo; el cultivo requiere en total 12000 $M^3*ha^{-1}$ de agua por ciclo, donde el agua de riego y la precipitación deben establecer el balance hídrico que permita que la planta obtenga todo el producto potencial; así que, se considera agua en exceso a los valores del total de agua recibida por una hectárea por ciclo que están por encima de 12000 $M^3$": Programa de Agronomía de CENICAÑA.

Donde

| | |
|---|---|
| $Y_{ijkl}$: | Variable de respuesta correspondiente a la ijkl ésima unidad de estudio |
| M: | Efecto de media general |
| $V_i$: | Efecto de la variedad i (i=1,2,3) |
| $E_j$: | Efecto del grupo de exceso j (j=1,2,3,4,5) |
| $C_k$: | Efecto del grupo de corte k (k=1,2) |
| $VE_{ij}$: | Efecto de la interacción de variedad-grupo de exceso |
| $VC_{ik}$: | Efecto de la interacción de variedad-grupo de corte |
| $EC_{jk}$: | Efecto de la interacción de grupo exceso-grupo corte |
| $VEC_{ijk}$: | Efecto de la interacción variedad-grupo exceso-grupo corte |
| $X_{ijkl}$: | Cantidad fija observada de dosis de nitrógeno/ha |
| ß: | Coeficiente de regresión |
| $E_{ijkl}$: | "Error aleatorio" |

De los 8 grupos de exceso de agua antes mencionados se consideraron solamente 5 grupos debido a que ellos concentraron el 93.1% de la información.

El **Cuadro 1** muestra las variables independientes y los niveles de los factores considerados. Se utilizó el programa SAS para el análisis de la información.

**CUADRO 1.** Descripción de variables.

| | |
|---|---|
| Variables Dependientes: | TCH y RDTO |
| Variables Independientes (Factores): | |
| Variedad: | |
| | CP 57-603 |
| | MZC 74-275 |
| | PR 61-632 |
| Grupos de Exceso de Agua: | |
| | 1. (0-3000] |
| | 2. (3000-6000] |
| | 3. (6000-9000] |
| | 4. (9000-12000] |
| | 5. (12000-15000] |
| Grupos de Corte: | |
| | 1. Plantillas |
| | 2. Socas |
| Nitrógeno por Hectárea: | |
| | 79-186Kg * Ha$^{-1}$ |

## 2.3. RESULTADOS

El análisis de varianza para el modelo, mediante la utilización del procedimiento GLM del programa estadístico SAS, y tomando de la suma de cuadrados tipo III, dado que se está en presencia de un cuasiexperimento desbalanceado con 3 factores en el que existe por lo menos un dato en cualquiera de las celdas que conforman la combinación de niveles de los factores, permitió detectar que hay un efecto significativo de la aplicación de nitrógeno por hectárea tanto en la variable TCH como en el ARE; de tal manera que con la opción SOLUTION del procedimiento se pudo estimar que por cada incremento de una unidad de nitrógeno por hectárea se encuentra asociado de manera lineal un incremento promedio de 0.09 toneladas de caña por hectárea y una disminución del ARE de 0.0063 unidades. Vale la pena aclarar que el rango de variación del nitrógeno por hectárea estuvo entre 79 y 186 kg*ha$^{-1}$ y que su efecto es independiente de los otros factores, tal como es la variedad, los grupos de exceso de agua, etc, al tenerse en cuenta que no hubo suficiente evidencia para afirmar lo contrario (No hubo interacciones significativas del nitrógeno por hectárea con los otros factores).

El análisis de varianza también detectó como significativa la interacción variedad-grupo de excesos, y no hubo suficiente evidencia para decir que las interacciones de variedad - grupo de corte y variedad - grupo de exceso - grupo de corte son significativas, esto quiere decir que las variedades responden de diferente manera según la cantidad de agua que se le aplique. En el **Cuadro 2** se muestran los promedios de TCH y ARE asociados a cada grupo de exceso de agua para cada una de las variedades.

**CUADRO 2.** Respuestas de la producción de caña y del azúcar recuperable de las variedades según grupos de exceso de agua recibida.

| Variedad | Grupo de Exceso de Agua ($m^3$) | Promedio de TCH | Promedio de ARE |
|---|---|---|---|
| | **1.** (0 -3000] | 116,3 b[1] | 11,31 a |
| | **2.** (3000-6000] | 119,1 b | 11,52 a |
| **CP 57-603** | **3.** (6000-9000] | 117,9 b | 11,30 a |
| | **4.** (9000-12000] | 124,8 a | 11,24 a |
| | **5.** (12000-15000] | 121,3 a | 11,47 a |
| | **1.** (0-3000] | 139,6 b | 11,35 a |
| | **2.** (3000-6000] | 139,4 b | 11,42 a |
| **MZC 74-275** | **3.** (6000-9000] | 149,2 a | 11,31 a |
| | **4.** (9000-12000] | 146,8 a | 11,24 a |
| | **5.** (12000-15000] | 148,3 a | 11,41 a |
| | **1.** (0-3000] | 151,7 b | 10,90 a |
| | **2.** (3000-6000] | 161,5 a | 10,76 a |
| **PR 61-632** | **3.** (6000-9000] | 162,0 a | 10,62 a |
| | **4.** (9000-12000] | 163,6 a | 10,86 a |
| | **5.** (12000-15000] | 167,0 a | 10,81 a |

[1] Promedios con la misma letra no son significativamente diferentes (P=0.10)

Se observa que para la variedad CP 57-603 se obtienen diferencias significativas en peso en caña para los grupos de exceso de agua por debajo de 9000 m$^3$ al ser comparados con los grupos de exceso que están por encima de este valor. Para la MZC 74-275, estas diferencias se dan siginificativamente en los niveles por debajo de 6000 m$^3$ comparados con los resultados obtenidos en niveles de exceso superiores a éstos; mientras que la PR 61-632 el valor de diferencia está determinado por 3000 m$^3$ de agua en exceso. Las tres variedades responden de diferente manera a la aplicación de agua en exceso en cuanto a la variable TCH, pero sobre la variable ARE no tuvo ningún efecto significativo.

No sólo es importante determinar que existe un efecto significativo de los grupos de exceso de agua sino que es necesario determinar cómo es su comportamiento, con este fin se hizo un análisis de varianza descomponiendo las sumas de cuadrados debidos a los grupos de exceso y su interacción con variedades en componentes lineales, cuadráticos, cúbicos y de grado cuatro, el cual permitió caracterizar una respuesta lineal significativa ante el incremento del agua en exceso para cada variedad.

La estimación de los coeficientes de regresión lineal permite mostrar que por cada incremento de 1000 m$^3$ de agua en exceso que recibe una hectárea con caña de azúcar durante su periodo de cultivo con respecto a lo requerido, la variedad CP 57-603 tiene una productividad marginal de 0,5 toneladas, la MZC 74-275 de 0,83 toneladas y la PR 61-632 de 1,1 toneladas de caña por hectárea.

En los **Gráficos 1, 2 y 3** se muestra el comportamiento para cada una de las variedades, la tendencia en producción de caña por hectárea cosecha se expresa como una función del exceso de agua, medido en 1000 m$^3$ por hectárea durante el ciclo de cultivo.

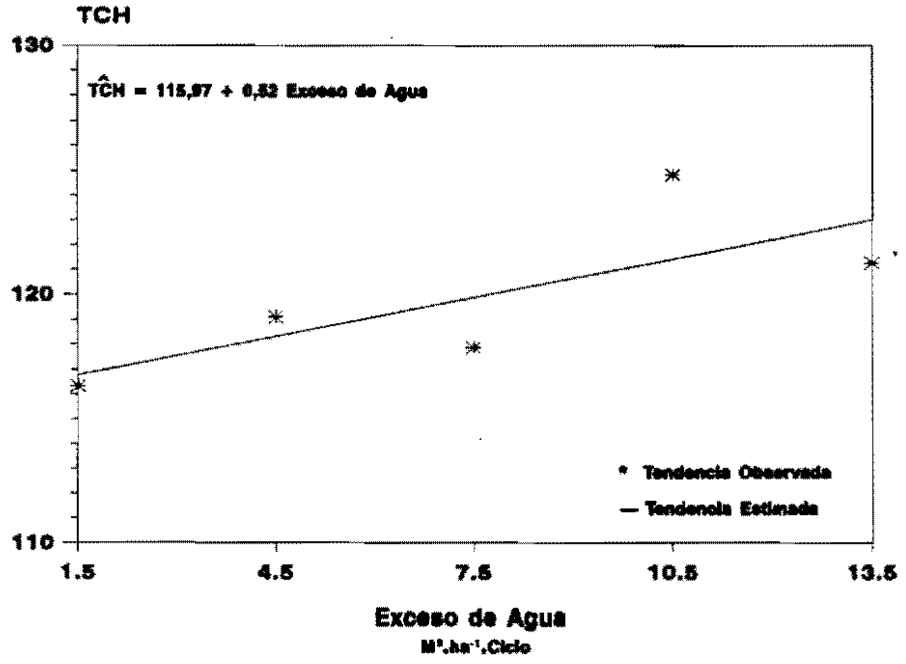


**GRAFICO 1.** Modelo para TCH en función del exceso de agua. Variedad CP 57-603.

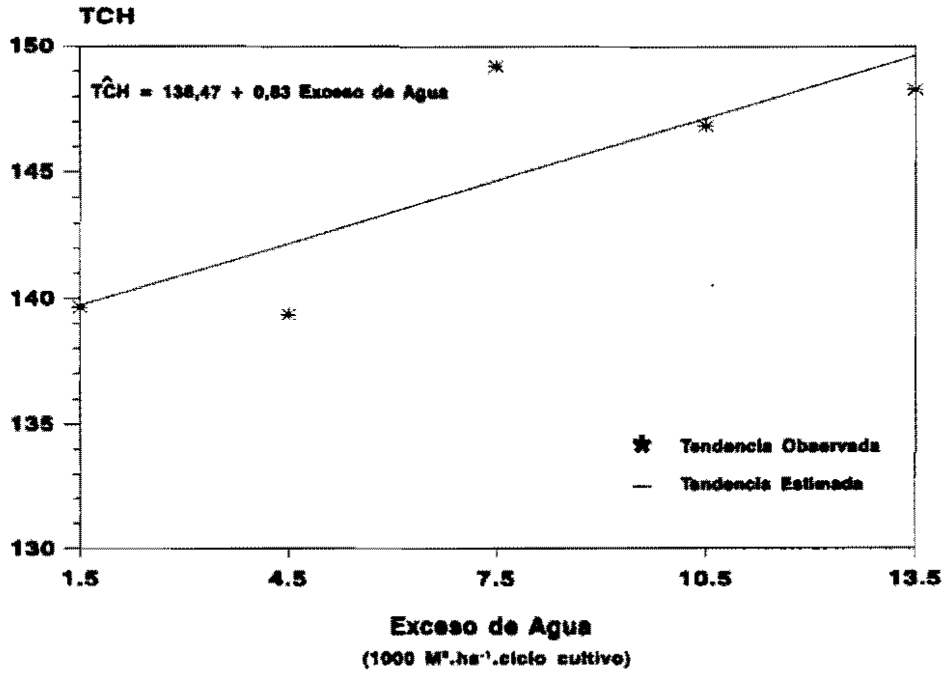


**GRAFICO 2.** Modelo para TCH en función del exceso de agua. Variedad MZC 74-275.

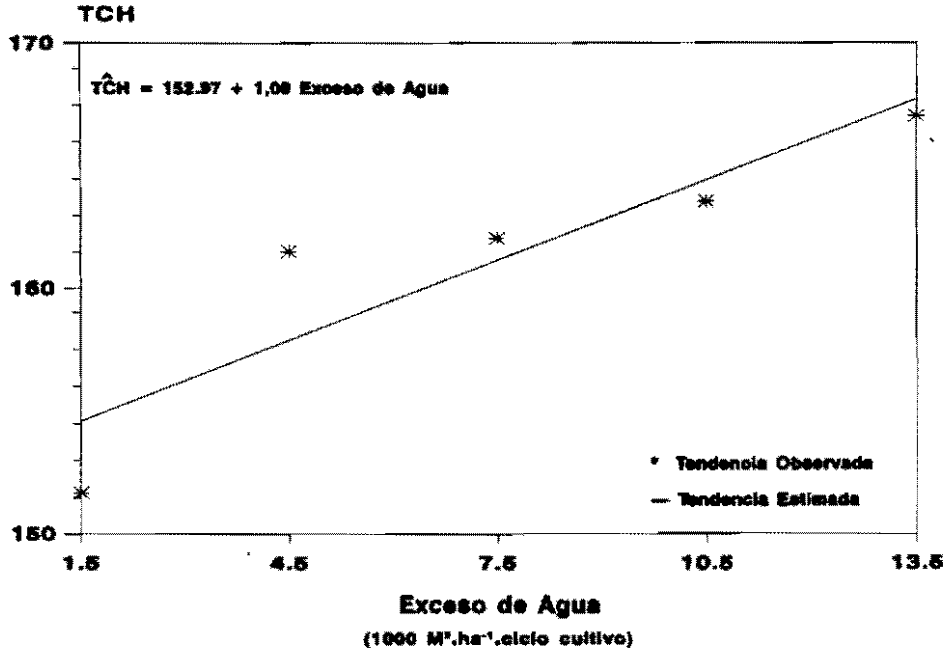


**GRAFICO 3.** Modelo para TCH en función del exceso de agua. Variedad PR 61-632.

Debido a que la edad de cosecha explica en cierto grado los resultados de producción, se calcularon estadísticas descriptivas para la edad de corte asociadas a cada grupo de exceso de agua; con el ánimo de ver si este factor presenta una tendencia con los incrementos del agua en el exceso. Un análisis de varianza detectó diferencias significativas entre variedades; no hubo suficiente evidencia para afirmar que estas diferencias dependen de los grupos de exceso de agua o viceversa, como tampoco hubo significancia para las otras interacciones. En el cuadro 3, se puede observar que la edad de corte es similar para todos los grupos de exceso de las variedades, por lo tanto no existe un efecto confundido de la edad de cortes sobre los incrementos de exceso de agua de las variedades consideradas.

**CUADRO 3.** Promedios de edad de cosecha de las variedades según grupos de exceso de agua.

| Variedad | Grupo de Exceso de Agua ($m^3$) | Edad de Corte (mes) |
|---|---|---|
| | **1.** (0 -3000] | 13,33 |
| | **2.** (3000-6000] | 12,97 |
| **CP 57-603** | **3.** (6000-9000] | 13,07 |
| | **4.** (9000-12000] | 13,35 |
| | **5.** (12000-15000] | 13,60 |
| | **1.** (0-3000] | 13,49 |
| | **2.** (3000-6000] | 13,29 |
| **MZC 74-275** | **3.** (6000-9000] | 13,43 |
| | **4.** (9000-12000] | 13,28 |
| | **5.** (12000-15000] | 13,41 |
| | **1.** (0-3000] | 14,85 |
| | **2.** (3000-6000] | 14,54 |
| **PR 61-632** | **3.** (6000-9000] | 14,73 |
| | **4.** (9000-12000] | 14,51 |
| | **5.** (12000-15000] | 15,59 |

## 2.4. EVALUACION ECONOMICA DE LOS EXCESOS DE AGUA

Acorde con las contribuciones marginales que obtienen cada una de las variedades ante los incrementos del agua en exceso; se valora en términos económicos la producción de caña adicional asociada a cada grupo de exceso y el costo del exceso de agua como tal. En el Cuadro 4 se puede observa que el valor de la producción adicional de caña por hectárea para todas las variedades y en todos niveles de exceso es menor que los costos incurridos por los cañicultores al regar una hectárea por encima de lo requerido; de tal manera que debidos al exceso de agua aplicado, este ingenio durante los tres años analizados presentó a nivel agregado unas pérdidas que ascienden a $1.361 millones de pesos corrientes de 1994, tomando como base para el cálculo el promedio de la cantidad de agua en exceso aplicada por hectárea durante el ciclo de cultivo; donde el promedio general de las pérdidas por hectárea son de $154.250 durante el ciclo de cultivo.

Dado que lo generalizado es encontrar excesos en la aplicación de agua, se toman las suertes que presentaron los menores niveles, con el ánimo de hacer unas cuentas conservadoras de las pérdidas incurridas por hectárea a nivel general y de las variedades; pérdidas que representan los posibles ahorros que se pueden lograr mejorando esta práctica agrícola. El 25% de las suertes menos intensivas en el consumo de agua en exceso tienen en promedio una pérdida $138.800 por hectárea, reduciéndose aproximadamente en 10% con respecto al promedio general y con base en este valor la sumatoria incurrida por el ingenio que no representó ninguna productividad adicional alcanzan un total de $1225 millones de pesos corrientes de 1994 en toda el área cosechada con las tres variedades.

**CUADRO 4.** Valor de la producción adicional de caña por hectárea y costos del exceso de agua aplicado por hectárea.

| Variedad | Grupo de Exceso de Agua | Area Cose-chada (ha) | Valor Producción de Caña adicional ($\$$*ha$_{-1}$) | Costo exceso de Agua ($\$$*ha$^{-1}$) | Pérdidas Económicas ($\$$*ha$^{-1}$) |
|---|---|---|---|---|---|
| **CP 57-603** | **1.** (0 -3000] | 365 | 16.470 | 76.540 | (60.070) |
| | **2.** (3000-6000] | 1.393 | 41.545 | 175.060 | (133.515) |
| | **3.** (6000-9000] | 1.424 | 67.250 | 274.170 | (206.920) |
| | **4.** (9000-12000] | 473 | 91.500 | 336.080 | (244.580) |
| | **5.** (12000-15000] | 151 | 123.950 | 438.900 | (314.950) |
| **MZC 74-275** | **1.** (0-3000] | 247 | 25.660 | 70.330 | (44.670) |
| | **2.** (3000-6000] | 1.000 | 66.870 | 171.300 | (104.430) |
| | **3.** (6000-9000] | 1.351 | 107.400 | 258.900 | (151.500) |
| | **4.** (9000-12000] | 684 | 149.600 | 339.900 | (190.300) |
| | **5.** (12000-15000] | 249 | 190.730 | 423.700 | (232.970) |
| **PR 61-632** | **1.** (0-3000] | 235 | 42.230 | 95.520 | (53.290) |
| | **2.** (3000-6000] | 225 | 85.550 | 182.750 | (97.200) |
| | **3.** (6000-9000] | 543 | 142.530 | 276.630 | (134.100) |
| | **4.** (9000-12000] | 427 | 200.160 | 363.960 | (163.800) |
| | **5.** (12000-15000] | 58 | 261.400 | 432.870 | (171.470) |

Se observa que las variedades se concentran en una mayor proporción en los niveles de exceso de agua que van de 3000 m$^3$ a los 12000 m$^3$, siendo la CP 57-603 la que agrupa un poco más de la tercera parte en un sólo grupo de exceso.

## 2.4. CONCLUSIONES

- Análisis estadístico de este tipo de información permite detectar que labores o prácticas agrícolas generalizadas no contribuyen a incrementar las productividades de los cultivos en caña de azúcar.

- Ningún exceso de agua se justifica económicamente.

- La utilización de la teoría de modelos lineales permite cuantificar los efectos de factores ausales y sirven de referencia para retroalimentar y orientar el proceso investigativo.

# PARTICIPATORY ON-FARM RESEARCH: STATISTICAL LITERATURE AND FUTURE POSSIBILITIES

by

Janet Riley

IACR - Rothamsted
Harpenden
AL5 2JQ
UK

A presentation prepared for the International Symposium on 'Statistics in Agriculture and Environmental Research' - Satellite Conference, Centro Internacional de Agricultura Tropical, Cali, Colombia. June 7-10, 1995.

## *Abstract*

In response to requests from agricultural research workers in developing countries, the author initiated a project to establish the current availability of statistical literature for participatory on-farm research. A description is given in this talk of the range of material available and the different applications for which it is suitable. Indications are given of areas where statistical development is necessary to provide a comprehensive set of statistical material for both general and specific applications.

IACR receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom.

# STATISTICAL METHODS

# IN

# MOLECULAR BIOLOGY

# STATISTICAL MODELS FOR THE DETECTION OF GENES CONTROLLING QUANTITATIVE TRAIT LOCI EXPRESSION

**Emilio A. Carbonell**
**IVIA, Apartado Oficial, 46113 Moncada (Valencia), Spain**

Many important traits in plant breeding exhibit continuous variation (yield, maturity, biotic and abiotic stress tolerance, etc.). The genetic principles underlying their inheritance are basically the same as those affecting Mendelian or qualitative traits, but since the segregation of the genes concerned could be followed individually, new methods and concepts had to be developed. In this presentation, the task of searching for genes affecting quantitative, or continuous, traits will be considered. The immediate hope is the possibility of identifying specific portions of the genome involved in the variation of these traits (called QTLs) in order to enhance breeding programs. Moreover, the long term hope is finding the location of these genes to characterize and manipulate them to our advantage.

Therefore, we will try to discuss the different statistical models used for QTL analysis and the power of different approaches. QTL analysis have been approached using different statistical strategies depending on the number of markers involved in the analysis. Early studies considered the relationship between a marker and a QTL; later, models considered a pair of markers flanking the QTL and studied the association between the QTL and the interval defined by the flanking markers. More recently, statistical methodologies focus their attention to consider the whole linkage group considering all markers of the group as being associated with the QTL. Although any type of classification of statistical approaches is always incomplete and biased, we will try to describe the current statistical models organized according to the number of markers studied and by the basic statistical methodology being employed.

The practical implications of the approach will be discussed presenting a study on genotype by environment interactions for 15 traits in almonds investigated during 2-3 years by means of isozymatic markers.

# MODELOS ESTADISTICOS PARA LA DETECCION DE GENES RESPONSABLES DE LA VARIACION DE CARACTERES CUANTITATIVOS

**Emilio A. Carbonell**
**IVIA, Apartado Oficial, 46113 Moncada (Valencia), Spain**

Muchos caracteres de interés agronómico muestran variación continua (rendimiento, precocidad, tolerancia a stress biótico y abiótico, etc.). Los principios genéticos que regulan su herencia son, básicamente, los mismos que afectan a los caracteres Mendelianos o cualitativos; sin embrago, dado que no se podía efectuar de forma individual la segregación de los genes responsables de su variación, ha sido necesario desarrollar nuevos métodos y conceptos. En esta conferencia, consideraremos el tema de la búsqueda de genes que afectan a los caracteres cuantitativos. El objetivo a corto plazo es la posibilidad de identificar regiones específicas del genoma involucradas en la variación de estos caracteres (o QTLs) para ser utilizados en programas de mejoramiento genético. A largo plazo, se espera llegar a localizar más finamente estos genes para caracterizarlos y manejarlos en nuestro beneficio.

Discutiremos los diferentes modelos estadísticos empleados en el análisis de QTLs y la potencia de los diferentes enfoques. Los análisis de QTLs se han efectuado siguiendo diferentes estrategias según el número de marcadores que se estudia simultáneamente. Los primeros estudios implicaban a un único marcador; posteriormente, se consideraron parejas de marcadores y se estudiaba la presencia de un QTL dentro del intervalo que definen. Más recientemente, los métodos estadísticos centran su atención en considerar todo el grupo de ligamiento, e incluso todo el genoma, y estudian las asociación de todos los marcadores simultáneamente. Aunque cualquier tipo de clasificación es incompleta y sesgada, describiremos los métodos existentes de acuerdo a número de marcadores que estudian y la técnica estadísitca que emplean.

Las implicaciones prácticas del enfoque del análisis de QTLs se discutirá presentando un estudio mediante marcadores isoenzimaticos sobre la interacción genotipo-medio en 15 caracteres de almendro evaluados durante 2-3 años

# STATISTICAL MODELS FOR THE DETECTION OF GENES CONTROLLING QUANTITATIVE TRAIT LOCI EXPRESSION

**Emilio A. Carbonell**

**IVIA, Apartado Oficial, 46113 Moncada (Valencia), Spain**

## 1. INTRODUCTION AND BASIC CONCEPTS

The set of hereditary material transmitted from parents to offspring is known as the genome, and consists of molecules of DNA arranged in the chromosomes. The DNA itself is characterized by its nucleotid sequence. DNA sequences therefore have lengths measured in base pairs. A physical map is an ordering of features of interest along the chromosome in which the metric is the number of base pairs between features. However, we are concerned with genetic mapping where the metric itself is under genetic control.

Genetic map distances depend on the level of recombination expected between two points. An individual receives one copy of each heritable unit from each parent, but the combination of units at different locations the individual transmits to the next generation need not be one of the parental sets. Recombination may have taken place during the process of meiosis, and recombination between two elements on the same chromosome is more likely the further apart are those elements, with a limiting value of 50%. Although there is generally a monotonic relation between physical and recombinational distances, it is not a simple one. The distance over which one recombinational event is expected to occur depends on the region of the genome.

Detailed maps have been developed in model species like *Drosophila* many years ago and more recently in crop plants. These maps consisting of identifiable features on the genome at known locations or *markers*, can be used in the search for genes affecting traits of interest of discrete nature like human diseases. These traits are the easiest to use because there is little ambiguity over which individuals had the disease.

Many important traits in plant breeding exhibit continuous variation (yield, maturity, biotic and abiotic stress tolerance, etc.) and therefore are named metric, polygenic, or quantitative traits (QTs). The genetic principles underlying their inheritance are basically the same as those affecting Mendelian or qualitative traits, but since the segregation of the genes concerned cannot be followed individually, new methods and concepts had to be developed. Why the segregation of genes involved in the variation of a QT in a population cannot be followed individually?. It is mainly due to two common well accepted aspects of the genetic control of QTs: a), "many" loci are involved, and b), the effect of the individual genes is "very small" compared to the environmental effects. Consequently, a branch of the science of Genetics concerned with quantitative traits grew up, the Biometrical or Quantitative Genetics.

In this discussion, the much more difficult task of searching for genes affecting quantitative, or continuous, traits will be considered. The fact that these traits are controlled by more than one gene, as well as by other non-genetic causes like the environment, further complicated the search. The immediate hope is that the possibility of identifying specific portions of the genome involved in the variation of this QTs in order to enhance breeding programs. Moreover, the long term hope is finding the location of these genes to characterize and manipulate them to our advantage. This basic problem is a very old one, In fact, Sax (1923) introduced for the first time the concept of using qualitative genes to locate genes of lesser effects controlling QTs. Thoday (1961) used single gene morphological markers to conduct detailed studies of QTs in *Drosophila melanogaster*. More systematic attempts to resolve QTs into their individual genetic components were initially limited by the scarceness of polymorphic qualitative markers that could cover large parts of the genome. These limitations were partly overcome by the use of isozymes, later by the restriction fragment length polymorphisms (RFLPs) which have provided a major source of markers with many desirable attributes and lately by RAPDs (random amplified polymorphic DNA), microsatellites and other markers at the DNA level. Thus, genetic or molecular markers (isozymatic loci, RFLPs and

RAPDs, etc.) linked to loci affecting a quantitative trait of interest can be used to follow their individual segregation and perform a kind of indirect selection named "marker assisted selection" to improve these traits more efficiently. The availability of detailed linkage maps of molecular markers makes it possible to dissect QTs into discrete factors, called QTLs (Quantitative Trait Loci) by Gelderman (1975); and, when a high enough number of markers are scored in a family segregating for a given polygenic trait, accurate estimates of genic effects and QTL locations can be obtained.

Molecular markers have been used to study many QTs in several species. Thus, great efforts have recently been focused on the construction of saturated linkage maps based on molecular markers to locate genes and QTLs affecting important traits by means of specific statistical methods. Four types of experimental designs are generally used in these studies: $F_2$'s, backcrosses, recombinant inbred lines and doubled haploid lines. The first two designs are the most frequently used for gene mapping mainly due to the less time involved in annual crops, but the last two allow unlimited replications. Some annual crops have been extensively studied (tomato, corn, etc.) others, like perennial crops or forest trees due to their inherent complexities like long juvenile periods have been much less studied.

Therefore, in this presentation we will try to explain and discuss the different statistical models seed for QTL analysis, the power of different approaches and some aspects of specifical applications in almonds.

## 2. STATISTICAL METHODOLOGIES

Genetic mapping of QTLs rests on the simple idea that genetic markers that tend to be transmitted together with specific values of the trait are likely to be close to a gene affecting that trait. In other words, an association is sought between marker variants and trait values, with higher levels of association suggesting closer genetic map distances.

QTL analysis have been approached using different strategies depending on the number of markers involved in the analysis. Early studies considered the relationship between a marker and a QTL; later, models considered a pair of markers flanking the QTL and studied the association between the QTL and the interval defined by the flanking markers. More recently, statistical methodologies focus their attention to consider the whole linkage group considering all markers of the group as being associated with the QTL.

Although any type of classification of statistical approaches is always incomplete and biased, we will try to describe the current statistical models organized according to the number of markers studied and by the basic statistical methodology being employed.

## 2. 1. A marker at a time.

### 2. 1. 1. Linear models: Comparison of means and ANOVA

Traditional methods to estimate the association between a marker locus and a quantitative trait are based on a very simple and appealing approach: individuals are classified according to the genotype of the marker; when the means for the quantitative trait for those groups are not statistically different it indicates that the classification into groups was somehow arbitrary as far as the quantitative trait is concerned; therefore, the marker and the QTL are independent (i. e., no QTL is linked with the marker). Conversely, if means are statistically different, it is an indirect proof of the association between the grouping structure (the genetic marker) and the quantitative trait. So, a group of statistical methodologies are based on the comparison of means or ANOVA approach. Formally, the theory behind this intuitive approach is as follows:

Let us consider a population derived from a cross between two inbred lines. Denoting by M the marker locus and Q the locus involved in the expression of a quantitative trait. Consider two codominant alleles for the marker locus and that both loci are linked, r being their recombination fraction (2r would be the probability of crossing over during meiosis) independent of the sex of the parents from which the gametes were produced. Defining *a* and *d* after Falconer (1989) as the genotypic value of the homozygote and heterozygote genotypes of the Q locus, the phenotypic value of an individual i could be expressed for a given QTL by the following linear model:

$$Y_i = \mu + aX_i + d\,(1-X_i^2) + \epsilon_i \qquad i = 1, \ldots, n \qquad (1)$$

where X is a dummy variable taking values of +1 (for the dominant homozygote), 0 (for the heterozygote) or -1 (for the recessive homozygote) in such a way that the expected genotypic values are:

$$\begin{array}{llll} \mu_{Q_1Q_1} = \mu + a & (X = +1) & for & Q_1Q_1 \\ \mu_{Q_1Q_2} = \mu + d & (X = 0) & for & Q_1Q_2 \\ \mu_{Q_2Q_2} = \mu - a & (X = -1) & for & Q_2Q_2 \end{array} \qquad (2)$$

and $\epsilon_i$ is random variable that is assumed to be normally distributed with mean 0 and variance $\sigma^2$ and includes, among other things, the effect of other QTLs affecting the trait. This general model can be particularized for other mating designs.

Other authors use and alternative but equivalent model in terms of mean, additive and dominance effects:

$$\mu_{Q_iQ_j} = \mu + a_i + a_j + d_{ij} \qquad (3)$$

$F_2$ and backcross populations have been extensively studied by many authors. Most of them used the ANOVA approach and presented the frequencies of the QTL genotypes (Table 1), and the expected values of the means and variances (Table 2) of the quantitative trait within each marker genotypic class. If both loci segregate independently (r = 0.5), the means of the quantitative trait for each genotypic marker class are both equal; therefore, a way to test if r = 0.5 is by testing the null hypothesis of equal class means by a t-test using non-pooled variances because they are different due to dominance and/or linkage between the marker locus and the QTL; hence, standard

ANOVA procedures should be avoided in this case. Variances will be equal in the case of no association (which is precisely what we want to test) or under no dominance of the trait. Moreover, even though the t-test is quite powerful for departures from non normality, it should be recognized from the structure of table 2 that within each marker class we have a mixture of normals where proportions are functions of the recombiantion fraction.

For the $F_2$ population, the contrasts of interest to estimate the genetic effects are:

$$\mu_{M_1M_1} - \mu_{M_2M_2} = 2\,a\,(1 - 2\,r)$$
$$2\mu_{M_1M_2} - (\,\mu_{M_1M_1} + \mu_{M_2M_2}\,) = 2\,d\,(\,1 - 2\,r\,)^2 \qquad (4)$$

Therefore, the estimates of the genetic effects are biassed due to the confounding effect of the recombiantion fraction. Since the quantity in parenthesis is always equal or lower than one, the bias in the estimate of the dominant effect is larger than that of the additive effect. Unless the QTL shows a complete dominance (a = d), there is no way to distinguish between a QTL tightly linked having a small genetic effect from another with a large effect but loosely linked. Hence, this type of approach could be considered as preliminary or exploratory to simply detect the presence of a QTL. Therefore, to solve the mixture problem and to obtain unbiased estimates of both the magnitude of the gene effects and the recombination fraction are needed, alternative statistical methods should be used.

### 2. 1. 2. Maximum likelihood estimation

To have a separate estimate of the recombination fraction and the biometrical parameters of the QTL, alternative methods have been developed. Maximum likelihood has been shown to be more powerful than moments method for certain problems of estimation; however, it requires that the type of distribution be known. Many studies apply the central limit theorem and assume an underlying normal distribution of the trait unless there is reason to assume otherwise.

For a $F_2$ population, the method is as follows: Let $Y_i$, $Y_j$ and $Y_k$ denote trait values for three random individuals belonging to the maker genotypic classes $M_1M_1$, $M_1M_2$ and $M_2M_2$, respectively. The likelihood of the entire population can be written as:

$$L = \prod_{i=1}^{n_1} f(Y_i) \prod_{j=1}^{n_2} f(Y_j) \prod_{K=1}^{N_3} f(Y_K) \qquad (5)$$

where L is the likelihood function; $f(Y_i)$ is the density function of $Y_i$; and $n_i$ is the total number of individuals in the ith marker class. In turn, the density functions are:

$$\begin{aligned} f(Y_i) &= (1-r)^2 f(Q_1Q_1) + 2r(1-r) f(Q_1Q_2) + r^2 f(Q_2Q_2) \\ f(Y_j) &= r(1-r) f(Q_1Q_1) + (1-2r+2r^2) f(Q_1Q_2) + r(1-r) f(Q_2Q_2) \\ (Y_k) &= (1-r)^2 f(Q_2Q_2) + 2r(1-r) f(Q_1Q_2) + r^2 f(Q_1Q_1) \end{aligned} \qquad (6)$$

That is, each density function is expressed as the sum of three disjoint events: since the linkage between marker and QTL is incomplete, the individuals within each marker class are a mixture of the three genotypes at the QTL. In turn, the probability of each event is a set of conditional probabilities: the product of the probability of having a particular genotype at the Q locus given the genotype of the marker (calculated from table 1), times the density of $Y_i$, conditional on the fact that the individual has that genotype at the QTL (f(QQ)). In order to obtain the form of the latter density, the underlying distribution of the genotypes at the QTL in the entire population (not within a marker genotypic class) is assumed to be normal. The variances include not only environmental variation but also genetic variation at other loci (QTL) affecting the quantitative trait.

For, say, $f(Q_1Q_1)$ the conditional density will have the form:

$$f(Q_1Q_1) = (2\pi\sigma_{Q_1Q_1})^{-1/2} \exp[-\frac{(Y_i - \mu_{Q_1Q_1})^2}{2\sigma^2_{Q_1Q_1}}]$$

(7)

In order to obtain the estimates of the recombination fraction and the means and variances, partial differentials of the log of the likelihood function L for all seven parameters are set to zero and solving the resultant equations. If an analytical solution is not readily evident, iterative solutions are available, but they tend to be cumbersome and time consuming and not necessarily provide solutions.

The problem with the approach is the large amount of computational resources that are required. Furthermore, because of the need to search the likelihood surface for each parameter separately and ignorance of the internal mathematical relationships among the parameters to be estimated, this method could not guarantee that the estimates obtained were in fact maximum likelihood estimates.

### 2. 2. Pair of markers

The basic idea about using a pair of markers flanking the putative QTL was put forward by Lander & Botsein (1986; 1989). Once a map of markers is available, we test whether a QTL lies in an interval of known size between two adjacent markers.

If the putative QTL is located within the interval defined by the two marker locus A and B with known recombination fraction between them equal to p, we denote by r and r' the recombination fractions to the left and right markers. If the distance between the two markers is very small, one may assume that the frequency of double cross-over is negligible; then p = r + r'. If not, p = r + r' -2rr'δ, where δ is the coefficient of coincidence. Under no interference, δ = 1, hence using the Haldane function p = r + r' - 2rr'

This approach combines information on all markers into one likelihood map. The integration of all markers makes this method very appealing, giving detailed insight into the part of the genome being investigated. The testing for the presence of a QTL (null hypothesis of $H_0{:}a{=}0$ corresponding to the hypothesis that no QTL is linked), is given by the significance of the linear model as shown in equation (1) using the Fisher-Snedecor F statistic.

When the QTL is located exactly at one of the two markers defining the interval, the problem is easily solved by standard regression procedures because in this case, the genotype of the QTL is the same as the marker; so, the values of $X_i$ in equation (1) are known. However, if the QTL is located somewhere between the flanking markers (which is the normal situation), these values cannot be directly obtained but only their probability distributions. In this case, the likelihood of

the parameters $\theta{=}(\mu,a,\sigma^2)$ based on the observations $Y{=}(Y_1,Y_2,\ldots\ldots.Y_n)$ is given for a $F_2$ population by:

$$L\,(\theta|Y) = \prod_{i}^{n}\,[G_i\,(-\,1\,)\,z\,(y_i|\,-\,1\,;\theta) + G_i\,(0)\,z\,(y_i|0;\theta) + G_i\,(+\,1)\,z\,(y_i|\,+\,1\,;\theta)]$$

(8)

Where $G_i(X)$ represents the probability distributions of X over the values of -1, 0, and +1 given the genotypes of the flanking markers for individual i. This probability is a function of p, r and r'; and, given that p is known and that r' can be expressed in terms of p and r assuming the Haldane function (p = r + r' - 2rr'), this probability has only r as unknown. Therefore, the recombination fraction between the QTL and the left marker of the interval, r, is included in the likelihood through $G_i(X)$ . Hence, for each value of r, a different likelihood will be obtained. Values of the probabilities for all combinations of genotypes of the markers are shown in Table 3. This probability will be calculated for each individual i according to the specific genotype of its markers. The function

$$z\,(y_i|x_i;\theta) = (2\pi\sigma^2)^{-1/2}\exp[-\frac{(y_i - \mu - a\,x_i - d\,(1 - x_i^2))^2}{2\sigma^2}] \qquad (9)$$

denotes the conditional density of $Y_i$ given $X_i = x_i$ (-1, 0, or +1) assuming normal distribution of the trait. In order to obtain the maximum likelihood estimates of θ, the function $L(\theta|r)$ may be maximized through choices of θ. This may be accomplished by using the EM algorithm (Dempster et al., 1977) for a fixed r.

The E and M steps should be repeated for convergence of the estimates of θ. Once the parameters are obtained, the most likely position may be estimated by the maximum values of the F statistic, through all possible values of r varying along the intervals. The F values could be plotted against r; then, the most likely position of the QTL is given by the value of r, say $r_M$, which gives the maximum of the curve provided it is higher than a given significant threshold. Probability statements when testing the hypotheses that certain effects are zero (such as a = 0 and/or d = 0) can be established.

The effect of two linked QTLs on the estimation of the recombination fraction and the genetic effects using the above model was studied with simulated data. Results from 100 simulations consistently gave bimodal distributions for the higher heritability of the trait of 0.50 as shown in Figure 1 (curve A for doubled haploids and curve B for backcrosses). For doubled haploids and heritability of 0.08, few cases showed a highly significant unimodal distribution (curve C) like the "ghost" QTL observed by Martinez & Curnow (1992), or a multimodal or plateaued curve (curve D) indicating that QTLs could be found anywhere within a long range of values of the recombination fraction with no clear location of the maximum of the likelihood function. More frequently, a bimodal curve (curve E), like in the high heritability case, indicated the presence of two QTLs; however, the predicted locations were biased in the sense that the leftmost QTL was shifted to the right and the leftmost one to the left (see curve F for the comparison of predictions).

Therefore, implicit in the above methods is the fact that a single QTL is investigated at a time and that it segregates independently from the remaining QTLs contributing to the quantitative trait. The full model is actually a summation of the $a_j$ and $d_j$ effects over the total number of QTLs (j = 1, 2,....q) so that when a particular QTL is investigated, it is assumed that its effect is not influenced by the rest of the QTLs. Therefore, it is not intended to resolve clusters of QTLs but only those that are not too close to make tenable the assumption of independent action on the measured quantitative trait.

2. 3. All markers simultaneously.

Multiple regression techniques have been suggested to identify multiple QTLs linked to several markers.

The basic procedure consists of assigning a coded value to each marker genotype and use stepwise regression procedures to identify which markers are associated with the variation of the quantitative trait The multiple regression model will contain quadratic terms to account for dominance at the QTLs. The problems with this approach are those of the stepwise selection of variables: several marked segments are usually available in the same chromosome for the analysis of the QTL effects; these segments do not likely segregate independently. Since stepwise regression procedures incorporate the variables into the model in a sequential manner, one by one, the estimates of the effects of those QTLs which are already in the model may be biased by other linked and unlinked QTL which may or may not be yet in the model. Some QTLs already in the model may prevent the entrance of a new QTL in the model; therefore, some QTL may cover the presence of another QTL that will remain undetected by the statistical methodology. Moreover, the method may produce very different results depending on the path of selection followed by the computer programs; i. e., if two variables have very similar F values to be in or out of the equation and are highly correlated, the final equation may depend on which of the two variables is chosen to enter in the equation.

Considering the genome as a whole, Rodolph & Lefort (1993) proposed a method based on the linear model that can be used in a very general genetic context. The purpose is to detect regions displaying QTL effects rather than to detect isolated QTLs and estimate their effects. Their linear model is actually an extension of the basic model in equation (1) summing over the set of m markers, that is:

$$E(Y_i) = \mu + \sum_{j}^{m} \begin{cases} +\alpha_j - \delta_j & \textit{for}\, M_i(j) = M_1 M_1 \\ \quad\quad +\delta_j & \textit{for}\, M_i(j) = M_1 M_2 \\ -\alpha_j - \delta_j & \textit{for}\, M_i(j) = M_2 M_2 \end{cases} \qquad (10) \qquad \text{for an } F_2$$

where $M_i(j)$ represents the value taken by the jth marker (j = 1,...m) on individual i. Parameters $\alpha_j$ and $\beta_j$ are statistical effects associated with the markers and are related to the additive and dominance effects.

The major advantage of the methods based on all markers simultaneously is the possibility of making global tests, recovering the whole or a part of a chromosome. In doing so, the existence of a set of linked QTLs can be detected. Conversely, the main characteristic of the interval mapping methods is their sequential nature. This is the reason why they are not well suited for the detection of closely linked QTLs, but they are more powerful in the case of an isolated QTL. Hence, a compromise should be reached using both approaches. In fact, recently, a joint approach has been proposed that makes the combined use of the conventional interval mapping with regression methods in order to detect multiple QTLs. It was proposed by Jansen (1993) suggesting to fit one QTL at a time in a given interval and simultaneously using some of the markers as cofactors to eliminate the effect of additional QTLs. Similar idea was put forward by Zeng (1993, 1994) and expanded by Jansen and Stam (1994) in a very general method. The method exploits two features, the use of additional parental and $F_1$ data, which fixes the joint QTL effects and the environmental error, and the use of markers as cofactors which reduces the genetic background noise. By using the EM algorithm, missing values of any kind (markers or quantitative traits) are easily estimated making full use of the data.

Most methods are based on segregating populations derived from crosses of inbred lines. For many species this is not practicable, but crosses can be made between outbred lines. Haley et al. (1994) used least-squares methods to regress trait phenotypes onto additive and dominance effects of putative QTLs in marker intervals. The work is for the situation of crosses between outbred lines in which the trait loci are segregating but in which the markers used are fixed for alternative alleles.

## 3. SAMPLE SIZE AND POWER STUDIES

The accuracy of any mapping procedure depends not only on the ability of the statistical method to determine the location and to estimate the genetic effect of the QTLs. Other factors have a very important influence on this accuracy as well: the heritability of the trait, the contribution of each QTL to the total genotypic variance, the number of QTLs, their interactions, their distribution over the genome, their distance to the markers, the statistical distribution of the random non-genetic factors, the type of segregating population studied, its size, the genome size, the number of marker loci employed, etc. Given the large number of factors, authors restrict themselves to partial studies investigating some of these factors at a time.

For the single marker approach, Soller et al. (1976) gave approximate formula to get the number of individuals per marker class required to detect a given difference between the values of the quantitative trait in each marker class. They assumed that the QTL was located exactly on the marker (which is quite unlikely to happen); for $F_2$ they wrongly assumed constant variance of the quantitative trait within the marker classes. According to their calculations they concluded that a few thousand offspring would be enough to detect close linkages involving QTLs contributing about 1% of the total phenotypic variance of the trait in the $F_2$. Relaxing both assumptions Asíns & Carbonell (1988) gave the following formula for the number of individuals per homozygote genotype in a $F_2$ population in order to detect the QTL at the 5% level and a probability of error Type II of 10%:

$$N \geq \frac{10.5\left[r(1-r)+k^2r(1-3r+4r^2-2r^3)-\frac{1}{4}(1+\frac{1}{2}k^2)(1-\frac{1}{h^2})\right]}{(1-2r)^2} \qquad (11)$$

where k is the degree of dominance ( $k=\frac{d}{a}$ ), and $h^2$ is the heritability of the trait. It is clear from the formula that the number of offspring is a function of the unknown recombination fraction; hence the maximum allowed recombination fraction should be stated in order to obtain the sample size. As a practical guide, Cowen (1988) indicated that the smallest additive effect *a* detectable is in the range of 1/2 to 1/12 the size of the corresponding LSD at the 5% for the quantitative trait. He used 100 recombinant inbreds or doubled haploid lines and 6 replicates. With $S_1$ lines, the within-population dominance effects *d* can be also detected. The smallest *d* detectable is 3-4 times the above size.

The comparison between the maximum likelihood approach and the comparison of means seems to be influenced by the poor convergence during maximum likelihood estimation when the effects of the QTL are small and for recombination fractions approximately 0.5.

For the interval mapping approach, Carbonell et al. (1993) used 100 replicated sets of 250 simulated individuals to investigate the power of the interval mapping to detect and estimate the genetic effects of several QTLs. The simulated individuals had eight linkage groups with six markers each separated by 20 cM distance. Six unlinked QTLs were involved in the expression of the quantitative trait, with different gene action under the assumption of dominance in some QTLs or a completely additive model. They concluded that doubled haploid populations could be used with smaller sample sizes because they show much higher power than backcrosses. Moreover, more accurate estimates of the location of the QTL and with less variance were obtained. This result could be expected from the fact that the model uses more spread-out values (+1 and -1) than backcrosses (0 and -1), and that the absolute difference between the means of both homozygous genotypes is larger than that between the heterozygous and the recessive homozygous genotypes. When dominance is present backcrosses not only give biased estimation of the effects, because additive and dominant effects are completely confounded in this mating design, but also some QTL could not be detected due to similar the genotypic values of the homozygote and the heterozygote at those loci. That could be the case for complete dominance when the $F_1$ population is backcrossed with a "high" producing parental line and also when, independent of the direction of the backcross, the recurrent parental line had the "high" allele at some QTL. For doubled haploids, the power of detecting a given QTL is clearly related to its relative contribution to the heritability of the character; i. e., proportional of to the square of its additive value $a_j^2$. The higher its contribution, the higher the probability of being detected. Even QTLs having small effects were identified by the model; the power of the test was about 90% for heritabilities of the QTL as low as 5%. To obtain similar power for backcrosses, the heritability attributable to the individual QTL should be around 14%. For a given type of gene action, it is difficult to compare with $F_2$ populations results (Carbonell et al., 1992) because two different criteria were used for this population but it seems that doubled haploids have a similar power than $F_2$. In fact, similar values of 5% for the heritabilities of the QTL were obtained by van Ooijen (1992) for this type of population. However, if dominance is present, doubled haploids will only detect the additive component of a particular QTL. This fact could be of an extreme importance for QTLs showing overdominant effects or to design the most efficient breeding strategy in order to exploit the non-additive variation hidden in some QTLs. The major technical advantage for doubled diploids, independent of any effect of replication on the required number of offspring, lies in the fact that the lines can be reproduced independently and continuously evaluated with respect to additional quantitative traits and markers with all the information being cumulative (Burr et al., 1988). If the effect of replication is taken into account, replicated progenies can bring about a major reduction in the number of individuals that need to be scored. Reductions are the greatest when heritability of the trait is low, but are much less when heritability is moderate to high under the assumption of codominance at all QTL (Soller & Beckmann, 1990).

Regarding the statistical approach it is interesting to compare the classical "single marker" methodology with the interval mapping. The methodologies based on the study of all markers simultaneously are much more specifically designed for the case of multiple linked QTLs.

Lander & Botsein (1989) concluded that even in the worst case where the QTL is located in the middle of the interval, interval mapping decreases the number of offspring by a factor of (1 - p) as compared with the single marker approach; so for maps having markers at distances of 10, 20, 30 and 40 cM, the savings are about 9%, 16%, 23% and 28%, respectively (assuming the Haldane mapping function). However, Knott & Haley (1992) compared both maximum likelihood procedures for simulated $F_2$ data and found that the use of flanking markers only provides substantially more power for QTL detection when markers are widely spaced and the QTL effects are large; however, interval mapping has the additional advantage of correctly positioning the QTL and estimates its effect on the average. The estimates obtained using single markers are not close to the simulated values with both the distance from the nearest marker and the additive effect overestimated. In this comparison, the authors have ignored the problem of the level of significance of the comparison and did not give the exact distributions of the test statistics they used. Similar results were obtained by Rebai et al. (1994) when analytically compared the power of the maximum likelihood using interval mapping and the ANOVA approach for backcross populations. They showed that the ANOVA test is less powerful for all non null values of the distance between the flanking markers; however, for small intervals (distances from 0 to 30 cM) the difference in power between the two tests do not exceed 5%. For large intervals and quite large QTL effects the advantage of interval mapping is 10% to 30%. For simplicity in the analytical derivations, their approach assumed absence of double crossing over; therefore, the results for large intervals would be approximate and conclusion should be taken with caution. For small effects of the QTL (less than 1% of the phenotypic variance) powers of the tests are close to the significance level and could not be compared.

Single marker methodologies based on ANOVA and comparison of means are much more simple to implement using standard statistical computer packages but they can be only used to the detection of several unlinked QTLs. They can be used as an exploratory phase of the study where we are simply interested in the detection of association. If an estimate of the localization of the QTL and an unbiased estimation of effect is required, interval mapping should be used although previous mapping of the markers is needed. If precise estimation of the recombination fraction between the flanking markers is not available, Knott & Haley (1992) found that the use of an incorrect value for the recombination frequency between the flanking markers, does not affect the ability to map the QTL as long as the order of the markers along the chromosome is correct. On the other hand, when a QTL is located in the distal part of the linkage group not flanked by markers, the maximum of the LOD score is likely to arise in the interval near to the QTL, or even at the position of the marker; that will result in a mislocalization of the QTL and an under-estimation of its effect. Therefore, as Rodolphe & Lefort (1993) conclude, the comparison between results obtained from different statistical methods can give an interesting insight especially on the number of QTLs; if the purpose is a precise dissection of the genetic determinism of the trait under study, this cannot be achieved with only one experiment; the search for QTLs should be an iterative process. The multimarker model enables the experimenter to focus on the interesting segments detected. New experiments must be made, with more markers on these segments and more individuals in order to get more recombinant genotypes between these closely linked markers. These new experiments can be analyzed using different models and methods adapted to specific situations and purposes. Then, simultaneous use of several methods should lead to better efficiency in QTL technology.

## 4. PERSPECTIVES AND FURTHER REMARKS

In the preceding sections, it has been shown that there are still some unsolved matters that deserve further research. They are mainly related to the presence of closely linked and interacting QTLs, the development of efficient ways to detect QTL showing small effects, unveiling QTLs masked by other QTLs having very large effects, the presence of epistatic effects, methods not only directed towards increasing the power of the designs but also to decrease the probability of false positives, better knowledge of the distributional properties of the test statistics being used, models for non-normally distributed traits, models for populations of allogamous plants, etc.

There has been a fast development of techniques in the areas of genetic markers and pollen-derived homozygous plants. Both technologies assisted by specific statistical methods provide powerful means for QTL identification. Carbonell et al. (1993) concluded for interval mapping that doubled haploid populations will allow to conduct experiments with less number of individuals because, similar to $F_2$ populations, they show much higher power than backcrosses and the estimates are more accurate. However, if dominance or overdominance play an important role at some QTLs, this information will not be unveiled using the doubled haploid design. As it has been discussed previously there are still left several situations where location of QTLs may fail or even providing false positives; therefore, further research efforts have to be focused on the statistical methods. Sometimes, it may be a problem not related to the statistical model itself but to the experimental design used. If the purpose is a precise dissection of the genetic determinism of the trait under study, we do not believe it can be achieved with only one experiment or methodology; QTL analysis has to be an iterative process.

In addition to QTL location and estimation of gene effects, important aspects of QTL analysis have been recently studied. Therefore, there is a considerable body of knowledge that should enable us to deeply investigate other well known and important phenomena related to the nature of quantitative variation such as: pleiotropy, gene interactions, heterosis, transgressive segregations, and genotype by environment interaction. Very recently, several authors have focused studies towards these subjects (de Vicente & Tanksley, 1993; Bretó et al., 1994; Asíns et al., 1994) and we will discuss now some of these aspects.

## 5. APPLICATION TO STUDY THE INTERACTION GENOTYPE BY ENVIRONMENT IN ALMONDS (ASÍNS ET AL., 1994)

The genotype by environment interaction (G x E) is the differential genotypic expression across environments. This interaction reduces the association between genotypic and phenotypic values and may cause selections from one environment to perform poorly in another, thereby forcing plant breeders to examine genotypic adaptation. In addition, the sampling problems associated with yearly variation suggest a necessary testing for many crop cycles. To save time, breeders opt to substitute temporal with spatial variation, assuming that testing over a wide geographic range can ensure a parallel degree of temporal buffering capacity in the germplasm.

QTL analysis based on genetic markers could be used as an approach to study G x E, and even more, to gain stability by designing a marker assisted selection scheme that takes into account all the QTLs involved in the trait for the different environments.

The objective of the investigation was to study the effect of years on QTL detection for 15 traits. The plant material consisted of 123 seedlings established in 1896 that had been derived by hand pollination of emasculated flowers of the Spanish almond variety "Ramillete" with pollen of the Italian variety "Tuono". Fifteen traits were evaluated in 1989 and 1990, and six of them also in 1991 according to Table 4.

The association of isozymatic markers and traits recorded as categorical variables was studied by chi-square contingency tables and by one-way ANOVA and t contrast for the remaining variables. These results are shown in Tables 5 and 6.

At least 22 putative QTLs were detected. Only 3 (or 5 at 10% significance) of them behave somehow homogeneously through the years, none of them were associated with traits recorded as categorical variables. Three factors may be involved in this lack of stability: 1), the test statistic used to detect association being dependent on the character definition of variables; 2), the contribution of the QTL to the total genetic variability of the character; and 3), a differential gene expression depending on the year, i. e. due to G x E. The first two facto factors are close related. As it has been shown (Carbonell et al,. 1992), the power to detect a given QTL is related to its contribution to the heritability of the trait. Thus, with respect to moderate heritabilities (Dicenta et al., 1993a, b), if specific the contribution of the QTL is low, it may remain undetected. That could explain the cases like rugosity of the seed, RS, (tables 4 and 5) and percentage of the kernel, PK, (tables 4 and 6). If heritability is low and the statistical test is not very powerful, it is a matter of chance that the association is detected or not. This could have happen in case like the QTL detected for the color of the tegument, CT. Most cases in which there is a lack of stability in QTL detection involve traits whose heritability estimates change drastically from one year to another like yield intensity, YI, and number of double kernels, ND, and duration of flowering, DF, and/or whose correlation coefficients between years are low, like YI (from -0.14 to 0.28), kernel weight, KW, (0.36), and DF (from 0.00 to 0.24). This suggests that there are important differences in the number and/or the relative contributions of the genes controlling the quantitative trait that are dependent on the year (G x E).

It is noteworthy that the variation for three traits related to yield (YI, KW and PK) in 1990 was associated to segregation at Prx$_c$-2. Temperature records showed that the winter of that year (December 1989 to February 1990) was the warmest, with no under 0°C temperatures. Therefore it is not due to resistance to adverse conditions but a matter of genotype by environment interaction playing an important role in the determination of the phenotypic value of the plant.

Most associations found in 1989 were also found in 1990; however, none of the associations found in 1991 were found in the other years. The coldest winter was that from December 1990 to February 1991 with temperatures of -3°C. Similarly, the lowest values of the coefficient of correlation per trait among years were between 1991 and 1989 or 1990 (data not shown). Winter temperature regimes affect flowering times and the duration of flowering. The way these traits become affected must be a change in gene expression: one or several QTLs linked to Got-2 are involved in IF, MF and FF in 1989 and 1990, while other(s) linked to Prx$_c$-2 are specially relevant for those traits in 1991. This difference might be just a difference in the level of gene expression at the QTLs involved in such a way that the final result is a change in their genic effects (losing or gaining importance in their contribution to the genotypic value of the trait).

Given that a common marker was associated with several traits, correlation coefficients among those traits per year were studied. High correlations (about 0.7) were found among IF, MF and FF for the three years; DF and FF for 1989 (0.71), and DF and IF for 1991 (-0.73). On the basis of these data and the simultaneous change of association of IF, MF and FF with other marker loci being dependent on the year, these traits must involve many QTL in common (QTLs with pleiotropic effects).

Hence, it has been shown that it is possible, not only to measure but also uncover the differential gene expression involved in G x E using the methodologies developed for QTL analysis and it is an essential step to establish the marker assisted selection scheme and to dissect the quantitative trait itself.

## REFERENCES

Asíns, M. J., P. Mestre, J. E. Garcia, F. Dicenta & E. A. Carbonell, 1994. Genotype x environment interaction in QTL analysis of an intervarietal almond cross by means of genetic markers. Theor. Appl. Genet., 89: 358-364

Asíns, M. J. & E. A. Carbonell, 1988. Detection of linkage between restriction fragment length polymorphism markers and quantitative traits. Theor. Appl. Genet. 76: 623-626.

Bretó, M. P., M. J. Asíns & E. A. Carbonell, 1994. Salt tolerance in *Lycopersicon* species. III. Detection of quantitative trait loci by means of molecular markers. Theor. Appl. Genet. 88: 395-401.

Carbonell, E. A., M. J. Asíns, M. Baselga, E. Balansard & T. M. Gerig, 1993. Power studies in the estimation of genetic parameters and the localization of quantitative trait loci for backcross and doubled haploid population. Theor. Appl. Genet. 86: 411-416.

Carbonell, E. A., T. M. Gerig, E. Balansard & M. J. Asíns, 1992. Interval mapping in the analysis of nonadditive quantitative trait loci. Biometrics 48: 305-315.

Cowen, N. M., 1988. The use of replicated progenies in marker based mapping of QTLs. Theor. Appl. Genet. 75: 857-862.

Dempster, A. P., N. M. Laird & D. B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., Series B 39: 1-38.

Dicenta, F., J. E. Garcia & E. A. Carbonell. 1993. Heritability of flowering, productivity and maturity in almond. J. Hort. Sci. 68: 113-120.

Dicenta, F., J. E. Garcia & E. A. Carbonell. 1993. Heritability of fruit characters in almond. J. Hort. Sci. 68: 121-126.

Falconer, D. S., 1989. Introduction to Quantitative Genetics. Longmans, London.

Gelderman, H., 1975. Investigation on inheritance of quantitative characters in animals by gene markers. 1. Methods. Theor. Appl. Genet. 46: 319-330.

Haley, C. S., S. A. Knott & J. M. Elsen, 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136: 1195-1207

Jansen, R. C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211

Jansen, R. C. & P. Stam, 1994. High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136: 1447-1455

Knott, S. A. & C. S. Haley, 1992. Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet. Res. Camb. 60: 139-151.

Lander, E. S. & D. Botsein, 1986. Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc. Natl. Acad. Sci. USA 83: 7353-7357.

Lander, E. S. & D. Botsein, 1989. Mapping mendelian factors underlying quantitative traits using RFLPs linkage maps. Genetics 121: 185-199.

Martinez, O. & R. N. Curnow, 1992. Estimating the location and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. 85: 480-485.

van Ooijen, J. W., 1992. Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. 84: 803-811.

Rebai, A., B. Goffinet & B. Mangin, 1994. Comparing power of different methods for QTL detection. Biometrics (in press).

Rodolphe, F. & M. Lefort, 1993. A multi-marker model for detecting chromosomal segments displaying QTL activity. Genetics 134: 1277-1288.

Sax, K., 1923. The association of size differences with seed coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8: 552-560.

Soller, M. & J. S. Beckmann, 1990. Marker-based mapping of quantitative trait loci using replicated progenies. Theor. Appl. Genet. 80: 205-208.

Soller, M., T. Brody & A. Genizi, 1976. On the power of experimental designs for the detection of linkage between marker loci an quantitative loci in crosses between inbred lines. Theor. Appl. Genet. 47: 35-39.

Thoday, J. M., 1961. Location of polygenes. Nature 191: 368-370.

de Vicente, M. C.& S. D. Tanksley, 1993. QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics 134: 585-596.

Zeng, Z. B., 1993. Theoretical basis of precision mapping of quantitative trait loci. Proc. Natl. Acad. Sci. USA 90: 10972-10976

Zeng, Z. B., 1994. Precision mapping of quantitative trait loci. Genetics 136: 1457-1468

Table 1. Frequencies of QTL genotypes within each genotypic class of the marker for an $F_2$ population.

| Markers | $Q_1Q_1$ | $Q_1Q_2$ | $Q_2Q_2$ |
|---|---|---|---|
| $M_1M_1$ | $(1-r)^2$ | $2r(1-r)$ | $r^2$ |
| $M_1M_2$ | $r(1-r)$ | $(1-2r+2r^2)$ | $r(1-r)$ |
| $M_2M_2$ | $r^2$ | $2r(1-r)$ | $(1-r)^2$ |

Table 2. Means and variances of the QTL within each genotypic marker class for $F_2$ populations.

| Marker genotype | Means | Variances [(a)] |
|---|---|---|
| $M_1M_1$ | $\mu+a(1-2r)+2dr(1-r)$ | $2a^2r(1-r)+2d^2ru-4adr(1-3r+2r^2)$ |
| $M_1M_2$ | $\mu+d(1-2r+2r^2)$ | $2a^2r(1-r)+2d^2ru$ |
| $M_2M_2$ | $\mu-a(1-2r)+2dr(1-r)$ | $2a^2r(1-r)+2d^2ru+4adr(1-3r+2r^2)$ |

(a) $u=1-3r+4r^2-2r^3$

Table 3. Marginal marker genotypic frequencies and expected and conditional frequencies of the two types of QTL genotypes given the genotypes of the flanking markers for a doubled haploid population.

| | | Conditional frequencies | | | |
|---|---|---|---|---|---|
| | | Under Haldane's | | Under no double | |
| Marker | Marginal | $Q_1Q_1$ | $Q_2Q_2$ | $Q_1Q_1$ | $Q_2Q_2$ |
| $A_1B_1/A_1B_1$ | (1 - p)/2 | (1 - s) | s | 1 | 0 |
| $A_1B_2/A_1B_2$ | p/2 | (1 - t) | t | r'/p | r/p |
| $A_2B_1/A_2B_1$ | p/2 | t | (1 - t) | r/p | r'/p |
| $A_2B_2/A_2B_2$ | (1 - p)/2 | s | (1 - s) | 0 | 1 |

p is the recombination fraction between the markers, s = rr'//1 - p), t = r(1 - r')/p, (assuming Haldane's function); and r and r' are the distance of the putative QTL to the left (A) and right (B) marker, respectively

Table 4.- Coding, heritabilities and mean values of the traits (taken from Dicenta and Garcia 1993)

| | | | Mean values | | $h^2$ | | |
|---|---|---|---|---|---|---|---|
| Traits | Code | Years | Ramillete | Tuono | 1989 | 1990 | 1991 |
| Initial Flowering | IF | 3 | 39.3 | 54.7 | 0.95±0.06 | 0.57±0.05 | 1.03±0.07 |
| Maximum Flowering | MF | 3 | 46.0 | 59.0 | 0.80±0.08 | 0.71±0.08 | 1.12±0.06 |
| Final Flowering | FF | 3 | 52.0 | 64.7 | 0.87±0.07 | 0.53±0.07 | 0.97±0.07 |
| Duration of Flowering | DF | 3 | 12.7 | 10.0 | 0.93±0.11 | 0.08±0.15 | 0.58±0.13 |
| Maturation Date | MD | 2 | 223.5 | 214.5 | 0.61±0.07 | 0.88±0.08 | |
| Duration of Maturation | DM | 2 | 170.5 | 150.0 | 0.56±0.08 | 0.89±0.12 | |
| Percentage of Failures | NF | 2 | 2.0 | 0.6 | 0.60±0.29 | 0.27±0.18 | |
| In-shell Weight | SW | 2 | 3.4 | 2.8 | 1.25±0.13 | 0.90±0.10 | |
| Kernel Weight | KW | 2 | 1.0 | 1.0 | 0.78±0.18 | 0.60±0.12 | |
| Percentage of Kernels | PK | 2 | 30.6 | 35.7 | 0.55±0.14 | 0.54±0.11 | |
| Number of Double Kernels | ND | 2 | 0.4 | 3.0 | 1.19±0.45 | 0.19±0.38 | |
| Flower Density | FD | 3 | 4.3 | 3.3 | 0.72±0.23 | 0.65±0.16 | 0.24±0.27 |
| Density of Yield | DY | 3 | 4.3 | 3.0 | 0.94±0.15 | 0.42±0.12 | 0.00±0.19 |
| Auto compatibility | AC | 1 | NO | YES | | | |
| Rugosity of the seed | RS | 2 | 1.0 | 2.5 | 0.56±0.11 | 0.64±0.11 | |
| Color of the seed | CS | 2 | 2.5 | 1.5 | 0.26±0.14 | 0.37±0.13 | |

Table 5.- Probabilities of the chi-square statistic calculated to compare the distributions of traits regarding the marker genotypes. *indicates probability lower than 0.05.

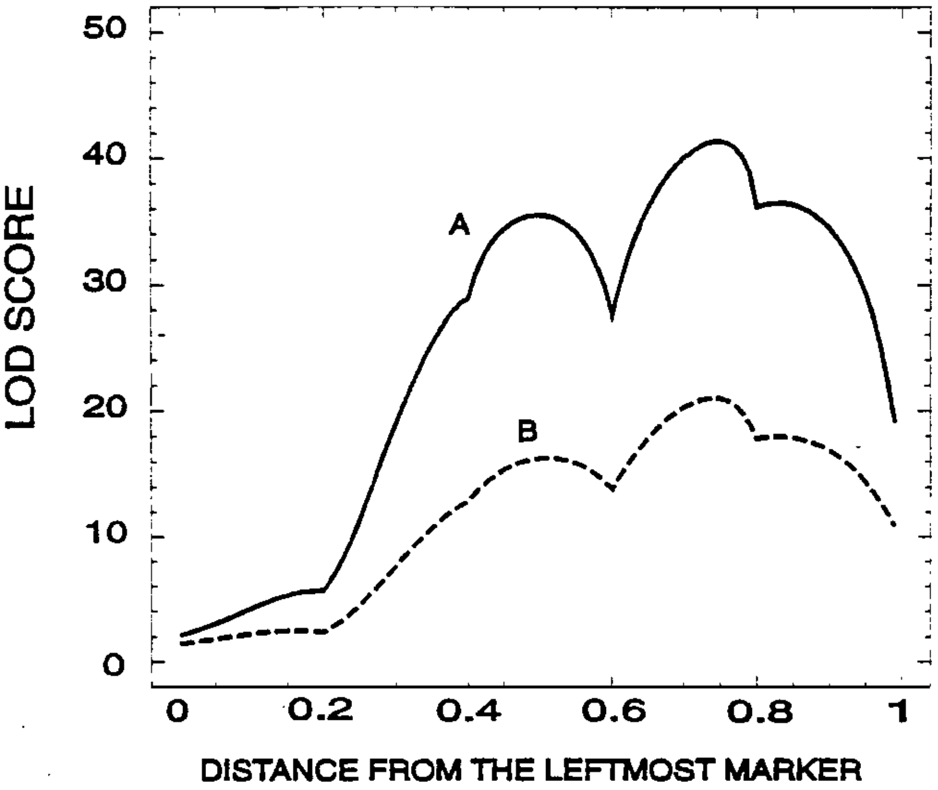
| | | Traits | | | | |
|---|---|---|---|---|---|---|
| Marker | Year | FD | DY | AC | RS | CS |
| Pgm-2 | 1989 | 0.27 | 0.68 | 0.28 | 0.75 | 0.02* |
| | 1990 | 0.47 | 0.49 | | 0.64 | 0.11 |
| | 1991 | 0.28 | 0.93 | | | |
| Pgi-2 | 1989 | 0.79 | 0.57 | 0.96 | 0.20 | 0.07 |
| | 1990 | 0.81 | 0.40 | | 0.93 | 0.13 |
| | 1991 | 0.25 | 0.35 | | | |
| $Prx_c$-2 | 1989 | 0.58 | 0.10 | 0.23 | 0.02* | 0.13 |
| | 1990 | 0.15 | 0.03* | | 0.27 | 0.03 |
| | 1991 | 0.69 | 0.74 | | | |
| Got-1 | 1989 | 0.34 | 0.97 | 0.89 | 0.46 | 0.36 |
| | 1990 | 0.91 | 0.68 | | 0.14 | 0.15 |
| | 1991 | 0.62 | 0.30 | | | |
| Got-2 | 1989 | 0.77 | 0.49 | 0.11 | 0.31 | 0.04* |
| | 1990 | 0.80 | 0.21 | | 0.18 | 0.23 |
| | 1991 | 0.54 | 0.73 | | | |

Table 6.- Probabilities of the F statistic obtained from the one-way ANOVA to test the association of the quantitative trait with a marker. * and ** indicate probability lower than 0.05 and 0.01, respectively.

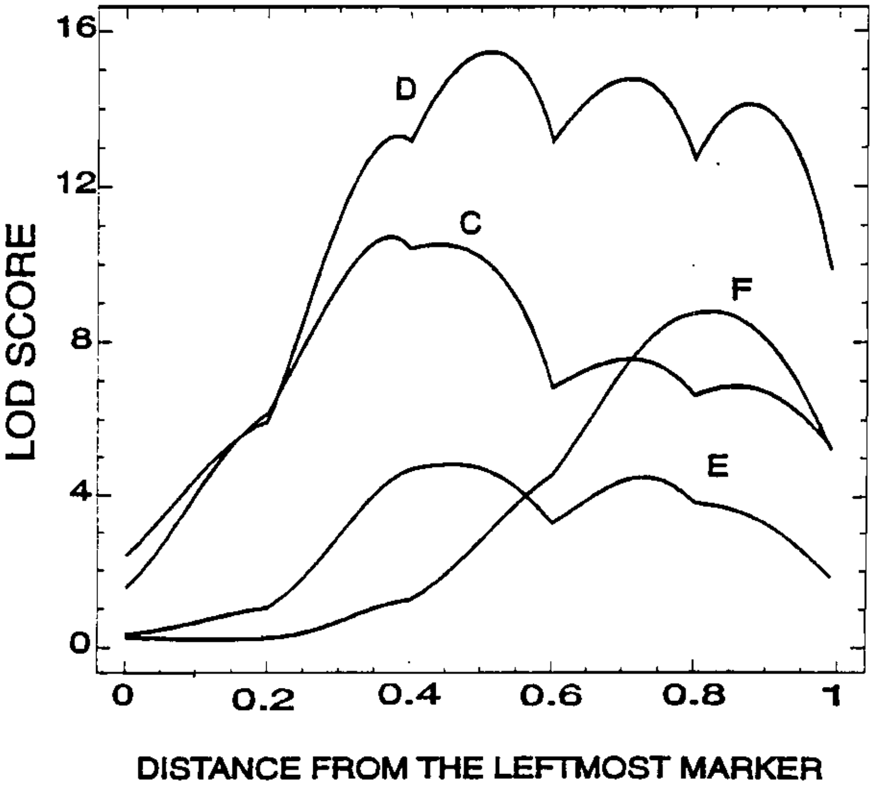
| | | IF | MF | FF | DF | MD | DM | NF | SW | KW | PK | ND |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pg -2 | 1989 | 0.07 | 0.32 | 0.73 | 0.23 | 0.16 | 0.12 | 0.18 | 0.17 | 0.03* | 0.89 | 0.23 |
| | 1990 | 0.05* | 0.10 | 0.51 | 0.06 | 0.57 | 0.45 | 0.61 | 0.28 | 0.02*.. | 0.68 | 0.12 |
| | 1991 | 0.58 | 0.11 | 0.25 | 0.60 | | | | | | | |
| Pgi-2 | 1989 | 0.08 | 0.06 | 0.72 | 0.23 | 0.21 | 0.16 | 0.87 | 0.05* | 0.04 | 0.39 | 0.73 |
| | 1990 | 0.27 | 0.20 | 0.48 | 0.52 | 0.85 | 0.92 | 0.82 | 0.03* | 0.27 | 0.13 | 0.74 |
| | 1991 | 0.25 | 0.15 | 0.28 | 0.51 | | | | | | | |
| Prx$_c$-2 | 1989 | 0.79 | 0.33 | 0.51 | 0.25 | 0.21 | 0.13 | 0.49 | 0.72 | 0.40 | 0.56 | 0.21 |
| | 1990 | 0.36 | 0.33 | 0.13 | 0.51 | 0.34 | 0.70 | 0.23 | 0.08 | 0.02* | 0.01** | 0.81 |
| | 1991 | 0.00** | 0.01** | 0.02* | 0.04* | | | | | | | |
| Got-1 | 1989 | 0.21 | 0.34 | 0.07 | 0.36 | 0.89 | 0.51 | 0.58 | 0.64 | 0.20 | 0.81 | 0.18 |
| | 1990 | 0.18 | 0.08 | 0.00** | 0.07 | 0.95 | 0.34 | 0.60 | 0.92 | 0.27 | 0.99 | 0.03* |
| | 1991 | 0.05* | 0.23 | 0.11 | 0.11 | | | | | | | |
| Got-2 | 1989 | 0.04* | 0.06 | 0.00** | 0.08 | 0.48 | 0.06 | 0.58 | 0.41 | 0.86 | 0.25 | 0.30 |
| | 1990 | 0.00** | 0.00** | 0.00** | 0.91 | 0.22 | 0.03* | 0.39 | 0.70 | 0.90 | 0.35 | 0.88 |
| | 1991 | 0.51 | 0.84 | 0.96 | 0.22 | | | | | | | |

## Figure Legends

Figure 1. Typical results found for two linked QTLs actually located at 42 and 82 cM from the leftmost marker. Curve F involving a single QTL at 82 cM is included for comparison purposes.



LOD SCORE
50
40
30
20
10
0
A
B
0
0.2
0.4
0.6
0.8
1
DISTANCE FROM THE LEFTMOST MARKER



LOD SCORE
16
12
8
4
0
D
C
F
E
0
0.2
0.4
0.6
0.8
1
DISTANCE FROM THE LEFTMOST MARKER

# CATEGORICAL DATA ANALYSIS IN BIOTECHNOLOGY RESEARCH

E. Barry Moser and Raúl E. Macchiavelli
Dept. of Experimental Statistics, Louisiana State University

Paper presented at the meeting "Statistical Methods in Environmental Research and Molecular Biology", Cali, Colombia, June 1995

## Abstract

Categorical responses are very common in biotechnology research. In this paper we present the main ideas of generalized linear models, with particular emphasis to logit models. We apply these models to two practical situations. The first problem studies the effect of chilling on bud development in azalea flowers, and the second situation analyzes the relationship between a genetic polymorphism in certain protein and obesity in humans. Both problems are analyzed with SAS PROC CATMOD.

## Resumen

Respuestas medidas en una escala categórica aparecen comúnmente en la investigación biotecnológica. En este trabajo se presentan las ideas principales de los modelos lineales generalizados, enfatizando los modelos *logit*. Se presentan dos aplicaciones prácticas. En la primera se estudia el efecto del tratamiento de frío sobre el desarrollo floral en plantas de azalea. El segundo problema estudia la relación entre un polimorfismo genético en cierta proteína y obesidad en humanos. Ambos problemas se analizan en PROC CATMOD en SAS.

## 1. Introduction

The linear model for normally distributed random variables is very commonly used to address hypotheses concerning the mean or expected value of the random variables. Let $y_{ij}$ be a normally distributed response measured on the $j$th individual of group $i$. Then $y_{ij}$ could be modeled as

$$y_{ij} = \mu_i + \epsilon_{ij} \text{ for } i = 1,2,\cdots,I; j = 1,2,\cdots,n_i \tag{1}$$

where $\mu_i$ is the $i$th group mean, $\epsilon_{ij}$ is the error or amount the $j$th individual of group $i$ differs from its mean, $I$ is the number of groups, and $n_i$ is the number of individuals in group $i$. Hypotheses of interest often include the null hypothesis $H_0$: $\mu_1 = \cdots = \mu_I$ and in general, these hypotheses can be tested using linear contrasts of the parameters. Alternative parameterizations of model (1) can be formulated and may lead to direct estimation of desired parameters and testing of desired hypotheses by the software employed. In fact a common parameterization of model (1) is the "effects" model

$$\mu_i = \mu + \alpha_i \text{ and } \sum_{i=1}^{I} \alpha_i = 0 \tag{2}$$

where $\mu$ serves as an "overall" mean and the $\alpha_i$ are the amounts by which the individual factor level means differ from the "overall" mean.

In categorical data analysis, we often observe the number of items out of a total that fall into each of several mutually exclusive categories $K$. Based upon random sampling principles these data are often treated as a sample from a multinomial distribution, although other distributions such as the Poisson and the negative binomial distribution are also used. It is tempting to model the category counts using the normal theory model proposed above. However, we must account for several qualities of the categorical data, and the normal theory approach may not adequately address them. First, we must define what the parameters of interest are. For multinomial data, the parameters of interest would be the proportion of individuals in the population that belong to each category or the probability that a randomly selected individual would belong to a given category. We must also account for the fact that the proportions or probabilities of the mutually exclusive categories must sum to 1. Further, we would like to restrict any predictions or estimates of proportions or probabilities to the interval [0,1].

A general class of models called generalized linear models (see e.g., Dobson 1983, Agresti 1990) have been developed to address issues such as those raised above for multinomial models. In generalized linear models, the model parameters enter the equation in a linear, additive manner, such as proposed in (1) and (2) above, yet the distribution of the response is accounted for through the estimation method and through the way in which the mean is included in the model (the link function). Generalized linear models address distributions that belong to the exponential family of distributions. A distribution belongs to the exponential family if it is of the form

$$f(y;\theta) = \exp(a(y)b(\theta) + c(\theta)+d(y)). \tag{3}$$

For $Y$ distributed as a normal$(\mu,\sigma^2)$ random variable we have

$$f(y;\theta) = \exp(y\frac{\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)-\frac{y^2}{2\sigma^2}). \tag{4}$$

Notice that $\sigma^2$ is treated as a nuisance parameter and is assumed to be known. For $Y$ distributed binomial$(n,\pi)$,

$$f(y;\theta) = \exp(y\log\pi - y\log(1-\pi)+n\log(1-\pi)+\log\binom{n}{y}) \tag{5}$$

or

$$f(y;\theta) = \exp(y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}). \tag{6}$$

If $a(y) = 1$ as it does for both the normal and binomial distributions, then the distribution is said to be in the canonical form and $b(\theta)$ is called the natural parameter of the distribution. In generalized linear models it is common to model the natural parameter as a linear function of the model predictors. Thus, the transformation specified by the natural parameter is applied to the observed means, which are then modeled as a linear function of the model parameters. For the normal distribution, the response function $F$ is taken to be $\mu/\sigma^2$, or usually just $\mu$, and is modeled as a linear function of the model parameters. Remember that $\sigma^2$ is assumed known. In practice it is estimated from the error in fit. For the binomial distribution, the response function $F$ is taken to be the logit transformation,

$$F = \log\left(\frac{\pi}{1-\pi}\right) \tag{7}$$

and is modeled as a linear function, say, of the groups,

$$F_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mu + \alpha_i \tag{8}$$

as we did before with the normal theory data, where $\pi_i$ is the probability of being in the $i$th group. The parameters of these models are usually estimated using maximum likelihood although alternatives based upon weighted least squares and chi-square goodness-of-fit criteria are also available (see e.g., Agresti 1990:445-477). If the parameters are estimated using maximum likelihood, then likelihood ratio and Wald tests can be used to test hypotheses about the model parameters. Pearson chi-square and likelihood ratio goodness-of-fit tests can also be used to evaluate model adequacy for restricted models.

Assume that we have observed a binomial response from each of two populations and we would like to test the null hypothesis that $H_0\colon \pi_1 = \pi_2$. This hypothesis implies the hypothesis $H_0\colon F_1 = F_2$ (or $H_0\colon F_1 - F_2 = 0$). Thus, we can consider the functions $F_1$ and $F_2$ and construct the linear hypothesis

$$F_1 - F_2 = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_2}{1-\pi_2}\right) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) \tag{9}$$

which is simply the log-odds ratio, the commonly reported statistic in contingency table analysis (Fienberg 1980:17).

If there are more than 2 groups, and these groups, for example, are arranged in a factorial treatment arrangement, then additional model parameters can be specified to focus upon the main effects and interactions of the treatment, experimental, and/or observational factors. For a 2-factor factorial treatment arrangement, the response function could be modeled as

$$F_{ij} = \log\left(\frac{1-\pi_{ij}}{\pi_{ij}}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{10}$$

with $\sum_{i=1}^{I} \alpha_i = 0,\ \sum_{j=1}^{J} \beta_j = 0, \sum_{i=1}^{I} (\alpha\beta)_{ij} = 0 \quad \forall j,\ \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0 \quad \forall i$ where the $\alpha_i$ compose the main effect of factor 1 for $i = 1,2,...,I$ levels, the $\beta_j$ compose the main effect of factor 2 for $j = 1,2,...,J$ levels, and the $(\alpha\beta)_{ij}$ compose the interaction of factor 1 with factor 2. The hypothesis $H_0\!: \alpha_1 = \alpha_2 = ... \alpha_I = 0$ tests for homogeneity of proportions across the levels of factor 1. The hypothesis $H_0\!: \beta_1 = \beta_2 = ... \beta_J = 0$ tests for homogeneity of proportions across the levels of factor 2. While the hypothesis $H_0\!: (\alpha\beta)_{ij} = 0 \quad \forall \quad (i,j)$,, tests that the distribution of proportions across the levels of factor 1 is the same across each of the levels of factor 2.

If more than 2 response categories exist, then the logit transformation can be generalized in several ways. In this paper we will focus on categories that are ordered. For an ordinal response the cumulative logit transformation is commonly applied. Let $i$ refer to one of the $I$ response categories, then this transformation computes the log of the odds of being in a category larger than $i$ to that of being in or below category $i$. If we have sampled from $J$ groups, then the response functions become

$$F_{ij} = \log\left(\frac{1 - \sum_{k=1}^{i} \pi_k}{\sum_{k=1}^{i} \pi_k}\right). \tag{11}$$

Since there are now more than 2 categories, we will need to add a modeling term, $z_i$, corresponding to the level of categories being compared in the odds ratio. This permits the odds to depend upon which categories are being compared. Now our model for the response functions can take the form

$$F_{ij} = \mu + Z_i + \beta_j \tag{12}$$

where $z_i$ behaves much like a covariate in an analysis of covariance model.

The categorical models developed in (10) and (12) will now be illustrated through 2 biotechnology experiments.

## 2. Chilling and Bud Development in Azaleas

A continuing problem in azalea production for flowering potted plants has been uniform flowering. During normal production cycle, azaleas will not bloom unless flower bud dormancy is overcome. Traditionally, breaking dormancy of azalea flower buds is accomplished by exposing plants to temperatures of 3-10$^\circ C$ (chilling) through natural cooling or refrigerated storage.

In this data set, obtained by Rebecca Moss (Dept. of Horticulture, Louisiana State University), the developmental stage of flower buds from plants which were chilled and non-chilled were recorded for three different shipments. The stages were classified in two categories ($b$,$c$). (Other shipments also included category $a$.)

The observed frequencies are presented in table 1. There are 12 cells.

**Table 1**: Observed frequencies of buds in two developmental stages. Data provided by R. Moss.

| Treatment | Shipment | Stage *b* | Stage *c* |
|---|---|---|---|
| non-chilled | 1 | 166 | 150 |
| non-chilled | 2 | 231 | 102 |
| non-chilled | 3 | 370 | 7 |
| chilled | 1 | 202 | 196 |
| chilled | 2 | 145 | 147 |
| chilled | 3 | 360 | 3 |

Let $\pi_{ij}$ be the probability that a bud will be in stage $b$ (in a plant from shipment $j$ treated with treatment $i$). Since the response has only two levels (stage b or c), the response function to consider is

$$F_{ij} = \log \frac{\pi_{ij}}{1-\pi_{ij}},$$

for $i$=1,2 (treatments) and $j$=1,2,3 (shipments). $F_{ij}$ is the logarithm of the odds in favor of a bud being in stage $b$ (instead of stage $c$) in a plant from treatment $i$ and shipment $j$.

The saturated model will be a model with one parameter for each $F_{ij}$ (or equivalently, for each $\pi_{ij}$). The fit for this model is trivially exact (no goodness of fit test is possible).

Notice that any model with 6 or more linearly independent parameters will be equivalent to this one. For example,

$$F_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}, \tag{1}$$

with the usual restrictions $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \alpha\beta_{ij} = \sum_j \alpha\beta_{ij} = 0$ has six parameters and is equivalent to the saturated one.

In order to fit this model using SAS PROC CATMOD, the following program can be used:

```
proc catmod;
model nwstg= trt shp trt*shp / freq pred prob;
```

The output obtained will include the frequency table presented as Table 1 (requested by the option **freq**), the observed probabilities (or relative frequency table, requested by **prob**), information about the iterative estimation procedure, an "analysis of variance table" presenting chi-squared tests for the main hypotheses of interest, and a list of estimated parameters. See Appendix A for a complete output.

A simpler model (with main effects only) was tried with the commands:

```
proc catmod;
model nwstg= trt shp / freq pred prob;
```

This model has only 4 parameters, leaving 2 degrees of freedom to test the goodness of fit. The likelihood ratio statistic for this test was 14.91, with a p-value of .0006. Therefore this model does not fit well, and the interaction term should be included (i.e., the saturated model will be used).

Like in ANOVA models, the presence of interactions makes the tests about main effects not very interesting. Contrasts can be used to test the hypotheses of interest (like simple effects). In this problem, 4 hypotheses were identified as important to the researcher. They were:

- Does chilling have an effect in shipment 1?
- Does chilling have an effect in shipment 2?
- Does chilling have an effect in shipment 3?
- Do non-chilled plants differ among shipments?

The last one will be a contrast with 2 df, while all the others will have 1 df each. In order to write the CONTRAST statements in SAS, we need to write them in terms of CATMOD full rank parametrization (which has "sum-to-zero" restrictions for all parameters).

For example, the first contrast is

$$H_0: F_{11} - F_{21} = 0.$$

Writing this contrast in terms of the parameters in the model (1),

$$\mu + \alpha_1 + \beta_1 + \alpha\beta_{11} - (\mu + \alpha_2 + \beta_1 + \alpha\beta_{21}) = 0.$$

Since $\alpha_2 = -\alpha_1$ and $\alpha\beta_{21} = -\alpha\beta_{11}$, the contrast is

$$2\alpha_1 + 2\alpha\beta_{11} = 0.$$

The other two simple effects of treatment can be obtained similarly. Notice that this form is not a very familiar one for a contrast, but it can be written directly in SAS:

```
contrast 'trt in shp1' trt 2 trt*shp 2 0;
contrast 'trt in shp2' trt 2 trt*shp 0 2;
contrast 'trt in shp3' trt 2 trt*shp -2 -2;
```

The last contrast (shipment effect in non-chilled plants) can be written as

$$H_0: F_{11} - F_{13} = 0 \text{ and } F_{12} - F_{13} = 0.$$

The corresponding SAS statement is

```
contrast 'nonchilled' shp 2 -1 trt*shp 2 -1, shp -1 2 trt*shp 1 -2;
```

The output for these contrasts are likelihood ratio tests:

```
        CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

Contrast                DF   Chi-Square    Prob
-------------------------------------------------
trt in shp1              1       0.22      0.6368
trt in shp2              1      24.82      0.0000
trt in shp3              1       1.40      0.2373
nonchilled               2      95.27      0.0000
```

From these results we can interpret the interaction very clearly: while chilling does not seem to have an effect in shipments 1 and 3, it does have a significant effect in shipment 2. Also, there are differences between shipments in the non treated plants.

## 3. Effects of genetic polymorphism in APO-AIV

In this example we will use CATMOD to model ordinal responses. The study involves determining the effect of genetic polymorphism in apolipoprotein AIV (APO-AIV) on body mass index in human subjects. The body mass index is defined as

$$BMI = \frac{\text{weight } (kg)}{\text{height }^2(\text{in m}^2)}.$$

APO-AIV is a protein associated with HDL (high density lipoprotein) in plasma and as "free" APO-AIV. It might have a role in reverse cholesterol metabolism or in triglycerides.

Several phenotypes have been identified for APO-AIV. Two of them were the most frequent and were used in this analysis (they were coded as 0, 1). The data were obtained by Dr. Michael Lefevre (Pennington Biomedical Research Center, Louisiana State University).

The problem is to check how these phenotypes differ on BMI. Since this variable, representing how "obese" or "lean" a person is, also depends on several other factors (like age, sex, race, etc.), the researcher did not use the actual BMI, but the BMI percentile. This BMI percentile was obtained from the raw BMI score by looking at tables constructed for each sex, race and age group. This is a way of eliminating the effects of these factors, transforming the response into an ordinal variable. The percentiles considered were 5, 10, 15, 20, 50, 75, 85, 90, 95 and 99.

The research question in this study is "do people in phenotype 1 tend to be in a higher BMI percentile than people in phenotype 0?". This question can be addressed by modelling the *cumulative logits*

(CLOGIT option). Let $\pi_{ij}$ be the probability that a subject will be in the *i*-th ordered class and the *j*-th group. The cumulative logit $F_{ij}$ is defined as the logarithm of the odds in favor of being in a class larger than *i*:

$$F_{ij} = \log \frac{1-\sum_{k=1}^{i} \pi_{kj}}{\sum_{k=1}^{i} \pi_{kj}} \qquad (2)$$

The linear model for this $F_{ij}$'s is

$$F_{ij} = \mu + \tau_i + \beta_j,$$

where $\tau$ represents the percentile effect and $\beta$ the phenotype effect. The usual restrictions apply $(\sum \tau_i = \sum \beta_j = 0)$. This model can be fitted with the following SAS program:

```
proc catmod data=colo.bmi;
response clogits;
model bmi_per = _response_ pheno / freq predict prob;
weight tot;
```

The results of this analysis are presented as Appendix B. The first thing to notice is the goodness of fit test, which has 8 d.f. In the output it is found under "residual". The $\chi^2$ statistic is 4.75, with a p-value of .7835. From this test, one concludes that the fit is good, and hence we can interpret the parameter estimates. Another part of the output to check for goodness of fit is the listing of residuals (obtained with the **predict** option). There is certain pattern of negative residuals for low function

numbers in PHENO=0 and high function numbers in PHENO=1. This is an effect of requiring a constant odds-ratio in the model (which we will discuss later).

The main research question (PHENO effect) can be answered by looking at the $\chi^2$ statistic for PHENO. Its value is 5.48, with 1 d.f. and a p-value of .0192. Therefore, there is a significant effect of phenotype on BMI. In order to interpret the nature of this effect, we need to look at $\hat{\beta} = .23$. (Notice that there is only one parameter associated with PHENO, since $\beta_2 = -\beta_1$.)

The difference of the functions associated with the same percentile at the two values of PHENO is

$$F_{i1} - F_{i2} = \mu + \tau_i + \beta - (\mu + \tau_i - \beta) = 2\beta.$$

If we write the definition of the cumulative logits we can conclude that

$$\log \frac{\dfrac{1-\sum_{k=1}^{i} \pi_{k1}}{\sum_{k=1}^{i} \pi_{k1}}}{\dfrac{1-\sum_{k=1}^{i} \pi_{k2}}{\sum_{k=1}^{i} \pi_{k2}}} = 2\beta.$$

This equation is saying that the odds ratio (the ratio of the odds in favor of being in a class larger than *i* for PHENO=0 and 1) is a constant:

$$\frac{\text{odds for PHENO} = 0}{\text{odds for PHENO} = 1} = \exp(2\beta).$$

Notice that this does not depend on the particular percentile ($i$). The estimated odds-ratio is obtained from $\hat{\beta}$ :

$$\exp(2\hat{\beta}) = 1.58$$

From this relationship, we estimate that the odds in favor of being in a class larger than $i$ (for any $i$) are 1.58 times larger in PHENO=0 than in PHENO=1. Hence, people with PHENO=0 are more likely to be in the higher percentiles (i.e., being more "obese") than people with PHENO=1.

In order to obtain the fitted probabilities we can proceed as follows. Consider PHENO=0.

$$\hat{F}_{11} = \log\frac{1 - \hat{\pi}_{11}}{\hat{\pi}_{11}} = \hat{\mu} + \hat{\tau}_1 + \hat{\beta} = 1.0382 + 2.9725 + .2300 = 4.2407 .$$

Solving this equation for $\hat{\pi}_{11}$ we obtain

$$\hat{\pi}_{11} = \frac{1}{1 + e^{4.2407}} = .0142 .$$

Similarly,

$$\hat{F}_{11} = \log\frac{1 - \hat{\pi}_{11} - \hat{\pi}_{21}}{\hat{\pi}_{11} + \hat{\pi}_{21}} = \hat{\mu} + \hat{\tau}_2 + \hat{\beta} = 1.0382 + 2.1580 + .2300 = 3.4262 .$$

From this equation,

$$\hat{\pi}_{11} + \hat{\pi}_{21} = \frac{1}{1 + e^{3.4262}} = .0315 ,$$

and $\hat{\pi}_{21} = .0173$ . Continuing in this fashion we can obtain all fitted probabilities.

## References

Agresti, A. (1990) *Categorical Data Analysis*. John Wiley & Sons, New York.

Dobson, A. J. (1983). *An Introduction to Statistical Modelling*. Chapman and Hall, London.

Fienberg, S. E. (1980). *The Analysis of Cross-classified Categorical Data*. Second edition. The MIT Press, Cambridge, MA.

SAS Institute, Inc. (1990). *SAS/STAT User's Guide, Version 6*. Fourth edition, Volume 1. SAS Institute, Inc., Cary, NC.

## A. Azalea Example: SAS Output

```
                         CATMOD PROCEDURE

Response: NWSTG                   Response Levels (R)=  2
Weight Variable: None             Populations     (S)=  6
Data Set: A                       Total Frequency (N)=  2079
Frequency Missing: 0              Observations  (Obs)=  2079


                        POPULATION PROFILES

                                             Sample
                    Sample  TRT  SHP          Size
                    --------------------------------
                       1     1    4           316
                       2     1    5           333
                       3     1    6           377
                       4     3    4           398
                       5     3    5           292
                       6     3    6           363



                        RESPONSE PROFILES

                         Response  NWSTG
                         ---------------
                             1       b
                             2       c
```

```
    RESPONSE FREQUENCIES


         Response Number
Sample         1         2
--------------------------
    1        166       150
    2        231       102
    3        370         7
    4        202       196
    5        145       147
    6        360         3



   RESPONSE PROBABILITIES


         Response Number
Sample         1         2
--------------------------
    1    0.52532   0.47468
    2    0.69369   0.30631
    3    0.98143   0.01857
    4    0.50754   0.49246
    5    0.49658   0.50342
    6    0.99174   0.00826
```

```
                 MAXIMUM-LIKELIHOOD ANALYSIS

                   Sub        -2 Log      Convergence
 Iteration      Iteration   Likelihood     Criterion
 -------------------------------------------------------
     0              0         2882.106        1.0000
     1              0        2056.5688        0.2864
     2              0         1951.538        0.0511
     3              0         1927.222        0.0125
     4              0        1923.5425      0.001909
     5              0        1923.3692     0.0000901
     6              0        1923.3686     3.1756E-7
     7              0        1923.3686     4.598E-12
```

```
                          Non Interaction Model

                               Parameter Estimates
Iteration          1            2            3            4
-------------------------------------------------------------------
    0              0            0            0            0
     1        0.8059       0.1235      -0.7301      -0.4076
     2        1.1519       0.1761      -1.0695      -0.7485
     3        1.4079       0.1894      -1.3238      -1.0046
     4        1.5540       0.1908      -1.4697      -1.1508
     5        1.5947       0.1909      -1.5105      -1.1915
     6        1.5973       0.1909      -1.5131      -1.1941
     7        1.5973       0.1909      -1.5131      -1.1941
```

```
           MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

         Source                   DF   Chi-Square      Prob
         --------------------------------------------------
         INTERCEPT                 1       201.41    0.0000
         TRT                       1        12.10    0.0005
         SHP                       2       167.10    0.0000

         LIKELIHOOD RATIO          2        14.91    0.0006




              ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                           Standard   Chi-
 Effect        Parameter  Estimate    Error      Square    Prob

-------------------------------------------------------------------
  INTERCEPT            1    1.5973    0.1126     201.41  0.0000
  TRT                  2    0.1909    0.0549      12.10  0.0005
  SHP                  3   -1.5131    0.1206     157.51  0.0000
                       4   -1.1941    0.1222      95.51  0.0000
```

## MAXIMUM-LIKELIHOOD PREDICTED VALUES FOR RESPONSE FUNCTIONS AND PROBABILITIES

| | | --------Observed------- | | ---------Predicted------- | | |
|---|---|---|---|---|---|---|
| | Function | | Standard | | Standard | |
| Sample | Number | Function | Error | Function | Error | Residual |
| 1 | 1 | 0.10135249 | 0.11265329 | 0.27517417 | 0.09723315 | -0.1738217 |
| | P1 | 0.52531646 | 0.02809112 | 0.56836271 | 0.02385387 | -0.0430463 |
| | P2 | 0.47468354 | 0.02809112 | 0.43163729 | 0.02385387 | 0.04304626 |
| 2 | 1 | 0.8174449 | 0.11888198 | 0.59415953 | 0.09787989 | 0.22328537 |
| | P1 | 0.69369369 | 0.02526037 | 0.64431896 | 0.02243133 | 0.04937473 |
| | P2 | 0.30630631 | 0.02526037 | 0.35568104 | 0.02243133 | -0.0493747 |
| 3 | 1 | 3.96759286 | 0.38152306 | 4.49543204 | 0.32481564 | -0.5278392 |
| | P1 | 0.98143236 | 0.00695245 | 0.98896331 | 0.00354532 | -0.0075309 |
| | P2 | 0.01856764 | 0.00695245 | 0.01103669 | 0.00354532 | 0.00753095 |
| 4 | 1 | 0.03015304 | 0.10026234 | -0.10666 | 0.08930642 | 0.13681301 |
| | P1 | 0.50753769 | 0.02505989 | 0.47336026 | 0.02226323 | 0.03417743 |
| | P2 | 0.49246231 | 0.02505989 | 0.52663974 | 0.02226323 | -0.0341774 |
| 5 | 1 | -0.0136988 | 0.11704389 | 0.21232539 | 0.09957698 | -0.2260242 |
| | P1 | 0.49657534 | 0.0292596 | 0.55288282 | 0.02461577 | -0.0563075 |
| | P2 | 0.50342466 | 0.0292596 | 0.44711718 | 0.02461577 | 0.05630748 |
| 6 | 1 | 4.78749174 | 0.5797509 | 4.1135979 | 0.32173498 | 0.67389384 |
| | P1 | 0.99173554 | 0.00475173 | 0.98391414 | 0.00509213 | 0.0078214 |
| | P2 | 0.00826446 | 0.00475173 | 0.01608586 | 0.00509213 | -0.0078214 |

```
                  Saturated Model

             MAXIMUM-LIKELIHOOD ANALYSIS

                  Sub          -2 Log       Convergence
Iteration      Iteration     Likelihood      Criterion
---------------------------------------------------------
    0             0           2882.106         1.0000
    1             0          2040.5745         0.2920
    2             0          1937.3417         0.0506
    3             0           1912.897         0.0126
    4             0          1908.7759       0.002154
    5             0          1908.4615       0.000165
    6             0          1908.4568      2.4755E-6
    7             0          1908.4567      8.935E-10
```

Parameter Estimates

| Iteration | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0. | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0.7975 | 0.1364 | -0.7318 | -0.4170 | -0.1008 | 0.2578 |
| 2 | 1.1449 | 0.1294 | -1.0791 | -0.7432 | -0.0938 | 0.2860 |
| 3 | 1.4013 | 0.0990 | -1.3355 | -0.9994 | -0.0634 | 0.3166 |
| 4 | 1.5550 | 0.0529 | -1.4892 | -1.1531 | -0.0173 | 0.3627 |
| 5 | 1.6082 | 0.0200 | -1.5424 | -1.2063 | 0.0156 | 0.3956 |
| 6 | 1.6149 | 0.0139 | -1.5492 | -1.2131 | 0.0217 | 0.4017 |
| 7 | 1.6151 | 0.0137 | -1.5493 | -1.2132 | 0.0219 | 0.4018 |

```
        MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

      Source                   DF   Chi-Square      Prob
      --------------------------------------------------
      INTERCEPT                 1       176.51    0.0000
      TRT                       1         0.01    0.9100
      SHP                       2       149.12    0.0000
      TRT*SHP                   2        14.50    0.0007

      LIKELIHOOD RATIO          0            .         .



                          Saturated Model

            ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                       Standard    Chi-
Effect        Parameter  Estimate     Error       Square       Prob
--------------------------------------------------------------------

INTERCEPT         1        1.6151     0.1216     176.51     0.0000
TRT               2        0.0137     0.1216       0.01     0.9100
SHP               3       -1.5493     0.1291     143.97     0.0000
                  4       -1.2132     0.1308      86.09     0.0000
TRT*SHP           5        0.0219     0.1291       0.03     0.8656
                  6        0.4018     0.1308       9.44     0.0021
```

MAXIMUM-LIKELIHOOD PREDICTED VALUES FOR RESPONSE FUNCTIONS AND PROBABILITIES

| | | --------Observed--------- | | --------Predicted--------- | | |
|---|---|---|---|---|---|---|
| | Function | | Standard | | Standard | |
| Sample | Number | Function | Error | Function | Error | Residual |
| 1 | 1 | 0.10135249 | 0.11265329 | 0.10135249 | 0.11265329 | 0 |
| | P1 | 0.52531646 | 0.02809112 | 0.52531646 | 0.02809112 | 0 |
| | P2 | 0.47468354 | 0.02809112 | 0.47468354 | 0.02809112 | 0 |
| 2 | 1 | 0.8174449 | 0.11888198 | 0.8174449 | 0.11888198 | 0 |
| | P1 | 0.69369369 | 0.02526037 | 0.69369369 | 0.02526037 | 0 |
| | P2 | 0.30630631 | 0.02526037 | 0.30630631 | 0.02526037 | 0 |
| 3 | 1 | 3.96759286 | 0.38152306 | 3.96759286 | 0.38152275 | 0 |
| | P1 | 0.98143236 | 0.00695245 | 0.98143236 | 0.00695244 | 0 |
| | P2 | 0.01856764 | 0.00695245 | 0.01856764 | 0.00695244 | 0 |
| 4 | 1 | 0.03015304 | 0.10026234 | 0.03015304 | 0.10026234 | 0 |
| | P1 | 0.50753769 | 0.02505989 | 0.50753769 | 0.02505989 | 0 |
| | P2 | 0.49246231 | 0.02505989 | 0.49246231 | 0.02505989 | 0 |
| 5 | 1 | -0.0136988 | 0.11704389 | -0.0136988 | 0.11704389 | 0 |
| | P1 | 0.49657534 | 0.0292596 | 0.49657534 | 0.0292596 | 0 |
| | P2 | 0.50342466 | 0.0292596 | 0.50342466 | 0.0292596 | 0 |
| 6 | 1 | 4.78749174 | 0.5797509 | 4.78749146 | 0.57953514 | 2.81698E-7 |
| | P1 | 0.99173554 | 0.00475173 | 0.99173553 | 0.00474996 | 2.30884E-9 |
| | P2 | 0.00826446 | 0.00475173 | 0.00826447 | 0.00474996 | -2.3088E-9 |

```
             CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

      Contrast                      DF    Chi-Square        Prob
      ----------------------------------------------------------
      nonchilled                     2         95.27      0.0000
      trt in shp4                    1          0.22      0.6368
      trt in shp5                    1         24.82      0.0000
      trt in shp6                    1          1.40      0.2373
```

B. APO-AIV Example: SAS Output

```
                          CATMOD PROCEDURE

 Response: BMI_PER                 Response Levels (R)=     10
 Weight Variable: TOT             Populations      (S)=      2
 Data Set: BMI                    Total Frequency (N)=    615
 Frequency Missing: 0              Observations   (Obs)=     20


                        POPULATION PROFILES

                                      Sample
                     Sample   PHENO     Size
                     ------------------------
                          1     0        520
                          2     1         95
```

```
                    RESPONSE PROFILES

                    Response   BMI_PER
                    ----------------
                          1          5
                          2         10
                          3         15
                          4         25
                          5         50
                          6         75
                          7         85
                          8         90
                          9         95
                         10         99



                  RESPONSE FREQUENCIES

                             Response Number
Sample           1          2          3          4          5
--------------------------------------------------------------
     1           8         10          6         28         76
     2           1          1          2          8         21


                             Response Number
Sample           6          7          8          9         10
--------------------------------------------------------------
     1         105         74         55         77         81
     2          18         16         12          8          8
```

```
                  RESPONSE PROBABILITIES

                           Response Number
Sample         1        2        3        4        5
------------------------------------------------------
     1   0.01538  0.01923  0.01154  0.05385  0.14615
     2   0.01053  0.01053  0.02105  0.08421  0.22105


                           Response Number
Sample         6        7        8        9       10
------------------------------------------------------
     1   0.20192  0.14231  0.10577  0.14808  0.15577
     2   0.18947  0.16842  0.12632  0.08421  0.08421


                Function     Response
            Sample  Number       Function
            ---------------------------
               1         1        4.15888
                         2        3.32823
                         3        3.02852
                         4        2.19722
                         5        1.11923
                         6        0.20844
                         7       -0.36556
                         8       -0.82905
                         9       -1.69005

               2         1        4.54329
                         2        3.83945
                         3        3.12457
                         4        1.93393
                         5        0.63063
                         6       -0.14764
                         7       -0.87249
                         8       -1.59686
                         9       -2.38647
```

```
          Function                                 DESIGN MATRIX
Sample     Number      1     2     3     4     5     6     7     8     9     10
--------------------------------------------------------------------------------
  1          1         1     1     0     0     0     0     0     0     0      1
             2         1     0     1     0     0     0     0     0     0      1
             3         1     0     0     1     0     0     0     0     0      1
             4         1     0     0     0     1     0     0     0     0      1
             5         1     0     0     0     0     1     0     0     0      1
             6         1     0     0     0     0     0     1     0     0      1
             7         1     0     0     0     0     0     0     1     0      1
             8         1     0     0     0     0     0     0     0     1      1
             9         1    -1    -1    -1    -1    -1    -1    -1    -1      1
  2          1         1     1     0     0     0     0     0     0     0     -1
             2         1     0     1     0     0     0     0     0     0     -1
             3         1     0     0     1     0     0     0     0     0     -1
             4         1     0     0     0     1     0     0     0     0     -1
             5         1     0     0     0     0     1     0     0     0     -1
             6         1     0     0     0     0     0     1     0     0     -1
             7         1     0     0     0     0     0     0     1     0     -1
             8         1     0     0     0     0     0     0     0     1     -1
             9         1    -1    -1    -1    -1    -1    -1    -1    -1     -1
```

```
                     ANALYSIS-OF-VARIANCE TABLE

         Source                     DF   Chi-Square      Prob
         ----------------------------------------------------
         INTERCEPT                   1        73.89    0.0000
         _RESPONSE_                  8       718.04    0.0000
         PHENO                       1         5.48    0.0192
         RESIDUAL                    8         4.75    0.7835
```

```
                  ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

                                         Standard     Chi-
Effect        Parameter  Estimate        Error        Square           Prob

---------------------------------------------------------------------------
INTERCEPT             1    1.0382        0.1208         73.89        0.0000
_RESPONSE_            2    2.9725        0.2656        125.26        0.0000
                      3    2.1580        0.1511        204.02        0.0000
                      4    1.8361        0.1243        218.06        0.0000
                      5    0.9614        0.0945        103.58        0.0000
                      6   -0.1554        0.0885          3.08        0.0792
                      7   -1.0428        0.0902        133.61        0.0000
                      8   -1.6358        0.0937        304.98        0.0000
                      9   -2.1206        0.0998        451.16        0.0000
PHENO                10    0.2300        0.0982          5.48        0.0192
```

## PREDICTED VALUES FOR RESPONSE FUNCTIONS

| | | -------Observed-------- | | --------Predicted-------- | | |
|---|---|---|---|---|---|---|
| | Function | | Standard | | Standard | |
| Sample | Number | Function | Error | Function | Error | Residual |
| 1 | 1 | 4.15888308 | 0.35630482 | 4.24069836 | 0.33545855 | -0.0818153 |
| | 2 | 3.32822836 | 0.23989078 | 3.42617757 | 0.22710874 | -0.0979492 |
| | 3 | 3.0285221 | 0.2090043 | 3.10430852 | 0.19600199 | -0.0757864 |
| | 4 | 2.19722458 | 0.14617634 | 2.22958043 | 0.13692893 | -0.0323559 |
| | 5 | 1.11923158 | 0.10180138 | 1.11280246 | 0.09796929 | 0.00642912 |
| | 6 | 0.20844376 | 0.08818257 | 0.22540855 | 0.08671575 | -0.0169648 |
| | 7 | -0.3655556 | 0.08917491 | -0.3675762 | 0.08765506 | 0.00202059 |
| | 8 | -0.8290492 | 0.09534959 | -0.8523541 | 0.09351136 | 0.02330488 |
| | 9 | -1.6900503 | 0.12092801 | -1.7050802 | 0.11720611 | 0.0150299 |
| 2 | 1 | 4.54329478 | 1.00530508 | 3.78060518 | 0.37031224 | 0.76268961 |
| | 2 | 3.83945231 | 0.71466964 | 2.96608438 | 0.27571165 | 0.87336793 |
| | 3 | 3.12456515 | 0.51087084 | 2.64421533 | 0.25053362 | 0.48034981 |
| | 4 | 1.93393396 | 0.308839 | 1.76948724 | 0.20748665 | 0.16444671 |
| | 5 | 0.63062682 | 0.215481 | 0.65270927 | 0.18662758 | -0.0220824 |
| | 6 | -0.147636 | 0.20575499 | -0.2346846 | 0.18434808 | 0.08704863 |
| | 7 | -0.8724881 | 0.22503257 | -0.8276694 | 0.18688966 | -0.0448188 |
| | 8 | -1.5968591 | 0.27415001 | -1.3124472 | 0.19183472 | -0.2844119 |
| | 9 | -2.3864666 | 0.36945129 | -2.1651733 | 0.20695884 | -0.2212932 |

# ANALISIS ESTADISTICO EN BIOLOGIA MOLECULAR: USOY APLICACION EN POBLACIONES VEGETALES

Orlando Martínez Wilches[1]

## RESUMEN

Cada vez que los investigadores de las ciencias básicas biológicas producen, innovan y proponen métodos y técnicas que describen, cualquiera que sea, la variabilidad de las poblaciones naturales y experimentales, es necesario analizar, revisar y evaluar los procedimientos estadísticos disponibles que se adecúan a tales circunstancias; ó bien si se requiere, desarrollar técnicas bioestadisticas alternas para el análisis e interpretación de los resultados provenientes de experimentos, que involucran los nuevos métodos biológicos.

En el caso de la Agronomia, la biología molecular y disciplinas afines han presentado reciéntemente los métodos de isoenzimas, RFLPS y los RAPDS, para determinar la variabilidad, composición y estructura genética de individuos, poblaciones naturales y experimentales. Se analiza y discute el uso de las distancias genéticas, índices de similitud, dendogramas y escalas multidimensionales como técnicas estadísticas para experimentos agronómicos que usan isoenzimas, RFLPS y RAPDS como marcadores genéticos.

[1] Profesor titular. Facultad de Agronomía. Universidad Nacional de Colombia, Bogotá.

# STATISTICAL ANALYSIS IN MOLECULAR BIOLOGY: USE AND APPLICATIONS IN AGRONOMIC POPULATIONS

## SUMMARY

When the researches in basic biological sciences propose, produce or introduce methods and techniques that describe the variability of natural or experimental populations, it is necessary to review, analyze and evaluate the statistical procedures available and adequate for these circumstances. Sometimes it is required to develop statistical techniques for the analysis and interpretation of the data originated in experiments involving biological innovations.

In agronomic research, the molecular biology and similar disciplines have proposed the isoenzymes, RFLP'S and the RAPDS to evaluate the variability, composition and genetic structure of natural and domesticated populations.

In this review, it is discussed and described the use of genetic distances, coefficients of similarity, dendograms and multidimensional scaling as statistical techniques in agronomic experiments which use isoenzymes RFLP'S and RAPDS as genetic markers.

## INTRODUCCION

McCalla (1994) señala cuatro grandes tendencias de la agricultura en los últimos años así: La interdependencia global e integral de los países por el mercado de bienes y servicios; el desarrollo acelerado de las comunicaciones y la información tecnológica en la agricultura tanto a nivel de productor como en las negociaciones de las multinacionales; el consenso mundial y la preocupación de los países por la ecología y el medio-ambiente donde los recursos naturales disponibles ya son finitos; finalmente, la revolución de la biología molecular y su acelerado desarrollo en los últimos 20-30 años. Esta disciplina y otras afines a ella, han ampliado el conocimiento de la genética, evolución y el funcionamiento de los organismos biológicos. Inicialmente se preveía, que mediantes estas técnicas biotecnológicas se obtendría una rápida transformación de la agricultura. Sin embargo, tales observaciones estaban sobre estimadas y se considera que nos encontramos en los primeros estados del impacto y aplicación que estas tecnologías puedan causar en el desarrollo y la productividad agrícola de los países. Los próximos años se preveé serán promisorios y éxitosos.

## TECNICAS ESTADISTICAS MOLECULARES:

Los métodos y procedimientos estadísticos disponibles para el análisis de los resultados provenientes de ensayos biotecnológicos, se pueden agrupar en las siguientes categorías:

1. Aquellos que tienen como propósito el de evaluar la variabilidad, clasificación, estructura y composición genética de las poblaciones,

2. Los desarrollados para la construcción de mapas cromosómicos ó genómicos, cuando se utilizan marcadores genéticos moleculares, y

3. Lo denominados QLT (Quantitative trait loci), que son loci asociados con caracteres cuantitativos de importancia económica, como el rendimiento, y que proveen al fitomejorador de una herramienta molecular ágil, precisa y oportuna de selección indirecta por los caracteres cuantitativos de interés envueltos en el programa de fitomejoramiento.

Este escrito solo se ocupa de los primeros, es decir, de aquellos que en general describen la variabilidad genética de las poblaciones. En particular, se enfatiza su uso en poblaciones, que convencionalmente se reconocen como "recursos genéticos naturales", las cuales son indispensables, como su nombre lo indica, para el desarrollo y progreso futuro de la agricultura.

## MARCADORES GENETICOS

Son numerosas las variables, los caracteres ó parámetros, que se han utilizado para observar y detectar la variabilidad presente en los seres vivos. Los marcadores genéticos son una clase de estos y con ellos se espera que reflejen la variabilidad debida principalmente a los genes.

Los marcadores morfológicos - cuantitativos se consideran como el resultado de los efectos combinados de muchos genes y el ambiente p.e. altura de planta, número de pétalos, longitud de la mazorca. Para su evaluación se requiere de una medida, conteo ó calificación.

Los marcadores bioquímicos, están constituidos por las isoenzimas y las proteinas. Mediante la técnica de la electroforesis en gel, se hace posible el estudio de la variación de las proteinas y enzimas en organismos vivos así: Las muestras de tejidos se homogenizan (muelen) para liberar las enzimas y proteinas de las células. El sobrenadante del homogenizado (parte liquida), se coloca en un gel de almidón, agar, poliacrilamida ó alguna sustancia gelatinosa. El gel se somete durante horas a corriente eléctrica continua, cada proteína del gel migra en una dirección y velocidad que depende de la carga eléctrica neta de la proteína y del tamaño molecular. Después el gel se trata con una solución química con un sustrato específico para la enzima en estudio y una sal que produce una mancha (banda) coloreada, que refleja la migración de la enzima. La utilidad del método radica en el hecho de que el genotipo del locus genético que codifica la enzima puede ser inferido a partir del número y posiciones de las bandas observadas en los geles (Ayala y Kiger 1984).

Los marcadores moleculares de mayor uso en la detección de la variabilidad genética, lo constituyen los RFLPS y los RAPDS. Los RFLPS, son una clase de enzimas, llamadas enzimas de restricción. Son nucleasas producidas por diferentes microorganismos y tienen la capacidad de reconocer ciertos sitios (sitios de restricción) constituidos por secuencias de bases específicas en el ADN. Si una secuencia especifica de bases está presente en el sitio de restricción, la enzima de restricción corta al ADN en ese sitio. Por lo tanto, una cadena larga de ADN se puede reducir a una serie de fragmentos de tamaño finito según el corte de la enzima de restricción. El número de fragmentos producidos y el tamaño de cada fragmento refleja los sitios de restricción en la cadena del DNA. Los fragmentos de restricción producidos por el corte de la endonucleasa (p.e. Hind III) de un tejido, se someten al proceso de electroforesis en agar; los fragmentos migran con la presencia de la corriente eléctrica y la velocidad de migración depende del peso molecular de cada fragmento. Posteriormente, el gel se colorea con bromuro de etidio y el patrón de migración de los fragmentos se observa directamente mediante manchas coloreadas de una manera similar a las isoenzimas y proteinas (Kochet 1994).

Los marcadores moleculares conocidos como RAPDS ó AP-PCR tienen como base la reacción en cadena de la polimerasa (una enzima, que bajo ciertas circunstancias produce replicas de cadenas sencillas de ADN). Los RAPDS (segmentos, amplificados, aleatorios de ADN) es una técnica para estudiar la variabilidad genética, la cual permite la detección de secuencias polimórficas de ADN, utilizando cebadores (Primers) sencillos con secuencias arbitrarias de oligonucleótidos. Las secuencias se amplifican o se generan con la información ADN del tejido de la especie en estudio y mediante la reacción en cadena de la polimerasa. Al igual que las isoenzimas, el material procesado se somete a electroforesis en agar y los segmentos amplificados migran por la acción de la corriente eléctrica, la velocidad de migración depende de su peso molecular. Después el gel se colorea con bromuro de etidio y el patrón de la migración de los segmentos de ADN, se observa directamente mediante manchas coloreadas (Williams et al 1990, Welsh and Mcclelland 1991).

## CUANTIFICACION DE LOS MARCADORES BIOQUIMICOS Y MOLECULARES

Los resultados experimentales de un ensayo biológico donde se utilicen las proteinas, enzimas, RFLPS o RAPDS es el mismo: un conjunto de bandas coloreadas en el gel que representan el comportamiento de la variabilidad. Como ilustración, considere 5 colecciones de una especie agrícola, p.e. cacao, las que se sometieron a un estudio de diversidad enzimática. En la Figura 1 se presentan los resultados correspondientes a una corrida de la $\alpha$-$\beta$ esterasa; en la figura se observa el patrón (las bandas) de variación de las colecciones, la última columna corresponde al estandard el cual expresa todas las bandas posibles producidas por las cinco colecciones. El problema es cómo cuantificar las bandas y una vez cuantificadas proponer medidas estadísticas que expresen la variabilidad entre las colectas en estudio.

Las bandas de la Figura 1 se pueden cuantificar mediante una función indicadora, esto es, asignar el valor 1 si la banda está presente y cero si no lo está. Al aplicar dicha función al ejemplo de la α-β esterasa, se obtiene la Tabla 1; ella refleja la variabilidad de las bandas pero ya de una forma cuantitativa, numérica, a la cual se le pueden proponer medidas estadísticas que expresen la diversidad enzimática entre las colectas en estudio.

**FIGURA 1.** PATRON DE VARIABILIDAD DE CINCO COLECCIONES DE CACAO ASOCIADOS CON LA α-β ESTERASA

| | COLECCIONES | | | | | |
|---|---|---|---|---|---|---|
| **ORDEN** | **A** | **B** | **C** | **D** | **E** | **ESTAND.** |
| 1 | | - | | - | | - |
| 2 | - | | - | - | | - |
| 3 | - | - | - | - | | - |
| 4 | | - | | | | - |
| 5 | - | - | - | | - | - |
| 6 | - | - | - | | - | - |
| 7 | | - | | | - | - |
| 8 | - | | - | - | | - |
| 9 | - | | - | - | | - |
| 10 | | | - | - | | - |

**TABLA 1.** CUANTIFICACION DE LA α-β ESTERASA EN CINCO COLECCIONES DE CACAO.

| | COLECCIONES | | | | | |
|---|---|---|---|---|---|---|
| ORDEN | A | B | C | D | E | ESTAND. |
| 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 2 | 1 | 0 | 1 | 1 | 0 | 1 |
| 3 | 1 | 1 | 1 | 1 | 0 | 1 |
| 4 | 0 | 1 | 0 | 0 | 0 | 1 |
| 5 | 1 | 1 | 1 | 0 | 0 | 1 |
| 6 | 1 | 1 | 1 | 0 | 1 | 1 |
| 7 | 0 | 1 | 0 | 0 | 1 | 1 |
| 8 | 1 | 0 | 1 | 1 | 1 | 1 |
| 9 | 1 | 0 | 1 | 1 | 0 | 1 |
| 10 | 0 | 0 | 1 | 1 | 0 | 1 |

## INDICES O COEFICIENTES DE SIMILITUD:

Una medida de semejanza para comparar dos colecciones (la A y la B), utilizando los resultados de la Tabla 1, sería aquella que relacionara el número de bandas (unos o ceros) que simultáneamente compartan las dos colectas. La siguiente tabla provee de la información necesaria para relacionar las ausencias y presencias comunes entre el par de colecciones.

| | | B | |
|---|---|---|---|
| | | 1 | 0 |
| A | 1 | a | b |
| | 0 | c | d |

$n = a+b+c+d$

Dos medidas de semejanza ($S_{AB}$) entre A, B serian:

$$S_{AB} = a/n$$
$$S_{AB} = (a+d)/n$$

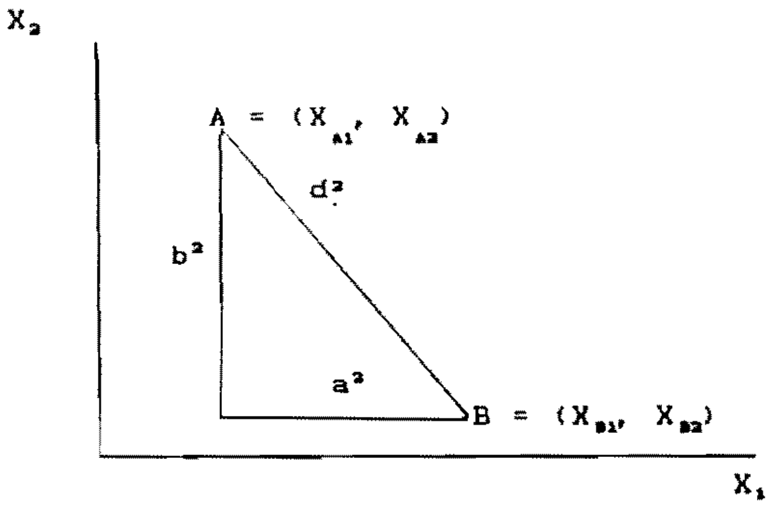
se puede entonces calcular una tabla (matriz) de coeficientes de similitud entre todas las colecciones.

Adicionales a las anteriores, se han propuesto diferentes índices de similitud. En la tabla 2 se expresan los más comunes, su interpretación y el autor. Estos índices fueron originalmente creados para estudios de poblaciones de insectos, ecología y en la especie humana donde la presencia y ausencia de características es común al evaluar el comportamiento ante una serie de estímulos cognocitivos.

TABLA 2: Coeficientes de Similitud

| | COEFICIENTE | INTERPRETACION | AUTOR |
|---|---|---|---|
| 1. | $\frac{a+d}{n}$ | Igual peso a 0-0 y 1-1 | Sokal, Michener 1958 |
| 2. | $\frac{a}{a+b+c}$ | No contabiliza 0-0 | Jackard, 1908 |
| 3. | $\frac{2a}{2a+b+c}$ | Doble peso a 1-1 no contabiliza 0-0 | Dice, 1945 |
| 4. | $\frac{2a}{b+c}$ | | Nei, 1987 |
| 5. | $\frac{a+d}{a+d+2(b+c)}$ | Doble peso a 0-1. 1-0 | Rogers y Tamimoto, 1960 |
| 6. | $\frac{a}{n}$ | No. 0-0 en numerador | |
| 7. | $\frac{a}{a+2(b+c)}$ | No. 0-0 en numerador doble peso a 100 y 0-1 | |
| 8. | $\frac{a}{b+c}$ | 0-0 excluído, relaciona presencia con ausencia | |

**DISTANCIAS:** Euclideana - Geométrica - Genética. La distancia Euclideana entre dos colectas es la aplicación del teorema de Pitagoras así:



$$D^2_{AB} = \quad d^2 = a^2 + b^2$$

$$D^2_{AB} = \quad (X_{A1} - X_{B1})^2 + (X_{A2} - X_{B2})^2$$

$$D^2_{AB} = \quad \Sigma_k ( X_{Ak} - X_{Bk} )^2$$

La distancia euclideana es intuitivamente atrayente, fácil de entender, es una medida geométrica que posee numerosas características algebráicas - matemáticas, de allí su amplio uso en investigaciones en las ciencias biológicas, económicas y sociales.

La distancia genética es una medida que expresa la divergencia, entre dos poblaciones, razas ó colectas, divergencia atribuible exclusivamente a genes ó a conjuntos de los mismos. Si pi es la frecuencia del i-ésimo gene de la población A y qi lo es para la población B entonces una medida de distancia genética entre A y B es la distancia euclideana aplicada a la frecuencia de los genes así:

$$D^2_{AB} = \Sigma_i ( p_i - q_i)^2$$

Se han propuesto diferentes medidas de distancia genética como son la de Rogers, Prevosti, Cavalli-sforza, Nei etc.; para su construcción se ha considerado aspectos geométricos, matemáticos y biológicos entre otros (Nei 1987).

**DISTANCIAS Y SIMILARIDADES.** Los índices ó coeficientes de similaridad, son medidas de semejanza entre bandas electroforéticas; algunos de ellos están relacionados con las distancias, fmediante funciones algebraicas. Es decir, bajo ciertas circunstancias es posible calcular distancias euclideanas a partir de los índices de similitud. Entre las expresiones que relacionan los coeficientes y las distancias se encuentran.

$$D^2{}_i = 1 - 2S_i$$

$$D^2_i = 2 (1 - S_i)$$

$$D^2_i = 1 \div (1 - S_i)$$

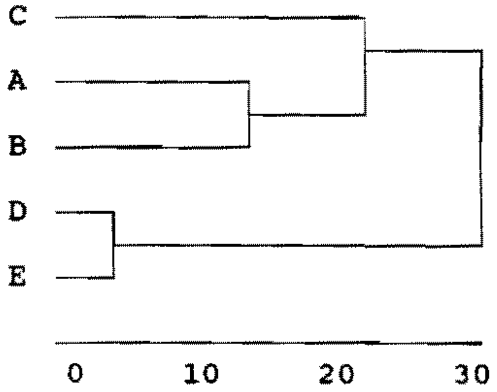
Sin embargo, tal como lo muestra Gower (1966, 1967) no siempre es posible calcular distancias euclideanas a partir de similaridades. La matriz de similaridad tiene que ser definida semipositiva para lograr la conversión. De las similaridades expresadas en la Tabla 2 solamente las definidas por Sokal y Michener (1958) y Jacckard (1908) poseen esta condición. Gower (1966) también enfatiza en usar la expresión $2(1 - S_i)$.

Hasta ahora, para cuantificar el patrón de bandas electroforéticas se han propuesto diferentes coeficientes de similaridad, distancias genéticas, geométricas y euclideanas. Sin embargo, cuando se estudian varias poblaciones, p.e. 20 que es un tamaño más bien intermedio, el número total de distancias entre pares de poblaciones sería de (20 x 19)/2 = 190. Por lo tanto, se torna dispendioso resumir estas 190 distancias y a partir de ellas realizar las inducciones y deducciones poblacionales. El siguiente paso, es entonces el manejo de una matriz de distancias y con ella hacer inferencias estadísticas. Se discuten los dendogramas y las coordenadas principales, que son dos métodos gráficos - estadísticos que proveen de buenas guías al investigador que usa marcadores moleculares y bioquímicos en la descripción de la variabilidad de poblaciones biológicas. Los dendogramas y las escalas multidimensionales resumen la matriz de distancias.

**DENDOGRAMAS - CONGLOMERADOS**: El propósito fundamental del análisis de conglomerados es el de proveer al investigador de "agrupaciones naturales" de un conjunto de individuos, razas, ó variedades. Se busca colocar conjuntos de individuos en grupos exhaustivos y mutuamente excluyentes, de tal forma que se puedan hacer inferencias estadísticas de semejanzas ó diferencias en y entre los grupos provistos por el análisis.

Los grupos establecidos por el análisis forman particiones y subdivisiones en conjuntos menores ó reagrupamientos en mayores y eventualmente se puede finalizar con una estructura jerárquica de agrupamiento. A esta estructura se le conoce como jerarquización en árbol.

La estructura jerárquica de agrupamiento ó la estructura en árbol, se puede representar en un diagrama ó figura bidimensional y a tal representación se conoce como dendograma así p.e.



Los dendogramas, en general, se construyen a partir de una matriz p x p de distancias ó de coeficientes de similitud. Entonces las p(p-1)/2 posibles distancias ó similitudes obtenidas de p poblaciones se condensan en el dendograma, lo que facilita y simplifica enormemente las inferencias de semejanza ó disimilitud entre los diferentes grupos y subgrupos de poblaciones en estudio.

Por lo tanto, al patrón de bandas electroforéticas provenientes de las isoenzimas, los RFLPS ó los RAPDS, se les ha provisto de métodos estadísticos formales (distancias, similitudes, dendogramas) de tal manera que su variabilidad, bioquímica, molecular y en general genética se puede discriminar y cuantificar.

## ESCALAS MULTIDIMENSIONALES

**COORDENADAS PRINCIPALES.** Es un conjunto de técnicas estadísticas -matemáticas, para encontrar una configuración de puntos a partir de una matriz de distancias. Para usar el escalamiento multidimensional se requiere necesariamente que las distancias sean euclideanas.

Como ilustración de la técnica considere el siguiente ejemplo: suponga un mapa de Colombia y un conjunto de ciudades; se solicita construir una tabla (matriz) de distancias entre las ciudades; simplemente con una regla se medirían las distancias en el mapa y luego se convertirían a distancias reales en kilómetros por ejemplo. Ahora considere el problema inverso: dada una matriz de distancias entre las ciudades construya el mapa (las coordenadas).

En primer término, dado un conjunto de distancias euclideanas, no existe una representación única de puntos que origine las distancias; así, si conocemos la distancia entre Cali - Ibagué no sabemos si Cali está al oriente - occidente - norte ó sur de Ibagué. Técnicamente significa que no conocemos la localización y orientación de la configuración. El problema de localización se resuelve colocando el centro de gravedad de la configuración en el orígen. El problema de orientación se resuelve mediante una transformación ortogonal, de tal forma que los ángulos y distancias no se modifiquen.

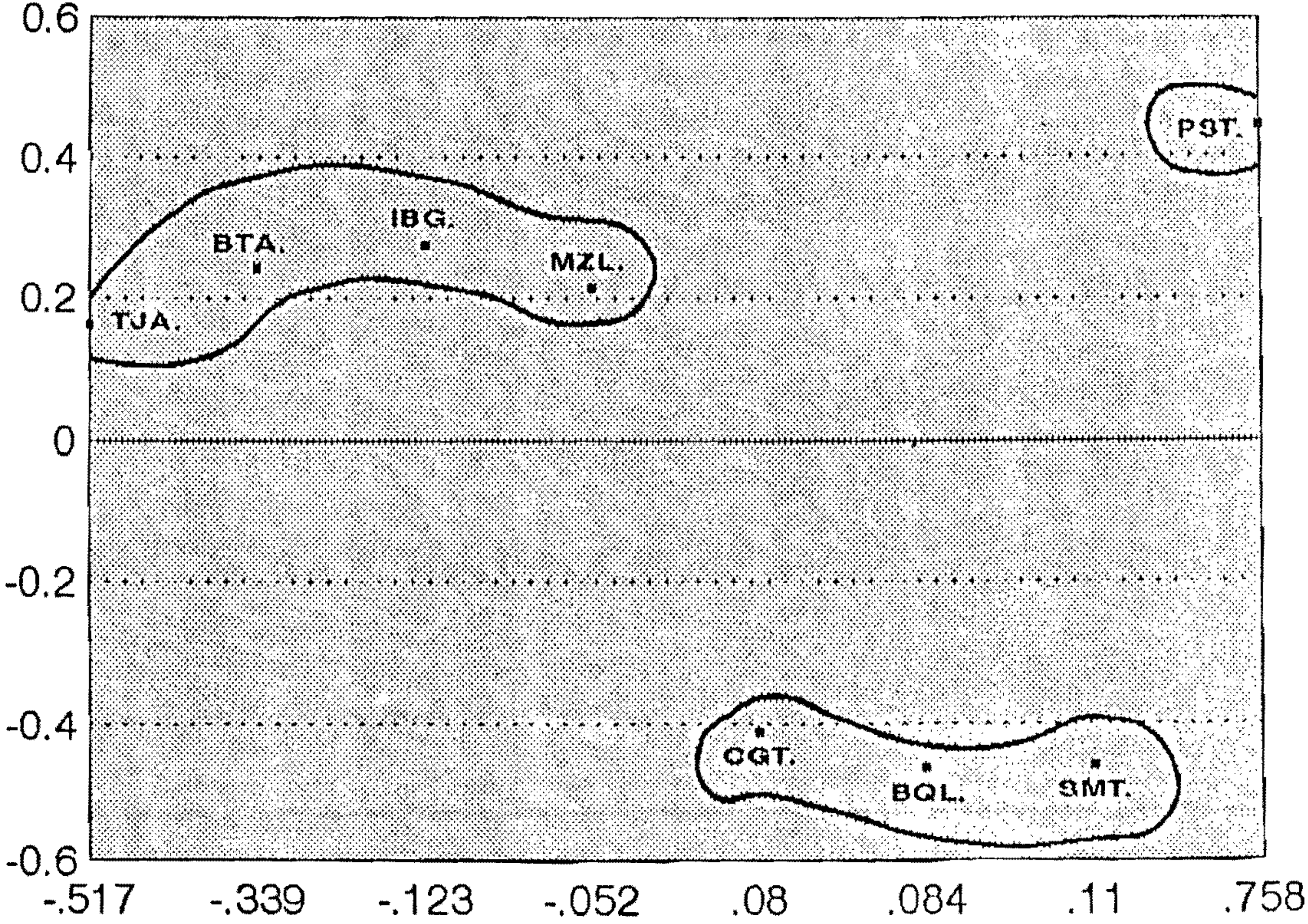
La aplicación de esta técnica estadística a los datos provenientes de ensayos biotecnológicos agrícolas, es inmediata ya que a partir de las bandas electroforéticas, se construyen índices de similaridad y con estos distancias euclideanas, a las cuales se aplican las escalas multidimensionales, para encontrar un plano de coordenadas principales donde las relaciones de semejanza y divergencia entre poblaciones biológicas se discriminan y cuantifican con cierto grado de sencillez.

A continuación, como ejemplo, se presenta una matriz de distancias entre algunas ciudades de Colombia y el uso de las escalas multidimensionales para reproducir un mapa (las coordenadas) de las ciudades. Se observa que el mapa preserva la distancia y localización de las ciudades. En la figura siguiente se observa la aplicación de las escalas multidimensionales a una colección de café, con datos provenientes de experimentos con RAPDS.

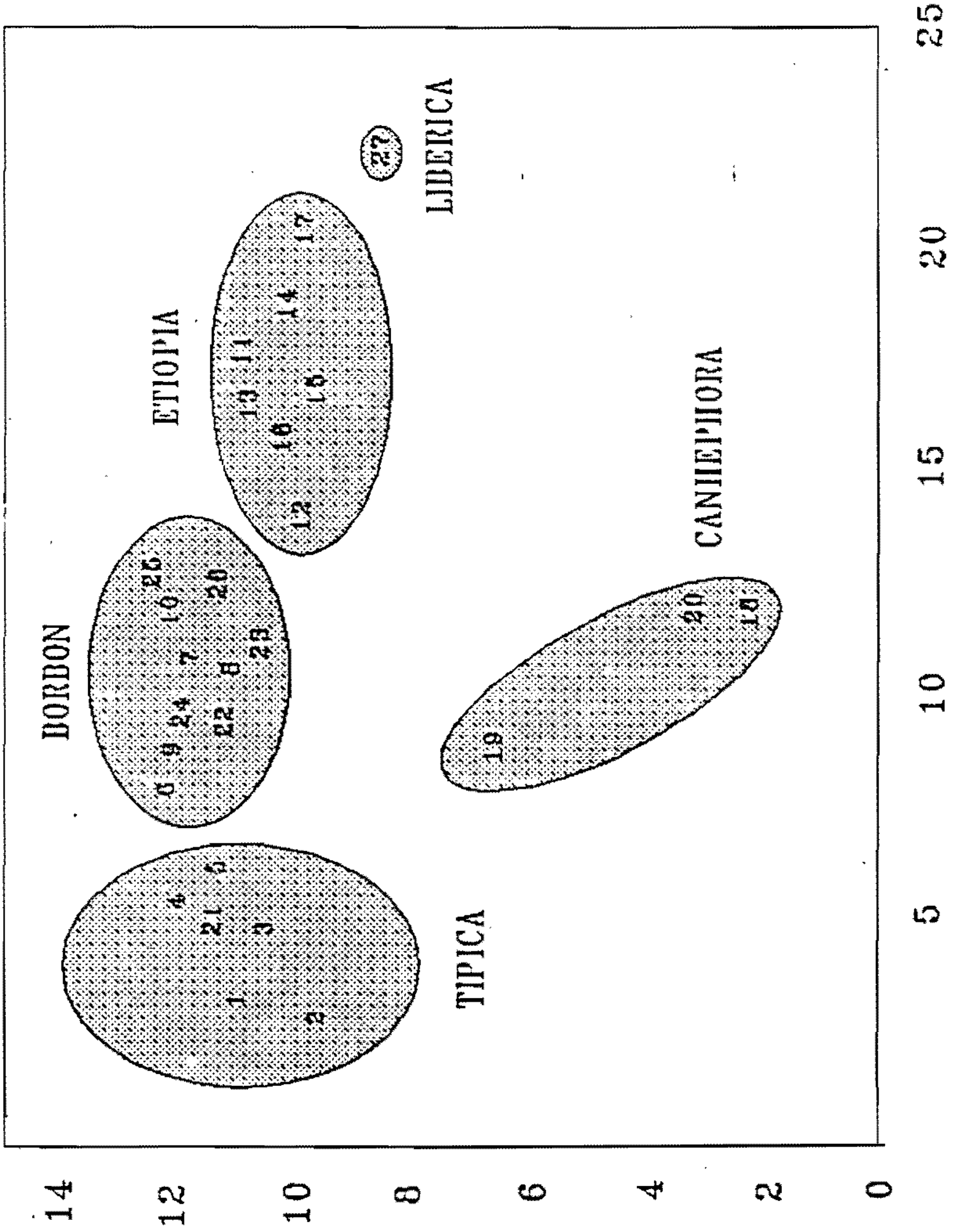
$$\begin{array}{l|llllllll}
\mathrm{BTA} & 0 & & & & & & & \\
\mathrm{IBG} & 1.2 & 0 & & & & & & \\
\mathrm{MZL} & 1.7 & 0.7 & 0 & & & & & \\
\mathrm{PST} & 2.6 & 2.2 & 2.4 & 0 & & & & \\
\mathrm{CTG} & 3.4 & 3.2 & 2.9 & 5.2 & 0 & & & \\
\mathrm{BQL} & 3.6 & 3.7 & 3.4 & 5.5 & 0.5 & 0 & & \\
\mathrm{STA} & 3.7 & 3.8 & 3.5 & 5.7 & 0.8 & 0.7 & 0 & \\
\mathrm{TJA} & 0.7 & 1.2 & 1.4 & 3.4 & 2.9 & 3.2 & 3.3 & 0
\end{array}$$

DISTANCIA ENTRE CIUDADES DE COLOMBIA

# MSD: COLOMBIA



COORDENADAS PRINCIPALES RAPDS



# BIBLIOGRAFIA

1. Ayala, J.F. y J.A. Kiger. 1984. Genética moderna.

2. Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis Biometrika: 53:325-328.

3. Kochet, G. 1994. Introduction to RFLP mapping and plant breeding applications.

4. McCalla, A.F. 1994. Priorites and Problems: the challenges facing world agriculture. V International Congress for Computer Technology in Agriculture. Royal Agricultural Society of England.

5. Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, N.Y.

6. Sokal, R.R. and Sneath, P.H.A. 1963. Principles of numerical taxonomy, London: Freeman.

7. William, J.G.K. et al. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers Nucleic Acids Research. 18-6531-6535.

8. Welsh, J. and M. McClelland. 1970. Finger printing genomes using PCR with arbitrary primers. Nucleic Acids Research. 18:7213-7218.

# ANALYSIS OF BLAST RESISTANCE IN RICE, COMPARING BIOTECHNOLOGY METHODS WITH TRADITIONAL BREEDING

*M.C. Amézquita[1], C.P. Martínez[2], F. Correa-Victoria[2], and G. Lema[1]*
*[1] Biometry Unit and [2] Rice Program*
*CIAT, Cali, Colombia*

Paper presented at the "International Symposium on Statistics in Agriculture and Environmental Research", Satellite Conference, CIAT, Cali, Colombia, June 7-9, 1995.

## SUMMARY

The relatively recent developments on statistical methods for the analysis of categorical variables allows parameter estimation and hypothesis testing using linear models on functions of response frequencies. A particular case is the Logit Model, which uses as dependent variable the logit function, expressed as ln ($\pi$/1-$\pi$), where $\pi$ is the proportion of success on a binary variable, and external factors as independent variables. The purpose of this paper is to illustrate an application of the logit model to compare biotechnology methods (Anther Culture) with the traditional breeding method (the Pedigree Method) in their efficiency to generate rice lines with stable resistance to blast, the most important disease of rice world-wide. An initial number of 17917 $F_2$ plants submitted to selection for blast resistance through the Pedigree Method (PM), together with 441 and 740 lines obtained by anther culture methods (AC-$F_2$ and AC-$R_2$) respectively, were evaluated for blast resistance and resistance stability. The analysis confirms that AC was more efficient than PM across cross-types in generating blast resistant lines; however, in terms of resistant stability, AC was superior to PM only on a given cross-type. Model assumption are discussed in the light of this particular study. General recommendations of methodological nature are made tending to guarantee valid statistical inferences.

## RESUMEN

Gracias a desarrollos relativamente recientes en métodos estadísticos para el análisis de datos categóricos, es posible realizar estimación de parámetros y prueba de hipótesis sobre variables de respuesta categóricas ajustando modelos lineales a funciones de las frecuencias de respuesta. El modelo Logit aplicado a una variable binaria, es un modelo lineal que usa como variable dependiente la función logit –ln($\pi$/1-$\pi$), siendo $\pi$ la proporción de 'éxito'– y como variables independientes uno o más factores cuyo efecto se desea probar. El objetivo de este trabajo es ilustrar el uso de este método estadístico para comparar el mejoramiento de arroz por el Método de Pedigree (MP) con el de Cultivo de Anteras realizado sobre plantas $F_2$ (CA-$F_2$) y sobre plantas $F_1$(CA-$R_2$) , en términos de su capacidad para generar líneas con resistencia estable a Piricularia. Una línea se consideró resistente si tenía calificación de reacción a Piricularia $\leq$4 en la escala 0-9. Una línea se consideró poseer 'resistencia estable' cuando mantuvo su resistencia a través de un ciclo de evaluación de 3 semestres. El análisis se realizó con 17917 plantas $F_2$ provenientes de 3 tipos de cruza que se sometieron a selección por el MP hasta obtener 681 líneas $F_4$ resistentes; y con 441 y 740 líneas de las mismas cruzas obtenidas por CA-$F_2$ y CA-$R_2$ respectivamente, de las cuales se lograron 64 y 67 líneas resistentes. Los tres conjuntos de líneas ($F_4$, CA-$F_2$ y CA-$R_2$) se evaluaron por 3 semestres consecutivos para determinar la estabilidad de su resistencia. El Modelo Logit detectó diferencias significativas a favor del método de Cultivo de Anteras en términos de resistencia a piricularia, a través de los 3 tipos de cruza. En cuanto a resistencia estable, CA fue superior en un tipo de cruza unicamente. Los supuestos del método y su aplicación específica a este caso, permitió hacer recomendaciones metodológicas tendientes a garantizar validez en la inferencia.

# ANALYSIS OF BLAST RESISTANCE IN RICE COMPARING BIOTECHNOLOGY METHODS WITH TRADITIONAL BREEDING

M.C. Amézquita[1], C.P. Martínez[2], F. Correa-Victoria[2], and G. Lema[1].

[1] Biometry Unit and [2] Rice Program, CIAT, Cali, Colombia

## Background

### *The rice blast disease. Breeding methods*

Rice blast, caused by the fungus *Pyricularia grisea* Sacc. is considered to be the single most important disease of rice world wide. Both the disease and the pathogen have been extensively studied. Development of resistant cultivars has been the most effective method to control this disease. Traditional breeding using the Pedigree Method (PM) has been extensively used (Rosero, 1979; Weeraratne et al., 1981; Roumen, 1992; Cuevas Pérez et al., 1992; Correa-Victoria & Zeigler, 1993a and 1993b). Initial success with blast resistant cultivars has been obtained; however, in many instances the resistance appeared to be ephemeral as new races of the fungus developed (Weeraratne et al., 1981; Correa-Victoria & Ziegler, 1993a and 1993b). More recently, Levy et al., (1991 and 1993), Correa and Zeigler, (1993b), suggested a breeding strategy for developing durable blast resistance by identifying combinations of genes that show complementary resistance to all the virulence spectrum observed within different genetic lineages of the blast pathogen. The production of doubled-haploids (Dhs) through a biotechnology method --anther culture (AC)-- has been proposed as an effective, efficient and economic breeding tool (Chu, 1982; Baenziger & Schaeffer, 1983; Caligari et al., 1987; Baenziger et al., 1989; Lynch et al., 1991; Sanint et al., 1993). Several studies in different crop species (Friedt et al., 1986; Pauk et al., 1988; Picard et al., 1988; Huang et al., 1988) compared the performance of AC derived lines with other breeding procedures in terms of yield and other characters. Most of these studies showed that the AC method was easier, faster and produced the same genetic variability. However, resistance to rice blast has not been considered in these comparisons.

***The measurement and analysis of blast reaction***

Reaction of the rice plant to blast is recorded under a 0-9 discrete scale according to the International System for Rice Evaluation (IRRI, 1988). Both , leave and neck blast reactions are measured in order to asses the degree of plant damage or its resistance/susceptibility to the disease. The first five categories of this scale (0-4) are of qualitative nature, as they reflect different types of lesion under which the plant does not show any evidence of damage. The last five categories (5-9) are quantitative in nature, but non-equally spaced, and they reflect the percentage of leaf area affected.

Very often, rice varieties are characterized by their blast reaction using the mean score. However, for statistical comparisons, the use of traditional ANOVA or other statistical methods for continuous variables is not appropriate given the nature of the measurement scale. When the main purpose is to analyze plant damage, the transformation of the disease-reaction scale into an appropriate "index of disease severity", of continuous nature, such as total plant area affected or percentage of leaf area affected-- is a good approach (Guimaraes et al, 1995). However, when the main purpose is to analyze plant resistance to blast, a dichotomous trait (resistant or not resistant), the conversion of the 0-9 scale into a binary response is necessary. Statistical methods for categorical response variables are useful tools in this case.

Many of the statistical methods for categorical variables are generalizations of those for continuous variables. This is the case of linear model methods, as typified by the Grizzle, Starmer and Koch approach (1969). The emphasis of these methods is estimation and hypothesis testing of the model parameters. Following parameter estimation, hypothesis about linear combinations of the parameters can be tested. Wald Statistics (Wald, 1943), which are approximately distributed as chi-square if the sample sizes are sufficiently large and the null hypothesis are true, are used for hypothesis testing. Linear model methods for categorical variables are a natural extension of the ANOVA approach for continuous variables. For example, analysis of variance, in the traditional sense, refers to the analysis of means and partitioning of variation among the means into various sources. In categorical methods, the term 'analysis of variance' is used in a generalized sense to denote the analysis of response functions and the partitioning of variation among those functions into various sources. The response functions are functions of the observed frequencies in the contingency table under analysis and incorporate the essential information from the response variable. Example of response functions are mean scores, marginal probabilities, cumulative logits, generalized logits (Agresti, 1984; Freeman, 1987; Agresti, 1990). A particular case of generalized logits, when the categorical variable is binary, is the logit function, expressed as $\ln (\pi/1-\pi)$, where $\pi$ is the proportion of 'success'.

For this particular study, two variables of interest, --both of binary nature-- were generated, based on leaf and neck blast reaction scores: general blast resistance and resistance stability. In the case of general blast resistance, its two levels are: 'resistant', with general blast reaction (GBL[1]) $\leq 4$, and 'non-resistant', with GBL $\geq 5$. For the second response variable, resistance stability, its two levels are: 'stable resistant', when the line maintains its resistance throughout a given evaluation period --a 3-semester cycle in this study--, and 'non stable resistant' if otherwise. Both can be analyzed using linear models on functions of response frequencies.

The objective of this paper is to illustrate the statistical methodology adopted by the authors to compare AC methods with the traditional breeding method, PM, in their capacity to produce rice lines with blast resistance and resistance stability.

[1] GBL is expressed for each line as the maximum score between leaf and neck blast reactions.

### Genetic Material. Data Source

Three types of crosses were made between the rice cultivar Fanny (Japonica and highly susceptible to blast) and 11 cultivars (either Japonica or Indica) with different degree of resistance/susceptibility to blast. The eleven crosses were:

***Japonica/Japonica - susceptible x resistant*** (6): CT5782 (Fanny/IRAT13), CT8813 (Fanny/TOX1011-4-1), CT8816 (Fanny/0S6), CT8817 (Fanny/LAC23), CT8819 (Fanny/IAC165), CT8820, (Fanny/ITA235).

***Japonica/Indica - susceptible x resistant*** (2): CY8814 (Fanny/Ceysvoni), CT8818 (Fanny/Carreon)

***Japonica/Indica - susceptible x susceptible*** (3): CT5780 (Fanny/CICA4), CT8821 (Fanny/Colombia1), CT8815 (Fanny/Tetep).

Using the pedigree method, beginning in 1988, an initial number of 17.917 $F_2$ plants (1500 plants/cross approx.) were produced. $F_3$, and $F_4$ generations were evaluated and selected for blast resistance through subsequent semesters at Santa Rosa Experimental Station, located in Villavicencio, Colombia. A total of 681 $F_4$ blast resistant lines were obtained and subjected to a 3-semester evaluation cycle to determine stability of blast resistance (table 1) under high disease pressure. Panicles collected from $F_1$ and $F_2$ blast susceptible plants were used for anther culture to produce the AC-$R_2$ and AC-$F_2$ populations, respectively, according to the methodology described by Núñez et al., (1989). Thus 441 AC-$F_2$ and 740 AC-$R_2$ Dhs lines were produced. These Dhs were planted at Santa Rosa in 1990 and evaluated for their leaf and neck blast reactions, resulting in 64 AC-$F_2$ and 67 AC-$R_2$ lines with general blast resistance (GBL score $\leq$ 4) (table 1). These lines were subsequently submitted to a 3-semester evaluation cycle to determine stability of blast resistance. Therefore, three different sets of lines ($F_4$, AC-$R_2$, and AC-$F_2$) were planted together, for comparison, during three semesters, and evaluated for blast resistance stability under upland conditions.

**Statistical analysis methodology**

The statistical analysis methodology followed two steps: a) A descriptive analysis to visualize overall performance of genetic populations generated by each of the 3 methods (PM, AC-$F_2$, AC-$R_2$), in terms of both, the original 0-9 blast reaction score and the binary variables, blast resistance and resistance stability. b) An inferential statistical analysis to assess the effect of 'method', 'cross type' and their interaction on blast resistance and resistance stability.

***Descriptive Statistical analysis***. For initial populations as well as for final populations generated by each method, the following descriptive parameters were calculated:

1. For initial populations:
   * Number of lines generated.
   * Skewness, median, mode and range for blast reaction scores
   * Number and proportion of blast resistant lines
2. For final populations being submitted to the evaluation cycle for resistance stability.
   * Number and proportion of blast resistant lines produced.
   * Skewness, median, mode and range for their blast reaction scores.
   * Number and proportion of stable resistant lines produced at the end of the evaluation cycle.
   * Skewness, median, mode and range for their blast reaction scores.

The skewness coefficient defined as $S_k = M_3\surd(\sigma^2)^{3/2}$, where $M_3$ [2], is the third moment of the distribution, indicates the degree of symmetry of the distribution. If $S_k \geq 0.5$, the distribution is left-oriented indicating a higher proportion of resistant lines; if $-0.5 < S_k < 0.5$, the distribution is symmetric; and if $S_k \leq -0.5$, the distribution is right-oriented, indicating a higher proportion of susceptible lines. The mode, indicating the most frequent value in the distribution, and the median,

[2] $M_3 = [f_1(X_1 - X)^3 + f_2(X_2 - X)^3 + ... + f_n(X_n - X)^3] / N$
$\sigma^2$ = variance; X = mean score; $X_n$ = data; $f_n$ = frequency

indicating the value below which 50% of the population exists, served as summary statistics for the populations.

***Inferential Statistical Analysis***. In order to test for the effect of 'method', 'cross type' and their interaction on blast resistance (R) and resistance stability (RS), a linear model was fitted for each binary response, using the logit function as the dependent variable. The models used were:

$$\ln (\pi_R/1-\pi_R) = \mu + \text{Method (M)} + \text{Cross type (C)} + \text{MxC}$$

and, $$\ln (\pi_{RS}/1-\pi_{RS}) = \mu + \text{Method (M)} + \text{Cross type (C)} + \text{MxC},$$

$$where, \quad \sum M_i = 0, \quad \sum C_j = 0, \quad \sum\sum (MxC)_{ij} = 0$$

$\pi_R$ and $\pi_{RS}$ represent the proportion of blast resistant and stable resistant lines, respectively. Estimation of model parameters was performed, using the method of maximum likelihood. Parameter estimates were then used to calculate the probabilities, $p_R$ and $p_{SR}$, of generating resistant or stable resistant lines for each method within a given cross type. Hypothesis on the parameters were tested --comparing AC-$F_2$ and AC-$R_2$ vs. PM across and within cross types.

The contingency table under analysis corresponds in the case of blast resistance, to a (9x2) table with 9 populations (3 methods x 3 cross types) and 2 levels for the response variable. In the case of resistance stability, the contingency table is a (6x2), as cross type JxI - SxS was excluded from the analysis for presenting zero frequencies in anther culture methods (table 3). Assumptions for linear models on categorical variables (Fienberg, 1980; Freeman, 1987) were studied and considered to apply to this particular data set; that is:

* Each population (or row) in the contingency table is assumed to be a simple random sample from the overall population. In this study, the number of lines generated by each method in each cross type are a simple random sample from the entire population of possible lines generated.

* The frequencies in each row of the contingency table follow a binomial distribution with

probability of success $p_R$ or $p_{SR}$ respectively. Therefore, the frequencies in the whole table follow a product binomial distribution.

* In order to support the asymptotic normal distribution of the response functions, it has been suggested that the sample size per population (or row) needs to be of at least 25 for each response function being analyzed and with no more than 20% of the cells having small frequencies, of less than 4 or 5 (see CATMOD procedure, SAS, 1989). As we are dealing with binary variables, that is with one response function, the minimum number of lines per method/cross-type required is 25, which is the case in this study. However, the second requirement is barely met in the case of resistance stability.

All data processing and analysis was performed through SAS version 6.07 (1989), using the CATMOD procedure.

**Results and Recommendations**

Tables 1 and 2 show overall performance of genetic populations generated by each method, across crosses (table 1) and per individual cross (table 2). Descriptive statistics on initial populations confirm two facts which are inherent to this type of research. First, the relatively small number of lines generated through anther culture methods as compared to the large number generated through the Pedigree Method. Second, the susceptibility of initial populations, as indicated by their median and mode (all greater than 4) and the skewness value of their distributions, with a bias towards a higher proportion of susceptible lines in populations generated by anther culture methods. This fact has been cited in the literature (Guiderdoni, 1991) and confirms the hypothesis that rice lines of Japonica origin respond better to anther culture than Indica type lines. Therefore, as all cross types have as parental material the highly susceptible cultivar Fanny, of Japonica origin, then AC-derived populations present a higher proportion of susceptible lines. In spite of this initial bias against AC, inherent to the method under test, the observed proportions of blast resistant and stable resistant lines produced by AC methods seem to suggest their superiority over the pedigree method (table 1).

Table 3 shows the contingency tables under analysis and corresponding logit values.The 'analysis of variance' (table 4) and parameter estimates (table 5) indicate that 'method' and 'cross type' are significant factors in producing blast resistant and stable resistance lines. However, there is a strong interaction among them indicating that the effect of the method depends on the cross-type also.

The probabilities of each method to generate resistant or stable resistant lines (table 6) and comparisons of AC vs. PM across and within cross types (table 7) confirm the superiority of AC methods across cross-types in terms of blast resistance; however only in one cross-type (JxI, SxR) AC was superior to PM in resistance stability. Additionally, the analysis shows that crosses between susceptible and resistant Japonica cultivars (cross type JxJ - SxR) are more efficient.

Some general recommendations of methodological nature result from this analysis: a) In order to guarantee valid statistical inferences, an effort needs to be made to generate when possible, higher numbers of anther-culture derived lines, either by using more crosses of the same cross type or by improving the procedures involved in anther culture technique. b) It is suggested to identify a-priori, those crosses which do not respond to anther-culture in order to exclude them from this type of comparisons. By doing this, only crosses with known positive response to anther culture methods would be included in the research design. c) When the only purpose of the experiment is to assess plant resistance --a dichotomous trait--, then the 0-9 measurement effort is not necessary. However, if additionally, the degree of plant damage or the percentage of leaf area affected are also variables of interest, then the 0-9 scale for reaction scores serves all purposes. d) This analysis shows that linear models on categorical variables --binary, in this case-- are useful tools for understanding and testing hypothesis on cross-classified categorical data. The use of the logit response function, with the method of maximum likelihood for parameter estimation, makes possible the analysis of contingency tables with the presence of some small frequencies cells.

## References

Agresti, A. (1990), Categorical Data Analysis, New York: John Wiley & Sons, Inc.

Agresti, A. (1984), Analysis of Ordinal Categorical Data, New York: John Wiley & Sons, Inc.

Baenziger, P.S. & G.W. Schaeffer, 1983. Dihaploids via anthers cultured in vitro. In: L.D. Owens (Ed.), Genetic Engineering: Applications to Agriculture, pp. 269-284. Bestville Symposia in Agricultural Research. Rowman & Allanheld Pub., New Jersey.

Baenziger, P.S., D.M. Wesenberg, V.M. Smail, W.L. Alexander & G.W. Schaeffer, 1989. Agronomic performance of wheat doubled-haploid lines derived from cultivars by anther culture. Plant Breeding 103:101-109

Caligari, P.D.S., W. Powell & J.L. Jinks, 1987. A comparison of inbred lines derived by doubled haploidy and single seed descendent in spring barley (*Hordeum vulgare*). Ann. Appl. Biol. 111:667-675.

Chu, C.C., 1982. Haploids in plant improvement. In: I.K. Vasil, W.R. Scowcroft & K.J. Frey (Ed.), Plant Improvement and Somatic Cell Genetics, pp. 129-158. Academic Press, New York.

Correa-Victoria, F.J. & R.S. Zeigler, 1993a. Field breeding for durable rice blast resistance in the presence of diverse pathogen populations. In: Durability of Disease Resistance- p 215-218. Th. Jacobs and J.E. Parlevliet (eds). Kluwer Academic Publishers 375pp.

Correa-Victoria, F.J. & R.S. Zeigler. 1993b. Pathogenic variability in *Pyricularia grisea* at a rice blast "hot spot" breeding site in eastern Colombia. Plant Disease (accepted for publication).

Fienberg, S.E. (1980), The Analysis of Cross-Classified Categorical Data, 2nd Edition, Cambridge, MA: The MIT Press.

Freeman, D.H. (1987), Applied Categorical Data Analysis, New York: Marcel Dekker Inc.

Friedt, W., J. Breun, S. Zuchner & B. Foroughi-Wehr., 1986. Comparative value of androgenetic doubled haploid and conventionally selected spring barley lines. Plant Breed. 97:56-63.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969), "Analysis of Categorical Data by Linear Models", Biometrics, 25, 489-504.

Guiderdoni, E., 1991. "Gametic Selection in anther culture of rice (Oryza Sativa L)". Theoretical and Applied Genetics (1991) 81:406-412

Guimaraes, E.P., Amézquita, M.C., Lema. G., and Correa-Victoria, F. (1995), "A methodology to determine the minimum evaluation period length for disease-resistance characterization in rice (*Oryza sativa L.*) CIAT, (in preparation for publication).

Huang, C.S., H.S. Tsay, C.G. Ch
ern, C.C. Chen, C.C. Yeh & T.H. Tseng, 1988. Japonica rice breeding using anther culture. Journal Agric. Res. China 37(1):1-8.

IRRI, 1988. Standard Evaluation System for Rice 3rd. Edition. International Rice Testing Program. IRRI. Los Baños, Laguna, Philippines.

Levy, M., J. Romao, M.A. Marchetti & J.E. Hamer, 1991. DNA fingerprinting with a dispersed repeated sequence resolves pathotype diversity in the rice blast fungus. The Plant Cell 3:95-102.

Levy, M., F.J. Correa-Victoria, R.S. Zeigler, S. Xu, & J.E. Hamer, 1993. Genetic diversity of the rice blast fungus in a disease nursery in Colombia. Phytopathology 83: 1427-1433.

Lynch, P.T., R.P. Finch, M.R. Davey & E.C. Cocking, 1991. Rice tissue culture and its application. In: G.S. Khush & G.H. Toenniessen (Ed.), Rice Biotechnology, pp 135-156. C.A.B. Internat, Wallingford, U.K.

Núñez, V.M., W.M. Roca & C.P. Martínez, 1989. El cultivo de anteras en el mejoramiento de arroz. Serie 04SR-07-02-CIAT, Cali, Col:60p.

Pauk, J., Z. Kertesz, & Z. Barabas. 1988. Production of wheat lines from anther culture and their achievements in performance tests. Novenytermeless (Hungary). 37(3):197-203.

Picard, E., C. Parisot, P. Blanchard, P. Brabant, M. Causse, G. Doussinault, M. Trotlet, & M. Rousset, 1988. Comparison of the doubled haploid method with other breeding procedures in wheat (*Tritricum aestivum*) when applied to populations. In: T.E. Miller & R.M.D. Koebner (Ed.), Proc. Seventh Int. Wheat Genetic Symp. July 13-19, 1988, pp. 1155-1159. Inst. Plant Sci. Res. Cambridge, U.K.

Rosero, M.J., 1979. Breeding for blast resistance at CIAT. In: Proc. Rice Blast Workshop, p 63-67. Int. Rice Res. Inst., P.O. Box 930 Manila, Philippines.

Roumen, E.C. 1992. Small differential interaction for partial resistance in rice cultivars to virulent isolates of the blast pathogen. 64:143-148.

Sanint, L.R., C.P. Martinez, A. Ramírez & Z. Lentini, 1993. Economic analysis of rice anther culture versus conventional breeding. In: CIAT (Ed.) Trends in CIAT Commodities: Working document No. 128. July 1993: 74-96

SAS/STAT User's guide, Version 6, Fourth Edition, Volume 2, Cary, N.C: SAS Institute Inc., 1989. 846 pp.

Wald, A. (1943), "Tests of Statistical Hypotheses Concerning General Parameters When the Number of Observations Is Large", Transactions of the American Mathematical Society, 54, 426-482.

Weeraratne, H., C. Martínez & P.R. Jennings, 1981. Genetic strategies in breeding for resistance to rice blast *Pyricularia oryzae* in Colombia. In: IRAT-GERDAT (Ed.),Proc. Symp. Rice Resistance to Blast., March 18-21, 1981. pp 305-311. Montpellier, France.

Table 1. Overall performance of populations generated by each method across crosses in relation to Blast Resistance and Resistance Stability

| Descriptor | METHOD | | |
|---|---|---|---|
| | Pedigree | AC-F2 | AC-R2 |
| **1. Initial populations** | | | |
| Number of lines generated | 17917 | 441 | 740 |
| Median | 6 | 7 | 8 |
| Mode | 6 | 7 | 9 |
| Range | 0-9 | 2-9 | 1-9 |
| Skewness | 0.2 | -0.2 | -0.9 |
| Number (and proportion) of resistant lines | 3491(19.5) | 64(14.5) | 67(9.1) |
| **2. Final populations** | | | |
| * Number (and proportion) of resistant lines produced | 681(3.8) | 64(14.5) | 67(9.1) |
| Median | 3 | 4 | 4 |
| Mode | 3 | 4 | 4 |
| Range | 1-4 | 2-4 | 1-4 |
| * Number (and proportion) of stable resistant lines at the end of the evaluation | 163(0.9) | 5(1.1) | 13(1.8) |
| Median | 3 | 3 | 3 |
| Mode | 4 | 3 | 3 |
| Range | 1-4 | 3-4 | 2-4 |

**Table 2. Performance, per individual cross, of populations generated by each method in relation to blast resistance (R) and resistance stability (RS)**

| | Pedigree | | | AC-F$_2$ | | | AC-R$_2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| CROSS | N[1] | R[2] | SR[3] | N | R | SR | N | R | SR |
| Cross type: JxJ - SxR | | | | | | | | | |
| 2. CT5782 | 1291 | 170 (13.2) | 79 (6.1) | 207 | 38 (18.4) | 2 (1.0) | 42 | 6 (14.3) | 1 (2.4) |
| 3. CT8813 | 1899 | 125 (6.6) | 51 (2.7) | 47 | 8 (17.0) | 0 (0.0) | 55 | 12 (4.8) | 6 (10.9) |
| 10. CT8820 | 1465 | 74 (5.1) | 6 (0.4) | - | - | - | 238 | 10 (4.2) | 1 (0.4) |
| 7. CT8817 | 1471 | 61 (4.2) | 2 (0.1) | 27 | 7 (25.9) | 2 (7.4) | 168 | 22 (13.1) | 0 (0.0) |
| 9. CT8819 | 1583 | 95 (6.0) | 1 (0.06) | 1 | 0 (0.0) | 0 (0.0) | 30 | 0 (0.0) | 0 (0.0) |
| 6. CT8816 | 1565 | 42 (2.7) | 0 (0.0) | - | - | - | 16 | 2 (12.5) | 0 (0.0) |
| Cross type: JxI - SxR | | | | | | | | | |
| 4. CT8814 | 1404 | 6 (0.4) | 1 (0.07) | 104 | 5 (4.8) | 1 (1.0) | 101 | 7 (6.9) | 4 (4.0) |
| 8. CT8818 | 1480 | 50 (3.4) | 0 (0.0) | - | - | - | 42 | 6 (14.3) | 1 (2.4) |
| Cross type: JxI - SxS | | | | | | | | | |
| 1. CT5780 | 1922 | 34 (1.8) | 15(0.8) | 42 | 5 (11.9) | 0 (0.0) | 2 | 0 (0.0) | 0 (0.0) |
| 11. CT8821 | 2057 | 24 (1.2) | 8 (0.4) | 3 | 1 (33.3) | 0 (0.0) | 30 | 2 (6.7) | 0 (0.0) |
| 5. CT8815 | 1780 | 0 (0.0) | 0 (0.0) | 10 | (0.0) | 0 (0.0) | 16 | 0 (0.0) | 0 (0.0) |
| Total (%) | 17917 | 681 (3.8) | 163 (0.9) | 441 | 64 (14.5) | 5 (1.1) | 740 | 67 (9.1) | 13 (1.8) |

[1] N = initial number of lines generated.
[2] R = number (and proportion) of blast resistant lines
[3] SR = number (and proportion) of stable resistant lines

**Table 3. Contingency tables and logit values**

| Cross-type | Method | Initial no. of lines | Resistance | | | Resistance Stability | | |
|---|---|---|---|---|---|---|---|---|
| | | | $N^R$ | observed $\pi_R$ | logit | $N_{SR}$ | observed $\pi_{SR}$ | logit |
| JxJ, SxR | AC-$F_2$ | 282 | 53 | 18.8 | -1.46 | 4 | 1.4 | -4.24 |
| | AC-$R_2$ | 549 | 52 | 9.5 | -2.26 | 8 | 1.5 | -4.21 |
| | PM | 9274 | 567 | 6.1 | -2.73 | 139 | 1.5 | -4.19 |
| JxI, SxR | AC-$F_2$ | 104 | 5 | 4.8 | -2.98 | 1 | 1.0 | -4.63 |
| | AC-$R_2$ | 143 | 13 | 9.1 | -2.30 | 5 | 3.5 | -3.32 |
| | PM | 2884 | 56 | 1.9 | -3.92 | 1 | 0.03 | -7.97 |
| JxI, SxS | AC-F' | 55 | 6 | 10.9 | -2.10 | - | - | - |
| | AC-$R_2$ | 48 | 2 | 4.2 | -3.13 | - | - | - |
| | PM | 5759 | 58 | 1.0 | -4.59 | - | - | - |
| Total | | 19098 | 812 | 4.3 | | 158 | 0.83 | |

1\ $N_R$ = number of resistant lines

2\ $N_{SR}$ = number or stable resistant lines

Table 4. 'Analysis of variance' table. Linear model with logit response function

| | General blast resistance | | | Resistance stability | | |
|---|---|---|---|---|---|---|
| **Source** | df | Wald chi-quare statistic | prob | df | Wald chi-square | prob |
| Intercept | 1 | 598.4 | 0.00001 | 1 | 313.34 | 0.00001 |
| Method | 2 | 63.2 | 0.00001 | 2 | 15.96 | 0.0003 |
| Cross type | 2 | 33.0 | 0.00001 | 1 | 4.13 | 0.0422 |
| Method x Cross type | 4 | 19.2 | 0.0007 | 2 | 16.34 | 0.0003 |

## Table 5: Parameter estimates for logit models

| Parameter | | Resistance | | Resistance stability | |
|---|---|---|---|---|---|
| | | Estimate | Prob. | Estimate | Prob. |
| Intercept | | -2.83 | 0.01 | -4.76 | 0.01 |
| Method (M) | AC-F$_2$ | 0.65 | 0.01 | 0.32 | 0.45 |
| | AC-R$_2$ | 0.27 | 0.16 | 0.99 | 0.01 |
| | PM | -0.92 | - | -1.31 | - |
| Cross Type (C) | JJ - SxR | 0.68 | - | 0.55 | - |
| | JI - SxR | -0.24 | 0.13 | -0.55 | 0.04 |
| | JI - SxS | -0.44 | 0.03 | - | - |
| Cross Type x Method | | | | | |
| JJ - SxR | AC-F$_2$ | -0.40 | - | -0.35 | - |
| | AC-R$_2$ | 0.06 | - | -0.99 | - |
| | PM | 0.34 | - | 1.34 | - |
| JI - SxR | AC-F$_2$ | -0.13 | 0.71 | 0.35 | 0.41 |
| | AC-R$_2$ | 0.50 | 0.04 | 0.99 | 0.01 |
| | PM | -0.37 | - | 1.34 | - |
| JI - SxS | AC-F$_2$ | 0.53 | 0.06 | - | - |
| | AC-R$_2$ | -0.56 | 0.03 | - | - |
| | PM | -0.03 | - | - | - |

Table 6. Probabilities of producing resistant ($p_R$) and stable resistant lines ($p_{SR}$)

| Cross-type | Method | $p_R$ | $p_{SR}$ |
|---|---|---|---|
| JxJ, SxR | AC-$F_2$ | .18 | .019 |
| | AC-$R_2$ | .13 | .038 |
| | PM | .05 | .004 |
| JxI, SxR | AC-$F_2$ | .05 | .007 |
| | AC-$R_2$ | .08 | .013 |
| | PM | .02 | .0009 |
| JxI, SxS | AC-$F_2$ | .06 | - |
| | AC-$R_2$ | .05 | - |
| | PM | .01 | - |

**Table 7. Treatment comparisons. Logit models**

| Comparison | | General blast resistance | | Resistance stability | |
|---|---|---|---|---|---|
| | | chi-square | prob | chi-square | prob |
| Method (M) | AC-$F_2$ vs. PM | 48.1 | 0.00001 | 4.7 | 0.0296 |
| | AC-$R_2$ vs. PM | 18.9 | 0.00001 | 15.9 | 0.0001 |
| | AC-$F_2$ vs. AC-$R_2$ | 1.3 | 0.2628 | 1.1 | 0.2875 |
| Cross Type (C) | JJ-SxR vs. JI-SxR (3 vs. 2) | 21.2 | 0.00001 | 4.1 | 0.0422 |
| | JJ-SxR vs. JI-SxS (3 vs. 1) | 14.7 | 0.0001 | - | - |
| | JI-SxR vs. JI-SxS (2 vs. 1) | 0.4 | 0.5474 | - | - |
| Method within Cross Type | | | | | |
| JJ - SxR: | AC-$F_2$ vs. PM | 64.0 | 0.00001 | 0.01 | 0.9128 |
| | AC-$R_2$ vs. PM | 9.7 | 0.0018 | 0.01 | 0.9378 |
| | AC-$F_2$ vs. AC-$R_2$ | 14.2 | 0.0002 | 0.001 | 0.9646 |
| JI - SxR | AC-$F_2$ vs. PM | 3.8 | 0.0501 | 5.5 | 0.0188 |
| | AC-$R_2$ vs. PM | 25.5 | 0.00001 | 17.9 | 0.00001 |
| | AC-$F_2$ vs. AC-$R_2$ | 1.6 | 0.2083 | 1.4 | 0.2326 |
| JI - SxS | AC-$F_2$ vs. PM | 30.3 | 0.00001 | - | - |
| | AC-$R_2$ vs. PM | 3.9 | 0.0479 | - | - |
| | AC-$F_2$ vs. AC-$R_2$ | 1.5 | 0.2187 | - | - |

# ANALISIS CON AFLP DE LA ESTRUCTURA GENETICA DE LA COLECCION CORE DE FRIJOL SILVESTRE

Myriam Cristina Duque E.[1], Delkin Orlando González[2]
Steve Beebe[3] y Joseph Tohme[2]

## Resumen

El frijol silvestre es un grupo con mayor diversidad genética que el grupo de los cultivados. Ante la imposibilidad de estudiar las accesiones de frijol del Banco de Germoplasma del CIAT, se formó la Colección CORE para conocerla en detalle. De ella se han tomado las accesiones silvestres y se han estudiado con técnicas de AFLP (Amplified fragment length polymorfism) buscando un mayor conocimiento de la estructura genética de la población. El análisis estadístico se cumplió en dos fases: distancias genéticas y análisis de correspondencia múltiple. Los resultados señalan los grupos conocidos de Mesoamérica y Zona Andina y además sugiere posibles subgrupos dentro de cada una de ellas.

## Introducción

Diferentes estudios realizados sobre poblaciones silvestres de frijol (*Phaseolus vulgaris* L.) apoyan la idea de que en este grupo hay mayor diversidad genética que entre los frijoles cultivados. Pruebas de ello se han encontrado en expresiones morfológicas (Urrea, C.A. y Singh, S.P., 1991) y en términos de proteinas de la semilla, al encontrar solo entre los silvestres arcelina y algunas formas de faseolina (Romero-Andreas, 1984). La estructura de las poblaciones silvestres ha sido estudiada

[1] Unidad de Biometría, CIAT, AA 6713, Cali , Colombia

[2] Unidad de Biotecnología, CIAT

[3] Programa de Fríjo!. CIAT

utilizando además isoenzimas, enfoque que confirmó la existencia de los dos grandes grupos de genes: Mesoamérica y los Andes, y presentó uno nuevo formado por el Norte de Los Andes con características combinadas de los dos grupos mayores mencionados (Urrea y Singh, 1991). El presente estudio busca avanzar en la comprensión de la estructura genética de las poblaciones silvestres utilizando DNA mediante técnicas de AFLP.

## Colección CORE (C.C.)

La Unidad de Recursos Genéticos del CIAT (URG) dispone aproximadamente de 24.000 accesiones de frijol en su banco de germoplasma. La evaluación realizada sobre la colección completa sería impracticable, por lo tanto se ha formado un subconjunto llamado COLECCION CORE (C.C.), la cual constituye una proximación flexible y dinámica para estudiar la diversidad genética y promover un mejor uso del germoplasma del banco sin ser un substituto de él. La C.C. es algo más que un grupo de fuentes de resistencia o de tipos comerciales.

Para la conformación de la C.C. se tuvieron en cuenta en primer lugar el origen: centros primarios (75%) y centros secundarios (25%). Según el origen, la selección de genotipos de la C.C. se hizo así: para el material de los centros primarios se tiene en cuenta el país, la importancia de la zona como productora de frijol, las características agroecológicas del sitio de recolección y características morfológicas como hábito de crecimiento, tamaño y color de la semilla. Para los centros secundarios se busca mantener la representatividad regional, por hábito y por tipo de grano. La C.C. tiene accesiones de centros primarios, secundarios, silvestres, cultivares estándar y líneas mejoradas. (Tohme et. al, 1994).

## Técnicas AFLP.

AFLP es la sigla de "Amplified Fragment Length Polymorfism". Este proceso consiste en la amplificación selectiva de fragmentos de restricción, a diferencia del RFLP, en el cual la amplificación es sobre fragmentos aleatorios.

Para obtener la amplificación selectiva se requiere de "primers" o "sebadores" que son secuencias conocidas de DNA, de cadena sencilla que al ser homólogos de otras cadenas sencillas, actúan como promotores para que enzimas específicas produzcan las copias de la cadena molde. Mediante diferentes combinaciones de primers es posible amplificar diferentes fragmentos de DNA haciendo selecciones mas específicas. El resultado final es una auto-radiografía en la cual pueden detectarse "manchas" reveladoras de la presencia de diferentes tipos de moléculas que se ordenan según su peso y carga eléctrica. Este proceso es similar a una huella digital de los genotipos incluidos, al revelar qué tipo de moléculas posee. Manchas situadas al mismo nivel se refieren a moléculas del mismo tipo y para hablar de su existencia se dirá que la "banda está presente" en el genotipo.

## Materiales y métodos.

Los genotipos silvestres incluidos en este estudio provienen de la CC con orígenes geográficos distribuidos por Mesoamérica y Los Andes. (Tabla 1)

**Tabla 1: Origen geográfico de los genotipos silvestres de la C.C. analizados con AFLP**

| PAIS | FRECUENCIA | % |
|---|---|---|
| Argentina | 10 | 8.8 |
| Bolivia | 3 | 2.6 |
| Colombia | 12 | 10.5 |
| Costa Rica | 1 | 0.9 |
| Ecuador | 7 | 6.1 |
| Guatemala | 11 | 9.6 |
| México | 50 | 43.9 |
| Perú | 19 | 16.7 |
| El Salvador | 1 | 0.9 |
| TOTAL | 114 | 100.0 |

El DNA utilizado fue extraído de trifolios jóvenes y su proceso siguió el protocolo de laboratorio para AFLP de la Unidad de Biotecnología de CIAT. El DNA fue dividido en dos partes cada una de ellas sometida a una combinación de primers diferente para lograr amplificación de diferentes fragmentos.

**Análisis Estadístico**

Para el análisis de los datos se procedió a generar una matriz en la cual los individuos formaban filas, y las bandas evaluadas en cada uno de ellos constituían las columnas. La identificación de los genotipos estaba dada por el país de origen y un número consecutivo. Las variables de análisis, correspondientes a la presencia o ausencia de bandas fue referenciada con la siguiente nomenclatura:

banda 1 - banda 110 : Bandas de la primera combinación de primers;
banda 111 - banda 203 : Bandas de la segunda combinación de primers.

El análisis se realizó en los tres conjuntos:

1. Primera combinación de primers
2. Segunda combinación de primers
3. Conjunto total: Primera y segunda combinación de primers

La estrategia de análisis se cumplió en dos fases: distancias genéticas y análisis de correspondencia múltiple. Para estudiar la distancia genética entre genotipos se partió de la definición de similaridad de Nei-Li (1979), también conocida como similaridad de Dice (1945) o de Sφrensen (1948), (Legendre & Legendre, 1984).

$$S_{i,j} = \frac{2a}{2a+b+c}$$

$S_{i,j}$ : Similaridad entre el ecotipo i y el j
a : Número de bandas presentes simultáneamente en los genotipos i, j.

b : Número de bandas presentes en i y ausentes en j.

c : Número de bandas presentes en j y ausentes en i.

La conversión de la medida de similaridad en distancia se hizo con la expresión

$$D = 1 - S$$

Se elige esta medida de similaridad, que excluye el número de bandas ausentes simultáneamente en los dos genotipos, pues esta ausencia no constituye necesariamente una semejanza entre individuos en el caso específico que se está considerando. La doble ponderación en las bandas que coinciden, enfatiza la medida permitiendo diferenciar mejor genotipos con baja similaridad.

Además, este coeficiente tiene significado biológico directo, por ser un estimador de la proporción de fragmentos amplificados que se comparten. Finalmente, el coeficiente elegido tiene menor valor de sesgo cuando la similaridad es alta. (Lamboy, 1994).

El cálculo de la matriz de distancia permitió la clasificación de los genotipos con el método de unión media (UPGMA).Los cálculos de similaridad , distancia, y el análisis de clasificación fueron hechos con el programa CLASIFI (García et al, 1995) escrito en macrolenguaje SAS, versión 6.09 . Los dendogramas se obtuvieron con el programa de expresión gráfica TREE de NTSYS. (Rolfh, F.J., 1989).

Esta aproximación es un poco laxa (en resultados) ya que solo importa el número de bandas en común y no cuáles bandas son, permitiendo agrupar en algunos casos individuos con baja afinidad.

La segunda fase, el Análisis de Correspondencia Múltiple buscó asociar la representación de los individuos con la representación de sus variables en los ejes principales de variación. Para ello se definieron como variables activas las correspondientes a la presencia o ausencia de las bandas y como suplementarias el país de origen, conservando para cada genotipo su valor geográfico individual. De no hacerse así, aparece un único punto por país en su centro de masa, perdiéndose la capacidad de apreciar la dispersión espacial. El procedimiento CORRESP del paquete SAS, versión 6.09 fue utilizado para este enfoque.

## Resultados y discusión

### Fase 1: Análisis de distancias genéticas

*Conjunto de datos, primera combinación de primers*

Globalmente hay los siguientes aspectos importantes: (figura 1)

1. La separación en primer lugar de un grupo de genotipos originales del Ecuador y Norte de Perú muy aglutinado.
2. Otros dos grandes grupos: Uno principalmente mesoamericano en el cual aparecen algunos genotipos argentinos y otro grupo en su mayoría de origen andino con unos pocos mexicanos. En el subgrupo mesoamericano, Guatemala aparece como un bloque.
3. Colombia aparece en todos los grupos.

*Conjunto de datos: Segunda combinación de primers.*

Los aspectos relevantes de la clasificación (figura 2) son los siguientes:

1. Se forma inicialmente el grupo de Ecuador, y Norte de Perú pero pierde un poco la coherencia pues los genotipos aparecen mas dispersos.
2. Con similaridad de 0.70 se forman dos grandes grupos: El mesoamericano con mucha cohesión en los mexicanos y un poco menos entre los guatemaltecos; y el andino en el cual figuran algunos mexicanos. No es muy marcada la regionalización al interior de este grupo.
3. Colombia aparece dispersa, solo entre los andinos.

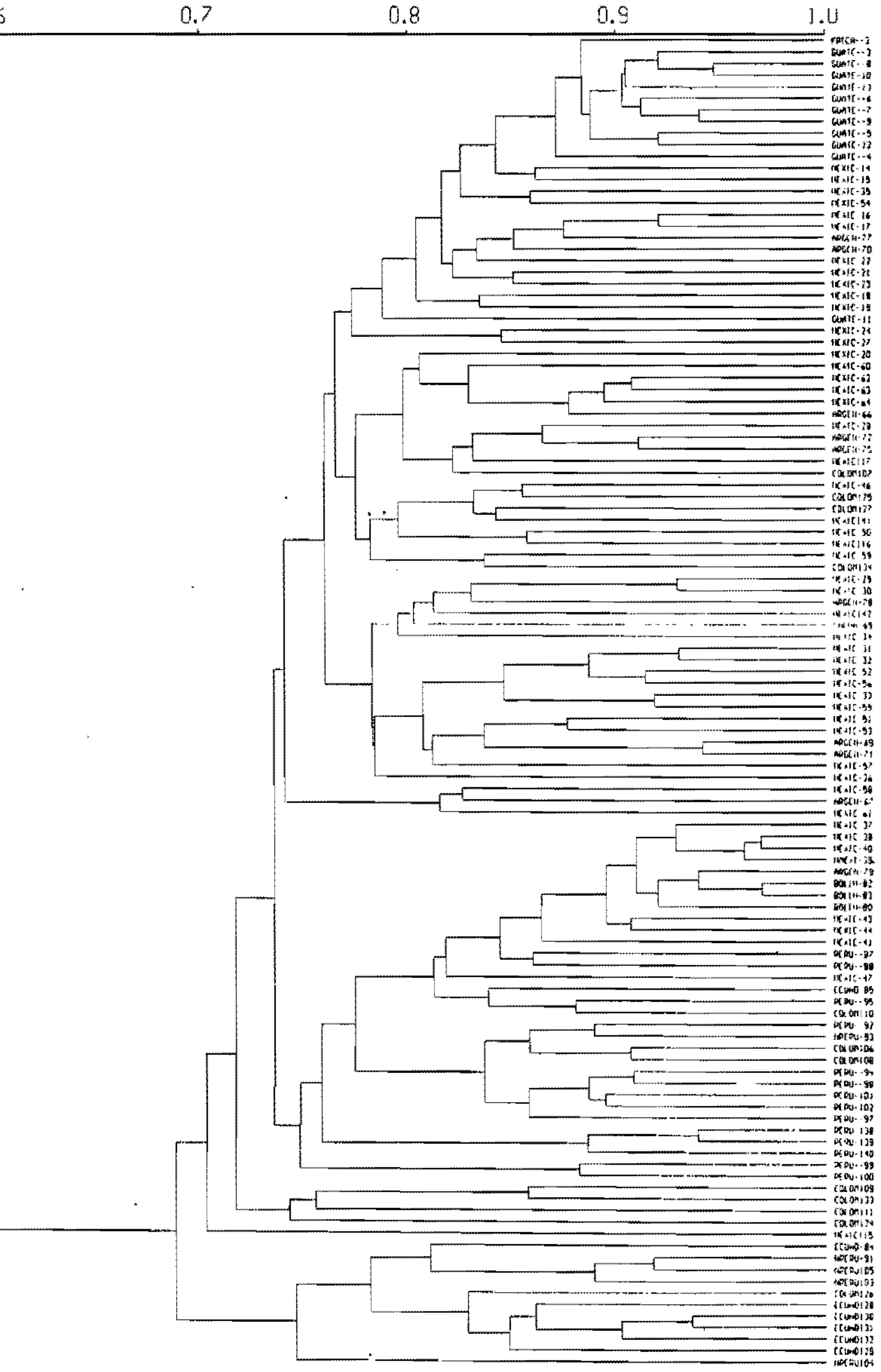


**Figura 1. Clasificación con distancias genéticas. Conjunto de datos: Primera combinación de Primers.**

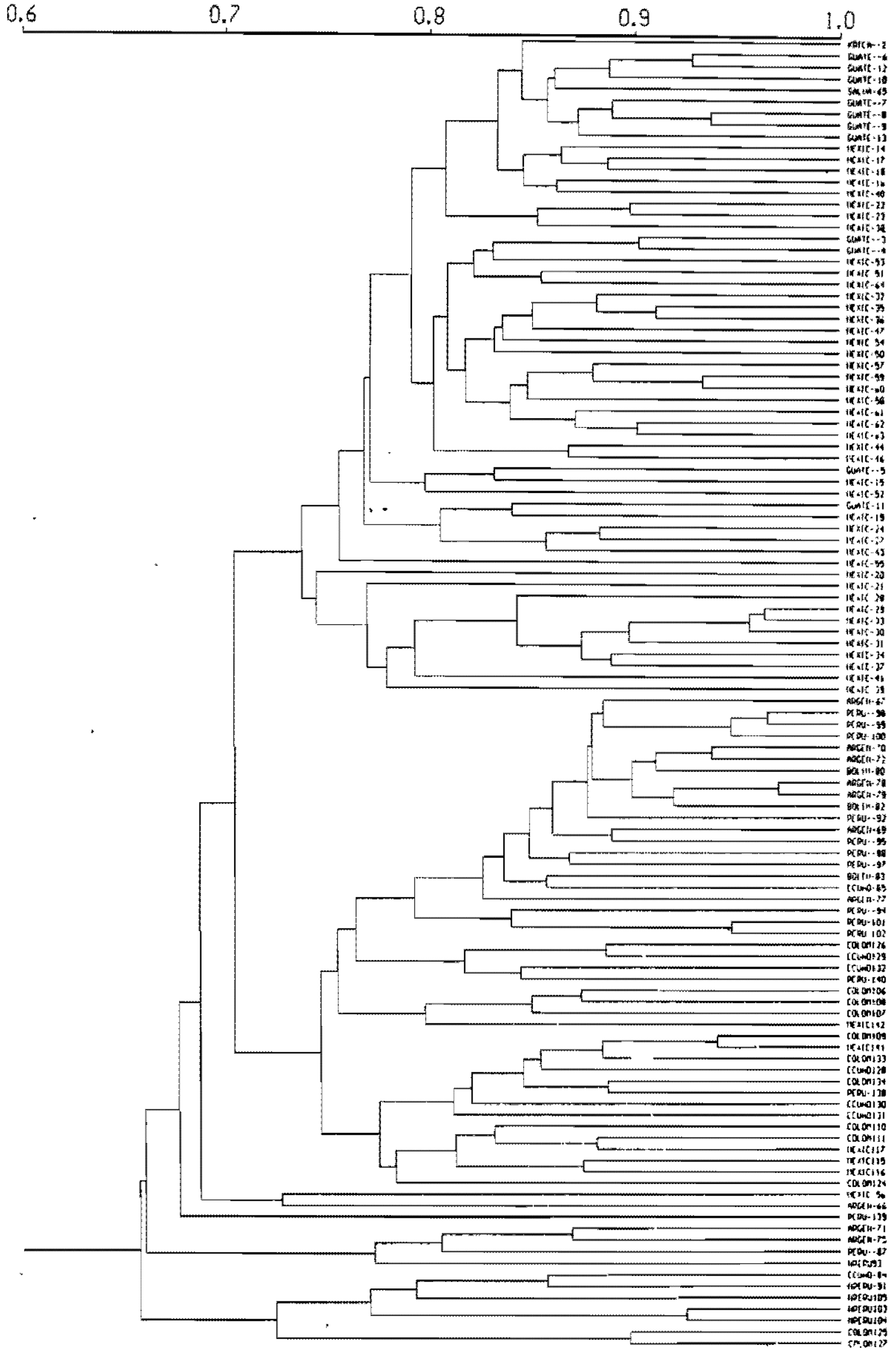


Figura 2. Clasificación con distancias genéticas. Conjunto de datos: Segunda combinación de Primers.

*Conjunto de datos: Combinación de primers 1 y 2*

La figura 3 muestra los aspectos mas destacados en este conjunto:

1. La separación del grupo del Norte de Perú y Ecuador es mas zonificada pues diferencia estas dos regiones.
2. Hay una distribución más ordenada de Colombia.
3. Se generan aparte del grupo conocido como andino con mayor concentración de Argentina y Bolivia, aunque persiste una zona de mezclas.
4. El grupo mesoamericano condensa bastante a Guatemala y a México.

Para evaluar comparativamente la información obtenida de los diferentes grupos de la matriz de distancias de los datos de la primera combinación de primers (D1) fue correlacionada con la respectiva matriz de la segunda combinación de primers (D2) y con la de ambas combinaciones (D12) con los resultados siguientes:

| | |
|---|---|
| Correlación entre D1 y D2: | 0.27 (***) |
| Correlación entre D1 y D12: | 0.80 (***) |
| Correlación entre D2 y D12: | 0.80 (***) |

lo cual confirma el carácter complementario de los dos grupos de primers y la importancia de considerarlos juntos.

## Fase 2: Análisis de Correspondencia Múltiple.

*Conjunto de datos: Primera combinación de primers.*

La magnitud de los primeros valores propios es baja indicando una estructura no muy diferenciada. $\lambda_1 = 0.30$ (9.8%), $\lambda_2 = 0.25$ (6.7%), $\lambda_3 = 0.23$ (5.9%).

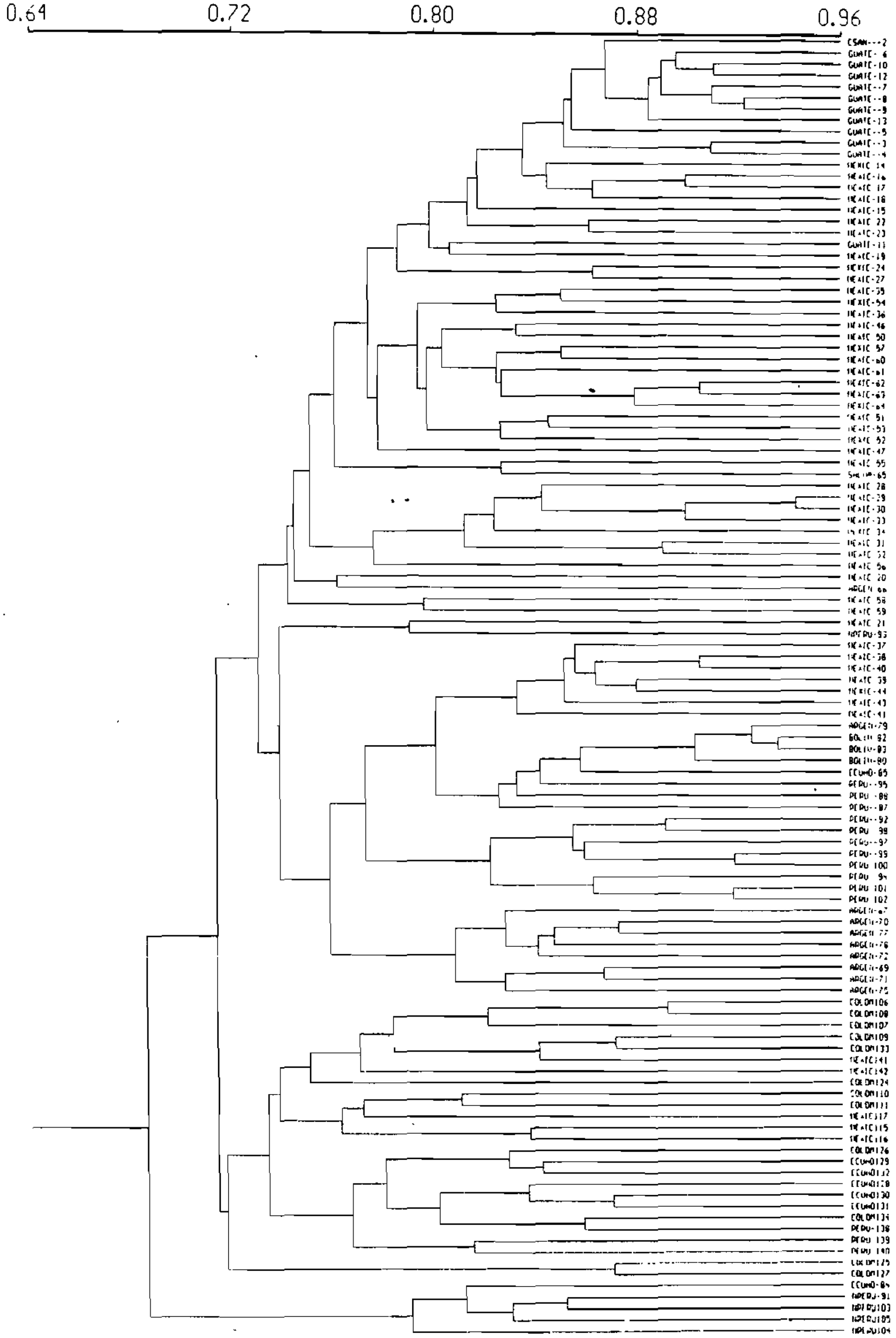


**Figura 3. Clasificación con distancias genéticas. Conjunto de datos: Combinación de Primers 1 y 2.**

No se encuentran valores muy altos en los cosenos cuadrados de las bandas de los primeros ejes principales, indicando que estos ejes están definidos mas por conjuntos de variables que por algunas pocas de singular importancia.

Tabla 2: **Bandas más importantes en la definición de los ejes principales. Primera combinación de Primers**

| Eje 1 | | | Eje 2 | | | Eje 3 | | |
|---|---|---|---|---|---|---|---|---|
| **Banda** | **Coseno²** | **Modalidad** | **Banda** | **Coseno²** | **Modalidad** | **Banda** | **Coseno²** | **Modalidad** |
| 4 | .48 | si | 70 | .44 | si | - | - | - |
| 100 | .46 | no | 71 | .49 | si | - | - | - |

Adicionalmente los bandas 1, 13, 14, 26, 46, 66, 77, 79, 92, 103 contribuyeron a la formación de varios ejes.

La gráfica tridimensional permite apreciar una marcada diferenciación en la dimensión 1 de los genotipos de Ecuador y Norte de Perú (extremo positivo) a los de origen guatemalteco en el extremo negativo. En la dimensión 2 se ordenan los de Ecuador y Norte de Perú en los mayores valores. Bolivia, Argentina y el resto de Perú tienen las menores calificaciones. La dimensión 3 colabora en la definición de los grupos anteriores pero los demás genotipos no son mayormente diferenciados. (figura 4)

*Conjunto de datos: Segunda combinación de primers*

Se obtienen unos valores propios un poco mayores, $\lambda_1 = .33$ (11.8%), $\lambda_2 = .27$ (7.8%), $\lambda_3 = .23$ (6.0%) por lo tanto se espera un poco de mejor diferenciación.

Atendiendo a los cosenos cuadrados puede hacerse la selección de las bandas mas importantes en la definición de los ejes (Tabla 3).

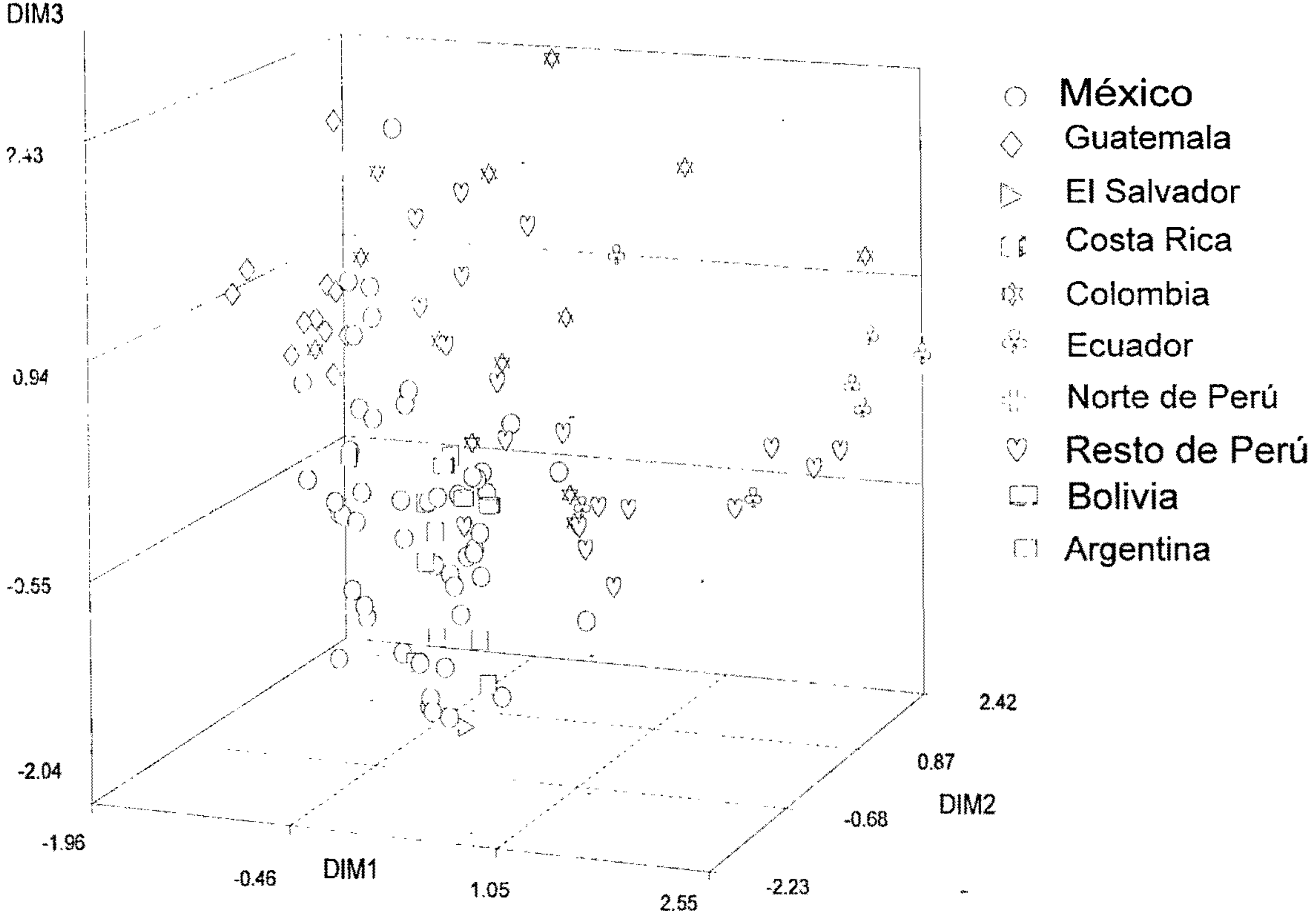


***Figura 4 . ACM --- Primera combinacion de primers***

**Tabla 3:** **Bandas más importantes en la definición de los ejes principales segunda combinación de primers**

| Eje 1 | | | Eje 2 | | | Eje 3 | | |
|---|---|---|---|---|---|---|---|---|
| Banda | Coseno$^2$ | Modalidad | Banda | Coseno$^2$ | Modalidad | Banda | Coseno$^2$ | Modalidad |
| 114 | .49 | no | - | - | - | 116 | .41 | si |
| 125 | .63 | si | - | - | - | - | - | - |
| 131 | .57 | si | - | - | - | - | - | - |
| 141 | .40 | no | - | - | - | - | - | - |
| 152 | .43 | si | - | - | - | - | - | - |
| 190 | .64 | no | - | - | - | - | - | - |
| 196 | .47 | no | - | - | - | - | - | - |

En este caso, las bandas 122, 123, 124, 135, 181, 188 participan en la definición de más de un eje.

La representación gráfica ofrece mas claridad: en el sector definido por los mayores valores de las dimensiones 1 y 2 y los menores de la dimensión 3 aparecen los genotipos de Bolivia, Argentina y Sur y Centro de Perú. En la misma región, un poco mas arriba se congregan los genotipos del Norte de Perú, Ecuador y algunos de Colombia, que desplazan hasta los lugares más altos. En el sector de los mayores valores de la dimensión 2 y menores para las otras dos dimensiones se aglutinan los genotipos mesoamericanos, ecuatorianos, sur de Perú y Colombia que amerita un poco de estudio adicional, los genotipos de Colombia presentan gran amplitud en su ubicación. (figura 5)

*Conjunto de datos: Primera y segunda combinación de primers.*

Al combinar los datos, los valores propios resultantes, como es de esperar sugieren la existencia de un conjunto no muy diferenciado pero en cuyo interior hay alguna estructura que perfilar. $\lambda_1$ = .28 (8 7%), $\lambda_2$ = .22 (5.3%), y $\lambda_3$ = .21 (4.9%).

Las bandas importantes en cada conjunto separado conservan su categoría pero con niveles ligeramente inferiores. Sin embargo, en la representación gráfica puede apreciarse mas balance en la distribución al permitir una regionalización más definida en términos generales. En todos los casos la dispersión de los genotipos colombianos es un hecho de resaltar. (figura 6)

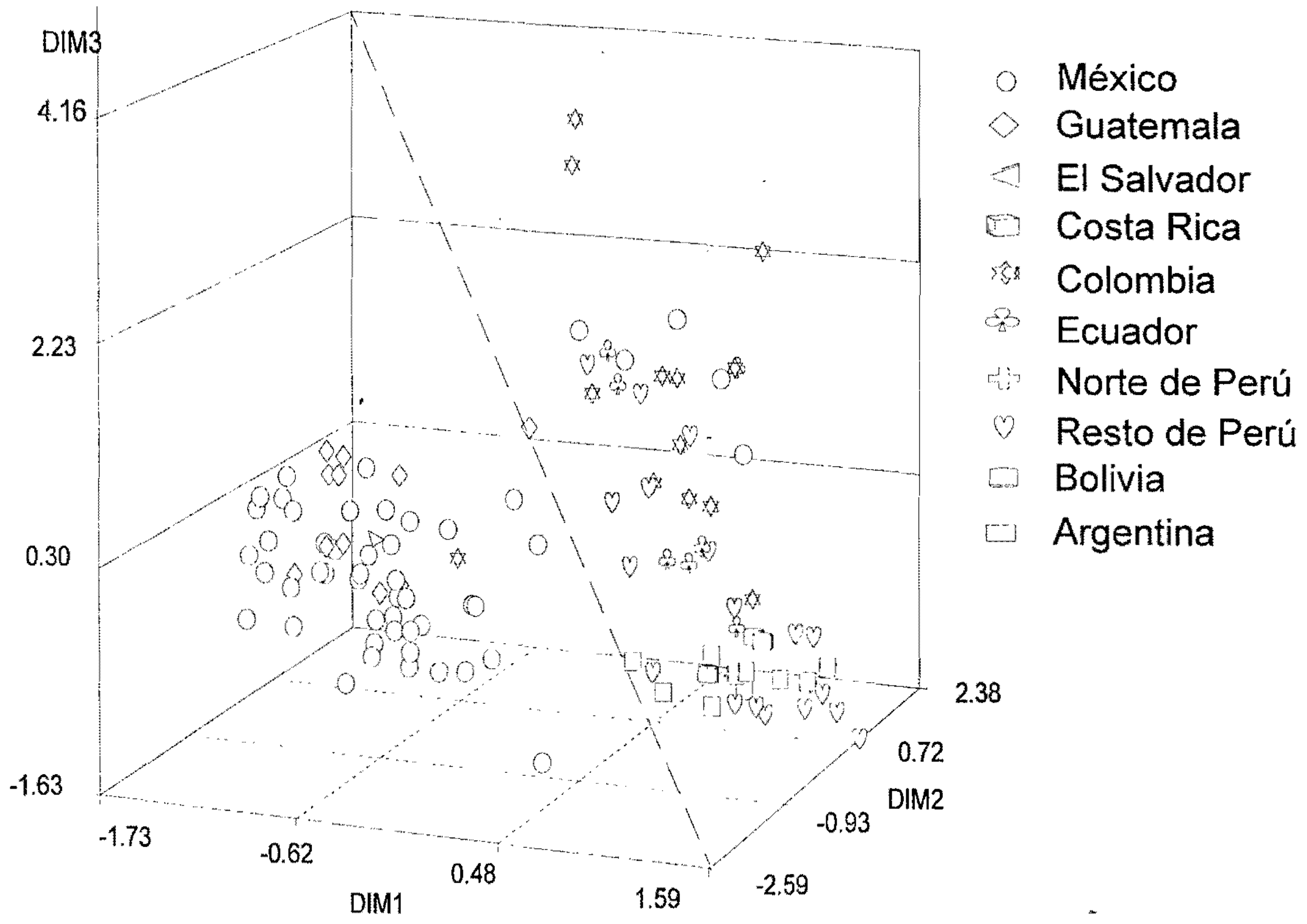


*Figura 5. ACM --- Segunda combinacion de primers*

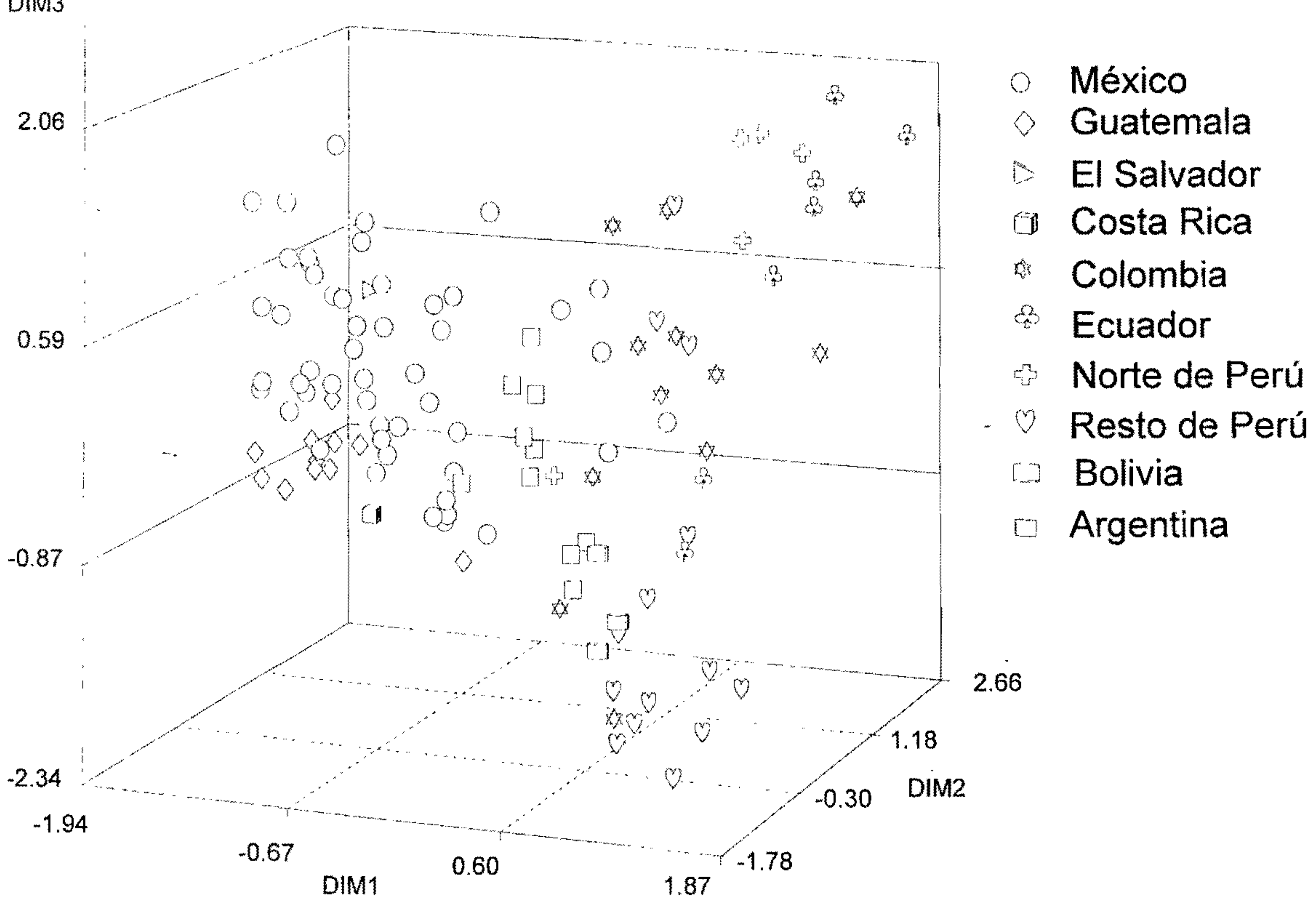


*Figura 6. ACM --- Combinaciones 1 y 2 de Primers*

## Conclusiones

1. Hay alta similaridad entre los genotipos silvestres de la C.C., tanto en los subconjuntos, como en el global y este hecho se percibe en los dos tipos de análisis.
2. La información suministrada por la combinación de primers 1 es complementaria de la suministrada por la combinación 2.
3. El conjunto completo ofrece mayor claridad en la interpretación de los grupos.
4. Los grupos andino y mesoamericano mantienen su identidad y se destacan en su interior Guatemala y el formato para Norte de Perú y Ecuador.
5. Colombia no presenta mucha afinidad con el resto de regiones de suramérica.

**Referencias**

García, J.A.; Duque, M.C.; Tohme, J.M.; Xu, S.; Levy, Morris, 1994. Un programa SAS para Análisis de clasificación. Documento de trabajo, CIAT.

Lamboy, Warren F. (1994) Computing Genetic similarity coefficients from RAPD data: The effects of PCR artefacts.

Legendre, L., y Legendre, P. 1984. Ecologie numérique 2. La structure des donées écologique. Collection d'écologie 13, 2a ed. Masson, Presses de l'université du Québec, Québec, Canadá.

Rolph, F.J., 1989. NTSYS, PC Numerical Taxonomy and Multivariate Analysis System.

Romero-Andreas, J. 1984. Genetic variability in the seed phaseolin of nondomesticated bean (*Phaseolus vulgaris L.* var aborigeneus) and the inheritance and physiological effects of arcelin, a novel seed protein. Ph.D. thesis. Madison, University of Wisconsin-Madison, 168p.

SAS Institute Inc., 1989. SAS/STAT User's Guide version 6, Fourth edition, Volume 1, Cary N.C.

Urrea, C.A.; Singh, S.P. 1991. Variation for leaflet shape in wild and cultivated landraces of common bean. (Variación de la forma del foliolo en biotipos de frijol silvestre y cultivado). Bean Improvement Cooperative. Annual Report 34:133. En. I1. CIAT, Apartado Aéreo 6713, Cali, Colombia)

# ANALYSIS OF GENETIC RELATIONSHIPS AND DIVERSITY BASED ON MOLECULAR MARKER DATA

**James Nienhuis**

Dept. of Horticulture, University of Wisconsin, Madison, WI 53706

Additional index words: Molecular markers, Random amplified polymorphic DNA, genetic relationship, multidimensional scaling, sampling variance, genetic diversity

## RESUMEN

Cultivares y variedades criollas de una especie cultivada, representan la principal fuente de genes disponibles para el mejoramiento de los cultivos. El conocimiento de la relación genética entre genotipos es útil en un programa de mejoramiento genético porque permite la organización del germoplasma existente. Basados en 80 bandas polymorficas obtenidas bajo la metodologia de RAPDs (Randon Amplified Polymorphic DNA), relaciones geneticas entre 76 genotipos de *Phaseolus vulgaris* y lineas de mejoramiento de América Central fueron estimadas. Se observó dos principales grupos entre los cultivares, los cuales corresponden a las clases rojo pequeño y negro.

Para que los mejoradores puedan tomar decisiones con respecto al muestreo de recursos de germoplasma, se requiere del conocimiento de las diferencias y variación entre y dentro grupos de genotipos. Este informe presenta un procedimiento basado en el analisis de varianza-ANOVA para determinar las diferencias entre grupos de individuos. Ademas se discute el método de Nei para medir la diversidad genetica (marker variance) y la prueba de homogeneidad de varianza. Finalmente, se revisa un procedimiento para la estimacion de la varianza de muestreo de los datos obtenidos con marcadores moleculares.

## ABSTRACT

Cultivars and landraces of a crop species represent the primary gene pool available to plant breeders for improvement of crop plants. Knowledge of relative genetic relationships among genotypes is useful in a breeding program because it permits organization of germplasm resources. The genetic relationships among 76 *Phaseolus vulgaris* genotypes and breeding lines from Central America was estimated based on 80 polymorphic random amplification polymorphic DNA (RAPD) bands. Two major clusters were observed among the cultivars which corresponded with small red and black market classes.

In order for plant breeders to make informed decisions regarding the sampling of germplasm resources knowledge of differences and variances between and within groups of genotypes is required. This report presents a procedure based on the ANOVA for determining differences between groups of individuals. In addtion Nei's method for measuring genetic diversity (marker variance) and test of homogeniety of variance is discussed. Finally, a procedure for estimation of sampling variance of molecular marker data is reviewed.

## INTRODUCTION

Cultivars and landraces of a crop species represent the primary gene pool available to plant breeders for improvement of crop plants. Genetic improvement of crop plants is based upon the identification of favorable genes in the promary gene pool the subsequent introgression of those genes into adapted cultivars. The identification of desirable geno'ypes in germplasm collections is often limited by our lack of knowledge of the organization of germplasm resources. If germplasm collections could be systematically organized based on genetic relationships, then the efficiency of sampling and utilization of germplasm resources could be greatly improved.

Patterns of genetic diversity have been studied in crop species using a variety of molecular, chemical and morphological descriptors. The most commonly used molecular tools for measuring genetic relationships have been isoyzmes, seed proteins and molecular markers. Although informative and practical, the use of variable protein and isozyme markers has often been limited by their low frequency in many crop species (Goodman and Stuber, 1980). Molecular markers provide an opportunity to more precisely measure genetic relationships compared to morphological and biochemical markers because they: 1) are potentially unlimited in number, 2) are not affected by the environment and 3) can be organized into linkage maps (Helentjaris et al, 1986). Estimates of genetic relationships based on restriction fragment length polymorphisms (RFLPs) have been shown to be consistent with expectations based on known breeding behavior and pedigrees in numerous crop species, including maize (Smith et al, 1990). More recently, scientists have used random amplified polymorphic DNA (RAPD) molecular markers as a tool for measuring genetic relationships (Skroch et al, 1992; Welsh et al, 1990, Williams et al, 1990). RAPDs are technically simpler and lower cost compared to RFLPs; however, reproducibility of banding patterns can be affected by different concentrations of reaction components and cycling conditions (Weeden et al, 1992). Nevertheless, in a study comparing molecular markers, the sampling variances associated with RAPDs and RFLPs were found to be similar for estimation of genetic relationships among *Brassica oleracea* genotypes (dos Santos et al, 1994).

The objective of this paper is to illustrate methods to reveal structure of germplasm collections based on molecular marker data. The principal objectives are 1) to test the significance of differences between groups of genotypes, 2) to measure genetic diversity (variance) , and 3) estimate sampling variance.

## EXAMPLES OF GERMPLASM ORGANIZATION

The use of RAPD markers for germplasm organization can be illustrated using data from our laboratory on common bean, *Phaseolus vulgaris* . In a cooperative study with Dr. Steve Beebe of C.I.A.T., Cali, Colombia, 80 polymorphic RAPD bands were identified among 76 Central American red and black beans (Beebe et al., 1995. The common bean genotypes represent cultivars and breeding lines developed in different countries, and do not represent a population in Hardy-Weinberg equilibrium. The procedures for DNA isolation and PCR amplification based on random primers have been previously described (Beebe et al, 1995; Nienhuis et al., 1995).

## GENETIC DISTANCE ESTIMATORS

Techniques commonly used in the measurement of genetic relationships using isozymes and RFLP data can be applied with slight modification to RAPD marker data. Options regarding the choice of genetic distance estimator for binary data have been summarized by Gower (1985) and Jackson et al. (1989). The four possible observations of the comparison between two genotypes for a RAPD marker are classified based on the presence ( 1 ), or absence ( 0 ), of a RAPD marker for each genotype (Table 1). Genetic distance estimators differ on how they combine A, B, C and D, into a genetic distance or similarity. Measures of co-occurrance, which employ various ratios of similarities or differences to total comparisons, are the most commonly used. In measures of co-occurance two variations are common: one which utilizes all contingency table outcomes, A, B, C and D, in the denominator, and the second which excludes the absence by absence comparison (contingency table D outcome) from the denominator. The use of (0,0) comparisons in the measure of genetic relationships is a subject of much debate. Dudley (1995), in a recent review, suggested that including (0,0) matches is appropriate only when two alleles exist at a locus, one of which produces a band and the other does not. The assumption of two alleles at a locus is most appropriate in intra-specific comparisons, e.g. among landraces or cultivars of a crop species. Exclusion of (0,0) outcomes is often done in cases where interspecific comparisons are of interest. The choice of genetic distance estimator to use with RAPD data depends on the relative amount and quality of information each comparison provides for the estimation of sequence polymorphism.

## GENETIC DISTANCE AMONG COMMON BEAN GENOTYPES

Each polymorphic RAPD band across all genotypes was assigned a number (1, 2, 3 ...n) according to decreasing molecular weights. Each band was treated as a unit character, and the genotype was scored for the presence or absence of a band and coded as 1 or 0, respectively. In this intra-specific comparison among bean cultivars genetic distance was calculated between all pairs of genotypes based on the following formula which is the compliment to the simple matching coefficient (Gower, 1985)

$$GD(i,j) = S^{N}(i \neq j) / [S^{N}(i \neq j) + S^{N}(i=j)]$$

where GD(i,j) is the measure of genetic distance between genotypes i and j, $S^{N}(i \neq j)$ and $S^{N}(i=j)$ are the total number of discordant and concordant scores between genotypes i and j, respectively. GD values of 0.0 and 1.0 indicate no and maximum difference between two genotypes, respectively.

## RELATIONSHIPS REVEALED BY MDS

The matrix of GD estimates was reduced to two dimensions and displayed as a Multidimensional scaling (MDS) plot (Wilkinson, 1992). Multidimensional scaling (MDS) is a method for estimating the coordinates of a set of objects in a space of specified dimensionality from data measuring the relationships between pairs of objects (SAS Institute Inc., 1992). Based on MDS analysis of RAPD marker data, the red and black beans were separated into two distinct groups, suggesting that some consistent pattern of variation exist between their genomes (Fig. 1).

The plotting of the genotypes was consistent with relationships bases on known pedigrees; for example, 'MCD 2004' is located in close proximity to 3 out of 4 of its progenies (DICTAs '09', '57' and '76'). Similarly, 'XAN 112' plotted in close proximity to its progenies, 'Turbo' and 'ICTA 12'.

In addition, pairs of sister lines plotted closely together or were indistinguishable: 'RAB's 50' and '205'; DOR's '481' and '482'; DOR's '474' and '475'; and RAB's '310' and '311' (Fig. 1).

## TEST OF DIFFERENCES AMONG GROUPS OF GENOTYPES

An F-test modified from the analysis of molecular variance (AMOVA) (Excoffier et al, 1992), and the analysis of variance (ANOVA) (Weir, 1990) was used to test the significance of differences between the groups of genotypes (Table 2). The relevant F-value to test the significance of differences between groups of genotypes was the ratio of the mean square for the interaction between marker frequency and group (M1) relative to the interaction between marker frequency and individual genotypes nested within each group (M2). The F-test revealed that the genetic distance between red and black beans (0.268) was significant (Table 3).

Although the red and black bean genotypes are statistically different based on the ANOVA the biological significance of that difference needs to be better understood. A perspective can be obtained by evaluating random permutations between the two market classes. This represents way to look at how common our observed pattern is relative to random patterns. In addition, to increase the probability of rare classification patterns we we allowed the number of individuals in one group to range from one to 75.

## RANDOM PERMUTATIONS

The mean of the distribution of 10,000 permutations was 0.245 with a standard deviation of 0.006 (Fig. 2A). Given the distribution of random permutations, the probability of obtaining a value of 0.268, the genetic distance between the red and black beans, p ( $Z \geq 2.68$) is less than 0.001.

The number of red and black beans in the original data set was 35 (46%) and 41 (54%), respectively; however, the random permutations allowed for a random percentage of the individuals to be classified into one group and the remainder into the other group. For percentages of individuals ranging from 30 to 70 percent, the values for random permutations clustered very tightly around a mean of 0.245 (Fig 2B). In contrast, for percentages of individuals greater than 70% or less than 30%, the variance of the random permutations was much greater, and few individual observations exceeded the value of 0.268. This illustrates that even in the extreme example of random permutations where drift is allowed, a value of 0.268 is a very rare event. This observation strongly supports the conclusion that the significant genetic relationship between red and black bean groups represents a true biological phenomenon.

### Genetic diversity

The basis of genetic diversity is sequence variation. RAPDs are molecular markers which sample and reveal sequence variation by the differential amplification of DNA fragments; thus, genetic diversity within groups of red and black bean genotypes was measured as RAPD marker variance. Genetic diversity (RAPD marker variance) was estimated as Nei's gene diversity at a locus, $\Sigma$ (2n*p*q) / (n-1), summed over all marker loci, where n = the number of genotypes in each group, and p and q refer to the frequency of the presence (1) or absence (0) of each band, respectively (Nei, 1987). The RAPD marker variance of the red and black bean groups was similar, 0.225 and 0.210, respectively (Table 4). Comparison of RAPD marker variance between populations was done using t-test (Nei, 1987). Standard errors for t-test were computed using gthe bootstrap. Based on the t-test, the magnitudes of the variances associated with red and black groups were declared homogeneous. Thus, representative sampling of genetic diversity among the accessions included in this study would require similar numbers of genotypes sampled from the red and black groups. Another measure of diversity within a group is the mean genetic relationship among all possible pairs of individuals within a group. The mean genetic relationships highly correlated (>0.95) with total marker variance, as both are based on band frequency within a group. These two measures are also highly correlated with Nei's $H_S$ statistic (Nei, 1987).

**Sampling variance**

Sampling variance in estimation of genetic relationships occurs when a random subset of marker bands does not equal the value obtained from all possible bands. Two hundred bootstrap samples each of size n (n=10, 20, 30...n) were drawn from the full data set (Tivang et al, 1994). The genetic relationship between all pairs of genotypes was calculated for each bootstrap sample. The variance among the 200 bootstrap samples for each pair of genotypes was standardized to the coefficient of variance (CV) by dividing the variance by the bootstrap sample mean.

The number of polymorphic molecular marker bands used to estimate genetic relationships varies widely from 61 (Nienhuis et al., 1992) to 1205 (Smith et al., 1990). Larger numbers of polymorphic bands will provide increasingly more uniform coverage of the genome and will minimize bias due to undersampling certain regions of the genome. Nevertheless, it would be more efficient to determine genetic relatedness using a smaller set of polymorphic bands (Smith et al., 1990). The question is how large a sample of polymorphic bands is required to provide a given level of precision? The plot of the relationship between the coefficient of variation (CV) and sample size (number of bands) indicates that CVs as low as 20% for estimating genetic relationship between the average individual genotypes can be achieved by sampling as few as 80 bands (Fig. 3).

Among the genetic relationships among all possible pairs of genotypes within this population, approximately 40 relationships between relatively closely (<0.01), intermediate (0.20) and distantly (>0.39) related pairs were selected (Fig.3). The rate of reduction in CV with increasing number of bands samples was constant for the three groups, whereas the intercepts were different (Tivang et al., 1994). Expressing the relationships between the variance of bootstrap samples vs. sample size in terms of CV is a convenient method for allowing the investigator to decide the level of precision desired or achieved. However, the CV can vary either due to increased variance or due to changes in the mean of the bootstrap samples. Because the variance was constant, the different relationships displayed for the three groups is due to the differences in their mean genetic relationship. Thus,

because fewer bands are polymorphic among closely related individuals compared to more distantly related individuals, more bands will be required to achieve the same level of precision (CV).

## CONCLUSIONS

Plant breeders can use molecular markers to organize genetic resources into related groups to make more informed decisions regarding choice of parents. Nevertheless, relationship matrices generated by comparing each genotype to all other genotypes for a set of RAPD markers are no more informative than the original data. Interpretation of the results requires data reduction. Moreover, the plant breeding interpretation of the genetic relationship matrix requires not only graphic displays, but also methods to reveal structure of the data. Thus, knowledge of differences and the magnitudes of variance between and within groups translates the data into useful plant breeding information.

Knowledge of genetic relationships when complemented by phenotypic data can reveal sources of desirable characteristics in more closely related genotypes. This combined knowledge may permit the recovery of the recurrent parents phenotype in fewer breeding generations than would be required for a more distantly related donor parent.

## LITERATURE CITED

Beebe, S.E., I. Ochoa, J. Nienhuis, P. Skroch and J. Tivang (1995). Genetic diversity among common bean breeding lines developed for Central America. Crop Sci. (in press).

Dudley, J.W. 1994. Comparison of genetic distance estimators using molecular marker data, p. 3-7. In: Analysis of molecular marker data. Crop Science Society of America, Madison, WI.

Excoffier, L., P.E. Smouse and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric relationships among DNA haplotypes: application to human mitochondrial DNA restriction data Genetics 131:479-491.

Goodman, M.M. and C.W. Stuber. 1980. Genetic identification of lines and crosses using isoenzyme electrophoresis. Proc. 35th Annu. Corn, Sorghum Res. Conf. 35:10-31.

Gower, J.C. 1985. Measures of similarity, dissimilarity, and distance, p. 297-405. In: Kotz, S. and N.L. Johnson (eds.) Encyclopedia of statistical sciences. vol. 5. Wiley, New York.

Helentjaris, T., M. Slocum, S. Wright, A. Schaefer and J. Nienhuis. 1986. Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor. Appl. Genet. 72:761-769.

Jackson D.A., K.M. Sommer and H.H. Harvey. 1989.Similarity coefficients: measures of co-occurance and associations or simply measures of occurrence? Am. Nat. 133:436-453.

Nei, M. 1987. Molecular evolutionary genetics. Colombia University Press, New York.

Nei M. 1972. Genetic distance between populations. American Naturlist 106:283-292.

Nienhuis, J., M.K. Slocum, D.A. DeVos and R. Muren. 1993. Genetic similarity among *Brassica oleracea* genotypes as measured by restriction fragment length polymorphisms. J. Amer. Soc. Hort. Sci. 118:298-303.

Nienhuis J.,. J.B. dos Santos, J. Tivang and P. Skroch. 1995. Genetic distance among cultivars and landraces of lima beans (Phaseolus lunatus L.) as measured by random amplified polymorphic DNA markers J. Amer. Soc. Hort. Sci. 120:300-306.

Santos, J.B. dos, J. Nienhuis, P. Skroch, J. Tivang and M.K. Slocum. 1993. Comparison of RAPD and RFLP genetic markers in determining genetic similarity among *Brassica oleracea* L. genotypes. Theor. Appl. Genet. 87:909-915.

SAS Technical Report P-229, SAS/STAT Software: changes and Enhancements, Release 6.07.

Skroch, P., Tivang, J and J. Nienhuis. 1992. Analysis of genetic relationships using RAPD marker data, p. 26-30. In: Applications of RAPD technology to plant breeding. Crop Science Society of America, Madison, WI.

Smith O.S., J.S.C. Smith, S.L. Bowen, R.A. Tenborg and S.J. Wall. 1990. Similarities among a group of elite maize inbreds as measured by pedigree, F1 grain yield, grain yield, heterosis and RFLPs. Theor. Appl. Genet. 80:833-840.

Tivang, J., J. Nienhuis and O.S. Smith. 1994. Sampling variances of molecular marker data sets using the bootstrap. Theor. Appl. Genet. (in press)

Weeden, N.F., G.M. Timmerman, M. Hemmat, B.E. Kneen and M.A. Lodhi. 1992. Inheritance and reliability of RAPD markers, p.12-17. In: Applications of RAPD technology to plant breeding. Crop Science Society of America, Madison, WI.

Welsh, J. and M. McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Research. 18: 7213-7218.

Weir, B. S. 1990. Genetic data analysis: methods for discrete population genetic data. Sinauer Assoc., Inc. Sunderland, MA.

Wilkinson, L. 1992. SYSTAT: Statistics, Version 5.2. SYSTAT Inc., Evanston, Il.

Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rafalski and S.V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research. 18: 6531-6535.

Table 1. Contingency table showing the four possible outcomes of the comparison of two genotypes for RAPD markers.

| | genotype 1 | |
|---|---|---|
| genotype 2 | 1 | 0 |
| 1 | A<br>[1,1] | B<br>[0,1] |
| 0 | C<br>[1,0] | D<br>[0,0] |

Table 2. Form of the analysis of variance to compare differences between groups based on the consistency of marker freqencies within vs. between groups.

| Source | df[z] | Mean square | Expected mean square |
|---|---|---|---|
| Group | (p-1) | | |
| Genotypes(group) | (g-1)(p) | | |
| Markers | (m-1) | | |
| Markers x group | (m-1)(p-1) | $M_1$ | $\sigma^2_{mg/p} + m\sigma^2_{mp}$ |
| Markers x genotypes(group) | (m-1)(g-1)p | $M_2$ | $\sigma^2_{mg/p}$ |

[z] p, g and m refer the the numbers of groups, genotypes and markers respectively.

Table 3. Comparison between Central American red and black bean cultivars.

| Comparison | Number of comparisons[z] | Mean genetic distance[y] | F-value[x] |
|---|---|---|---|
| Red vs. black | 1435 | 0.268 | 9.54** |

[z] (35*41), where, 35 and 41 refer to the number of individual genotypes in each group.

[y] The mean gentic distances between all comparisons. Genetic distance was calculated as the ratio of disordant to total bands scored.

[x] F-value was calculated from the analysis of variance (Table 2) comparing the mean squares for consistency of marker freqencies within vs. between populations.

Table 4. Total marker variance, mean genetic distance and test of homogeneity of variance between Central American red and black bean genotypes .

| Population | Number of individuals | Total marker variance[z] | Mean genetic distance among individuals within population[y] |
|---|---|---|---|
| Red | 35 | 0.225 | 0.223 |
| Black | 41 | 0.210 | 0.206 |
| Significance test [x] | | $0.281^{ns}$ | |

[z] Total marker variance = $\Sigma$ ( 2n*p*q)/(n-1). The sum over all RAPD 80 loci scored for the variance of a binomial; where n= number of individuals in a population and p and q refer to the frequency of the presence or absence of a band, respectively.

[y] Gentic distance calculated as the ratio of discordant to total bands scored between two individuals.

[x] Nei's test of homogeneity of variance.

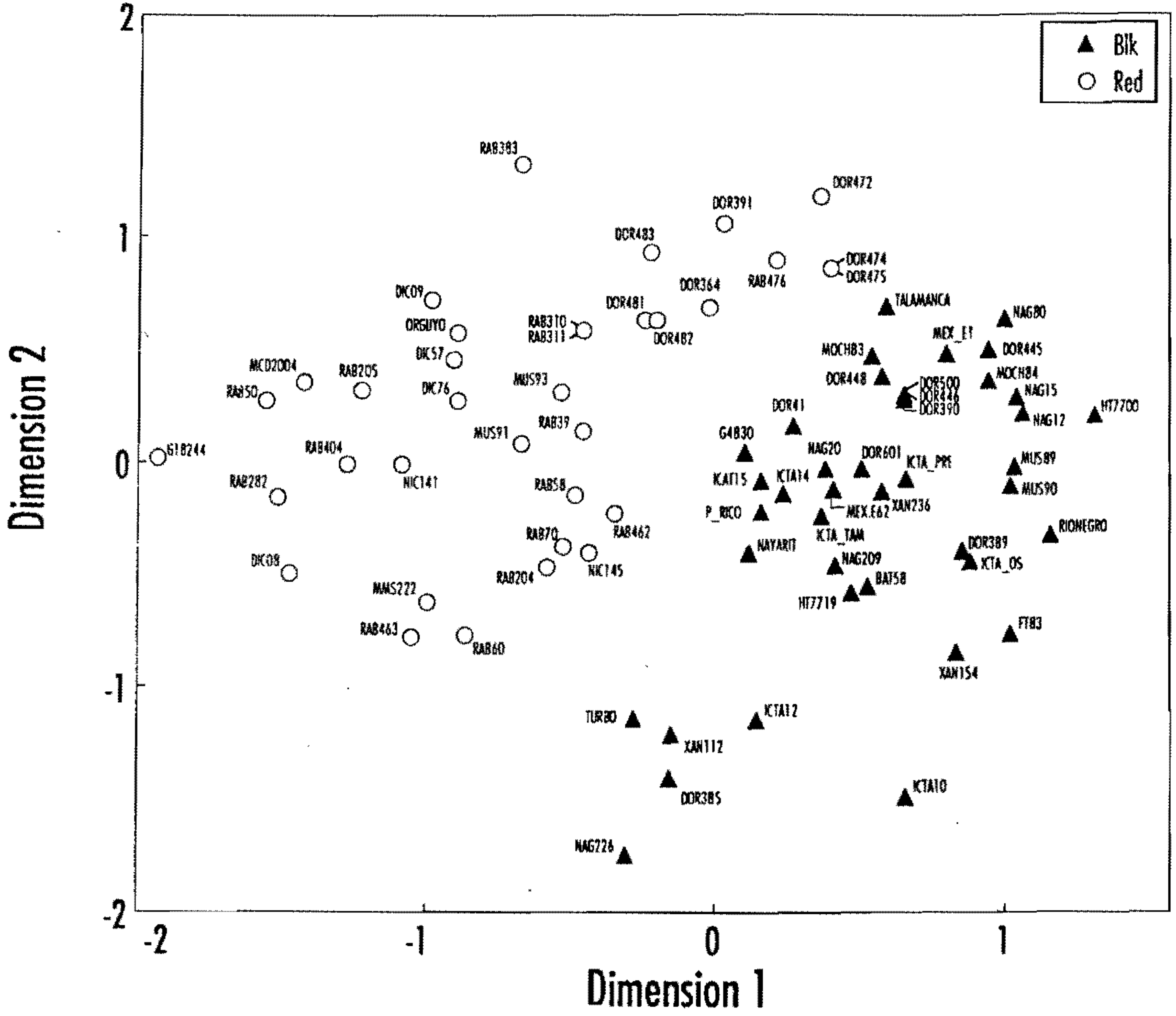


Fig. 1. Multidimensional Scaling of Central American Red and Black beans.

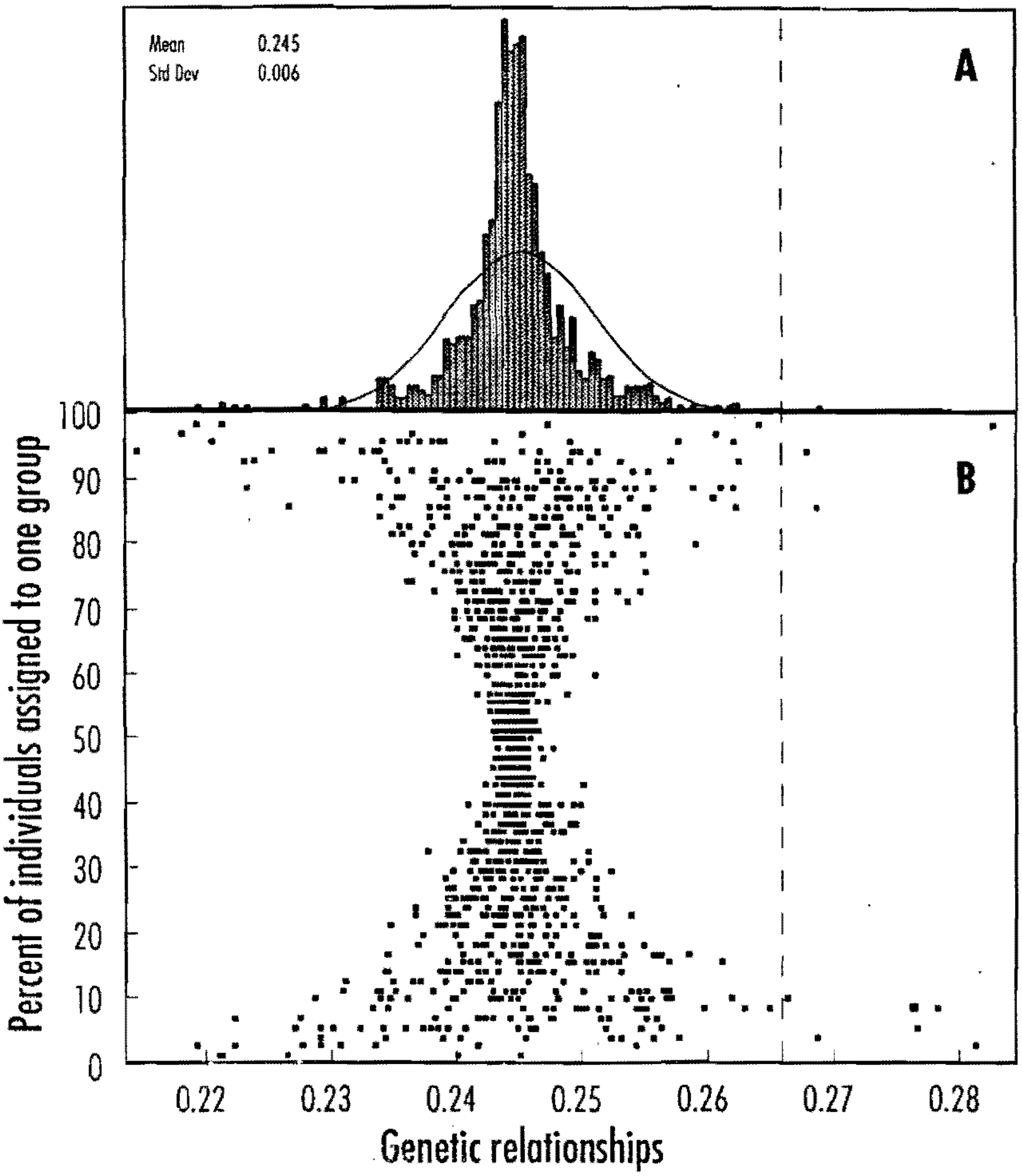


Fig 2 A and B. Both figures represent 10,000 random permutations of the genetic distance between two groups, in whic individual genotypes were randomly classified into either one of the groups. Figure A presents the distribution of genetic distance between the two groups. Figure B plots the genetic distance between the two groups vs. the percentage of individuals classified into one group. The observed distance between the red and black beans, 0.268, is indicated by the dotted vertical line.

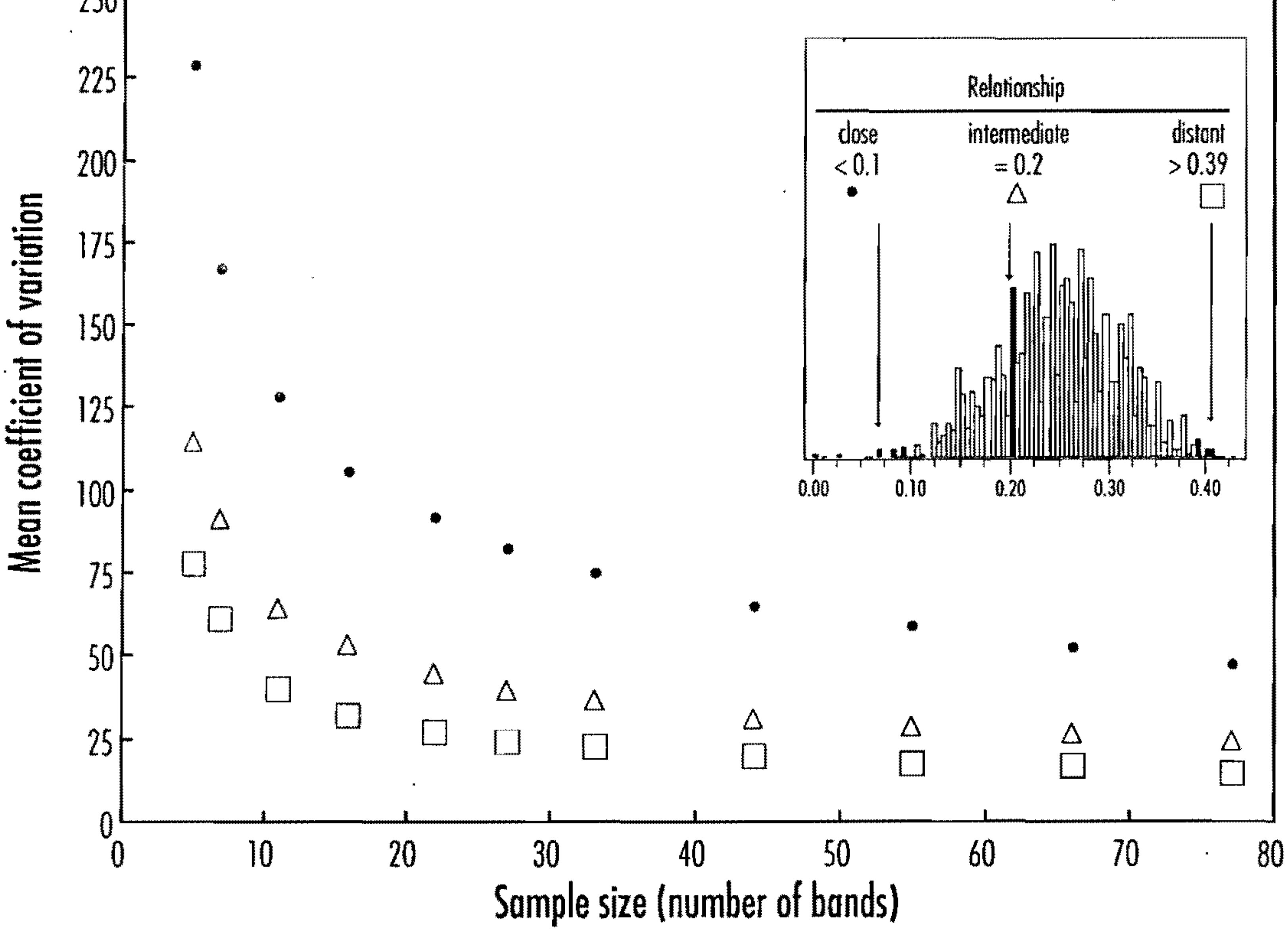


Fig. 3. Plot of the cefficient of variation for genetic relationships (GD) vs. sample size (band number) for three levels of relationship, close (GD<0.01), intermediate (GD=0.2) and distant (GD≥0.39). Approximately 40 distances were sampled at each level of relationship. Variance of GD at each sample size (band number) was calulated based on 200 bootstrap samples.

# STATISTICAL METHODS

# IN

# AGRICULTURAL EPIDEMIOLOGY

# Statistical Selection: Main Approaches and a Modification with a Preference Threshold

**F. P. A. Coolen**
**University of Durham**
**Department of Mathematical Sciences**
**Durham, England**
**and**
**P. van der Laan**
**Eindhoven University of Technology**
**Department of Mathematics and Computing Science**
**Eindhoven, The Netherlands**

## Summary

Statistical selection is discussed in general terms. In a certain sense statistical selection procedures are often more realistic than the usual testing and multiple comparison procedures in answering questions like "Which treatment can be considered to be the best?'. The approach of Bechhofer, the so-called Indifference Zone approach, as well as the approach of Gupta, the so-called Subset Selection approach, are presented. A comparison of these approaches is made using qualitative terms.

A short review is given on the use of a loss function as a generalization of selection of an $\epsilon$-best treatment. An $\epsilon$-best treatment is a generalization of a best treatment and is defined as a treatment on a "distance" less than $\epsilon$ of the best treatment. Finally, a generalization of the concept of the Indifference Zone selection is presented. The Indifference Zone approach is generalized by introducing a preference threshold. By this way there are three possibilities of decision: correct selection, false selection and no selection.

AMS Subject Classification: 62F07.

Key Words: Indifference Zone; Subset selection; probability of correct selection; probability of false selection; probability of no selection; loss function; $\epsilon$-best treatment; generalized selection goal; normal populations; preference threshold.

## 1. Introduction

In practice we are often confronted with the problem of selection of the best treatment or population. Especially in the field of biometry ( e.g. testing varieties ) selection is often an interesting feature. We assume that the populations are described by qualitative variables.

For all kinds of selection problems a quantitative methodology of selection is needed. The normal attack, using ANOVA techniques, is in some cases not completely adequate, in the sense that the formulation of the problem is not always realistic.

Let us consider the problem of selecting the best treatment from a number k ( integer $k \geq 2$ ) of treatments. The best treatment is defined as the treatment with the largest expected yield per unit plot. If there are more than one contenders for the best because there are ties, it is assumed that one of these is appropriately tagged.

To be sure, or almost sure, that we don't miss the best treatments, the probability of correct selection of the best treatment has to be taken into account. A better understanding of statistical selection will certainly improve the application of statistical selection procedures.

A short description of the basic approaches will be given in section 2. Some general remarks about modifications and generalizations as well as about comparison of both main approaches are made in section 3. In section 4 a short review is given of the use of $\epsilon$-best treatments and loss functions. A generalization of the Indifference Zone selection approach using a preference threshold is given in section 5. Finally, some concluding remarks are given in section 6.

## 2. Statistical selection: Main approaches

The two main approaches for selection of the best population or treatment are the Indifference Zone approach (see Gupta and Panchapakesan, 1979) and the Subset Selection approach (see Gupta and Panchapakesan, 1979). These two basic approaches will be shortly reviewed.

Assume k ( integer k ≥ 2 ) independent Normal random variables $X_1, \ldots, X_k$ are given. These variables are associated with the k populations or treatments indicated by $T_1, \ldots, T_k$, and are for instance sample yields. The assumed Normal distributions have common known variance $\sigma^2$ and unknown means $\Theta_1, \Theta_2, \ldots, \Theta_k$. The goal is to select the treatment with mean $\Theta_{[k]}$, where $\Theta_{[1]} \leq \Theta_{[2]} \leq \ldots \leq \Theta_{[k]}$ denote the ordered values of $\Theta_1, \Theta_2, \ldots, \Theta_k$. Let CS denotes correct selection.

The first approach is the so-called Indifference Zone approach, introduced by Bechhofer (1954). The goal is to indicate or select the best treatment. The selection rule is to select the treatment that resulted in the largest sample mean. The confidence or probability requirement is that the probability of a CS is at least P*, whenever the best treatment is at least $\delta^*$ away from the second best. In this context CS means that the best treatment with mean $\Theta_{[k]}$ produced the largest sample mean and consequently it is also selected as the best treatmemt. The minimal probability $P^*$ can only be guaranteed if the common sample size n is large enough.

The parameter space is defined as

$$\Omega = \{ \Theta \in \mathbf{R}^k : \Theta = ( \Theta_1, \Theta_2, \ldots, \Theta_k ) \}.$$

Bechhofer (1954) introduced the next measure of distance

$$\delta = \Theta_{[k]} - \Theta_{[k-1]}.$$

So the probability requirement

$$P(CS) \geq P^{*},$$

with $1/k < P^{*} < 1$, has to hold only for all $\Theta \in \Omega(\delta^{*})$, where the subspace $\Omega(\delta^{*})$ is defined as

$$\Omega(\delta^{*}) = \{ \Theta: \delta \geq \delta^{*} > 0 \},$$

the so-called Preference Zone.
$P^{*}$ and $\delta^{*}$ have to be specified by the experimenter.

The problem is to determine the common sample size n (= $n_i$ for i = 1, 2, ..., k) for which

$$\inf P(CS) \geq P^{*},$$

where the infimum is taken over the Preference Zone. For this location parameter case the Least Favourable Configuration (LFC) in the Preference Zone, where the P(CS) is minimal, is given by

$$\Theta_{[1]} = \Theta_{[k-1]} = \Theta_{[k]} - \delta^{*}.$$

In the case considered, where the observations $X_{ij}$ $(j = 1, 2, ..., n)$ on $X_i$ are Normally distributed with mean $\Theta_i$ and common known variance $\sigma^2$ (i = 1, 2, ..., k) and all $X_{ij}$ are independent of each other, one can find (Bechhofer, 1954) that

$$P_{LFC}(CS) = \int_{-\infty}^{\infty} \Phi^{k-1}(x+\tau)\, d\Phi(x),$$

with $\Phi$ (.) is the Normal cumulative distribution function and

$$\tau := \delta^* \sqrt{n} / \sigma.$$

If one requires that $P(CS \mid LFC) \geq P^*$, then the value of $\tau$ follows, and thus

$$n = (\tau \sigma / \delta^*)^2.$$

Tables for $\tau = \delta^* \sqrt{n} / \sigma$ can be found in for instance Gibbons, Olkin and Sobel (1977). With the chosen minimal n it can be guaranteed with minimal probability P* that the selected treatment is less than $\delta^*$ away from the best.

The second approach is subset selection, introduced by Gupta (1965). The subset selection procedure selects a subset, non-empty and as small as possible, with the probability requirement that the probability of a CS is at least P*. CS means in this context that the best treatment is an element of the selected subset. So

$$P(CS) \geq P^*,$$

where $1/k < P^* < 1$. The size S of the subset, defined as the number of treatments in the subset, is a random variable. The selection rule introduced by Gupta is defined as follows. Put treatment j, j = 1, 2, ..., k, into the subset if and only if

$$X^*_j \geq \max_{1 \leq i \leq k} X^*_i - d\sigma/\sqrt{n},$$

where $X^*_i$ denotes the sample mean of the ith sample and with d > 0. The Least Favourable Configuration is

$$\Theta_{[1]} = \Theta_{[2]} = \ldots = \Theta_{[k]}.$$

The probability requirement leads to

$$P(\ CS\ ) = P\ (\ X^*_{(k)} \geq X^*_{(i)} - d\ \sigma/\sqrt{n}\ ) \geq P^*,$$

with $X^*_{[i]}$ the ordered sample means and $X^*_{(i)}$ the sample means associated with the treatments corresponding with $\Theta_{[i]}$, i = 1, 2, ..., k.

Gupta (1954) has proved that the probability requirement leads to

$$P_{LFC}(CS) = \int^{\infty}_{-\infty} \Phi^{k-1}\ (x+d)\ d\ \Phi(x),$$

If we require that the left-hand side is equal to $P^*$, then we find the value of d. The selection constant d is in this way determined in such a way that $P(CS) \geq P^*$ for all possible parameter configurations. Tables with values of the selection constant d can be found in Gibbons, Olkin and Sobel (1977).

## 3. Modifications and a global comparison of both main approaches

In the literature different goals are considered. We mention a few:

i. Selecting the t best treatments, where integer t is larger than or equal to 2. We can do this with or without ordering of the treatments. In the first case we indicate a treatment as the best one, another treatment as the second best, etc.

In the second case we produce a collection of t treatments without ranking them.

ii. Selecting a subset that contains only good treatments.

iii. Selecting a collection of treatments which will contain at least the t ( $t \geq 2$ ) best treatments.

iv. Selecting a random number of treatments such that all treatments better than a standard treatment are included in the selected subset.

v. Selecting a subset whose size is smaller than or equal to m ( $1 \leq m < k$ ) and which will include at least one good treatment.

In the literature different generalizations and modifications have been proposed. We refer to Gupta and Panchapakesan (1979) for references. An interesting result has been achieved by Hsu (1981, 1984). Selecting a subset with confidence P* for containing the best, he also gives simultaneously confidence intervals for the differences of the treatments compared with the best one.

Subset selection is a flexible form of selection, because the number of replications has not to be determined in advance. After the experiment has been carried out, the selection can be prosecuted. The influence of the number of replications can be conducted from the (expected) size of the subset. A relatively large subset means, apart from random fluctuations, that the number of replications is small or the treatment means are close together, or both. If a correct selection CS is defined as the event that the best treatment is in the subset then the probability on CS can be compared with the power of a test. Both characteristics indicate the probability on a correct decision while the treatments may be (or are) different. Whereas subset selection can be used as a screening procedure, the indifference zone approach produces, in a certain sense, a more precise result. For the last method indicates the best treatment, at least we hope so. A condition is that a minimal number of observations have been done. The probability distribution of the size S, the number of treatments in the selected subset, can be found in van der Laan (1995). Also the distribution, expectation and variance of S for the LFC are given.

## 4. A short review of selection using the concept of an $\epsilon$-best treatment or a loss function

The requirement to select the best treatment may be a strong one if the best treatment is not far away from the other treatments. For example, the difference of the treatment mean of the best and the second best is relatively small. So the corresponding treatments are close together, say on a distance less than $\epsilon$, where $\epsilon > 0$. In such a situation it is often not of practical interest whether one selects the best one or the next best. Not every difference in the observed response is important. In many real world problems it is of interest to see whether the best or an almost best treatment can be selected. A treatment $T_i$ is called $\epsilon$-best if and only if :

$$\Theta_i \geq \max_{i \leq j \leq k} \Theta_j - \epsilon .$$

It is possible that there are more than one $\epsilon$-best treatment. This idea leads to the consideration that one may generalize the selection goal to selection of an $\epsilon$-best treatment. A consequence of this generalization is that the least favourable configuration becomes more difficult. Not an essential disadvantage using computers. The goal is to select a small but non-empty subset such that the selected subset will contain the best treatment or an $\epsilon$-best one with a certain confidence. In general, the generalization to selecting an $\epsilon$-best treatment will result in subsets of smaller expected size. Some aspects of selecting an $\epsilon$-best treatment has been considered in van der Laan (1990). The gain in efficiency, comparing selection of the best and selection of an $\epsilon$-best treatment, depends on the value of $\epsilon$.

The concept of $\epsilon$-best treatment has been generalized using a loss function by van der Laan and van Eeden (1995). They considered loss functions of the form:

$$L(\Theta,d,\epsilon)= h(\Theta_{[k]} - \epsilon - \max_{i\epsilon d} \Theta_i)\ I(\max_{i\epsilon d} \Theta_i < \Theta_{[k]} - \epsilon),$$

where h(.) is a non-decreasing function on $R^+$ and d is the selected subset of {1, 2, ..., k}. The special case of two Normal treatments, with k=2, $n_1 = n_2 = n$, $\sigma = 1$ and the next form of the loss function

$$L(\Theta,d,\epsilon) = ( \Theta_{[2]}-\epsilon-\Theta_{[1]})^p\ I( \max_{i\epsilon d} \Theta_i < \Theta_{[2]}-\epsilon )$$

has been considered by them in great detail. The decision rule considered by them was of the form:

$$\delta_c(x) = \begin{array}{lll} \{1\} & \text{if} & X^*_1 - X^*_2 > c \\ \{2\} & \text{if} & \quad " \qquad " \quad < -c \\ \{1,2\} & \text{if} & | \quad " \qquad " \quad | \le c. \end{array}$$

Among other things they give explicit expressions and tables for the Risk and the expected value of S. It is possible to find valus for c meeting the requirements

$R \le R^*$ fc- certain values of $\mu=|\Theta_1-\Theta_2|$

and

$ES \le 1+\gamma$ " "

with $\gamma \in (0,1)$.

## 5. A generalization of the indifference zone approach using a preference threshold

It is possible to generalize, in a certain sense, the concept of indifference zone selection by introducing a preference threshold ( Coolen and van der Laan, 1995 a,b). Again we consider k independent samples of common size from normal populations with equal known variance. Starting with the preference zone given by the parameter subspace

$$\Omega(\delta^* ) = \{ \Theta = (\Theta_1, \Theta_2, \ldots, \Theta_k) \in \mathbf{R}^k : \\ \Theta_{[k]} - \Theta_{[k-1]} \geq \delta^* > 0 \},$$

we accept three kinds of decisions, namely CS, FS ( = false selection) and NS ( = no selection). We apply the selection rule $R_c$ : Select population i if and only if

$$\Sigma^n{}_{j=1} (X_{ij} - X_{lj}) > c,$$

for $l = 1, 2, \ldots, k$; $l \neq i$, with the so-called threshold $c \geq 0$.

It is possible to require that, for instance, for the parameter configuration $\Omega(\delta^*)$ the following holds

$$P(CS) \geq P^* \quad \text{and } P(FS) \leq Q^*.$$

These two conditions turn out to be

$$\int^{\infty}_{-\infty} \Phi^{k-1} (z+\tau_{c,k}) \ d\Phi(z) = P^*$$

and

$$(k-1) \int^{\infty}_{-\infty} \Phi^{k-2} (z-\tfrac{1}{2}(\tau_{f,k}-\tau_{c,k})) \ \Phi(z-\tau_{f,k}) \ d\Phi(z) = Q^*$$

with $\tau_{c,k} = (n\delta^* - c)/(\sigma\sqrt{n})$ and $\tau_{f,k} = (n\delta^* + c)/(\sigma\sqrt{n})$. The two conditions determine these two constants, and lead to

$$n = \{ \sigma ( \tau_{c,k} + \tau_{f,k} ) / (2\delta) \}^2$$

$$c = \sigma^2 (\tau^2_{f,k} - \tau^2_{c,k} )/(4\delta).$$

These n and c are such that in $\Omega(\delta^*)$ both probability requirements are satisfied for given $\delta^*$, $P^*$ and $Q^*$, when using selection rule $R_c$. If c=0 we get the standard indifference zone selection procedure. In this last case, c=0, $\Omega(\delta^*)$ is the least favourable configuration for P(CS).

## 6. Some final remarks

During decades of years we are used to apply statistical tests, like analysis of variance tests, to problems that are real selection problems. We think it is important to investigate the possibilities to use statistical selection procedures for certain problems. The first thing we need is to formulate adequately the problem. If the problem we consider is a selection problem, then it must be formulated as a selection problem. A quotation of John Tukey said: `An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem'. Following this statement an exact formulation as a selection problem is worthwile and then an analysis (exact or approximate) is required. Not for all designs of experiments this problem has been solved. For a number of designs a simulation is feasible in order to find an accurate estimate of the selection constant required for the prosecution of the selection procedure.

We refer to Gupta (1977), Gibbons, Olkin and Sobel (1979) and Dudewicz (1980) for an introduction to statistical selection.

The essential aspect of a statistical analysis of observations from designed experiments is to keep the probability of an error under control. The use of selection procedures with a certain confidence requirement can help us to study practical problems in an realistic way.

## References

Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Annals of Mathematical Statistics **25**,16-39.

Coolen, F.P.A. and van der Laan, P. (1995a), On Indifference Zone selection with a preference threshold. Memorandum COSOR **95-02**, Department of Mathematics and Computing Science, Eindhoven University if Technology.

Coolen, F. P. A. and van der Laan, P.(1995b), On indifference zone selection with a preference threshold. Proceedings 50th Session of the International Statistical Institute, Beijing, 21 - 29.

Dudewicz, E.J. (1980). Ranking ( ordering ) and selection: An overview of how to select the best. Technometrics **22**, 113-119.

Gibbons, J.D., I. Olkin, M. Sobel (1977). Selecting and Ordering Populations: A New Statistical Methodology. Wiley, New York.

Gibbons, J.D., I. Olkin and M. Sobel (1979). An introduction to ranking and selection. The American Statistician **33**, 185-195.

Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. Technometrics **7**, 225-245.

Gupta, S.S. (1977). Selection and ranking procedures: a brief introduction. Commun. Statist. Theor. Meth. **A6**, 993-1001.

Gupta, S.S. and S. Panchapakesan (1979). Multiple Decision Procedures. Wiley, New York.

Hsu, J.C. (1981). Simultaneous confidence intervals for all distances from the `best'. Annals of Statistics **9**, 1026-1034.

Hsu, J.C. (1984). Constrained simultaneous intervals for multiple comparisoms with the best. Annals of Statistics **12**, 1136-1144.

Laan, P. van der (1990). Subset selection of an almost best treatment. Biom. J. **34**,, 647-656.

Laan, P. van der (1995). Distribution theory of subset selection. Submitted to J. Statist. Plann. Inf.

Van der Laan, P. and Van Eeden, C. (1993), Subset selection with a generalized selection goal based on a loss function. Memorandum COSOR **93-15**, Department of Mathematics and Computing Science, Eindhoven University of Technology.

Van der Laan, P. and C. Van Eeden (1995). Subset selection of the best of two normal populations using a loss function. Revised version of Van der Laan and Van Eeden (1993). Memorandum COSOR **95-15**, Department of Mathematics and Computing Science, Eindhoven University of Technology.

# THE IMPORTANCE OF EPIDEMIOLOGICAL RESEARCH FOR TROPICAL AGRICULTURE

**F. J. Morales** and **P. K. Anderson**, Head, Virology Research Unit, Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia, and Field Coordinator, Agricultural Institute of Canada (AIC) , Virus Epidemiology Program, National Agrarian University (UNA), Managua, Nicaragua, respectively.

## The Changing Agricultural Environment

The main characteristic of basic subsistence agriculture in the tropics, has been the cultivation of a wide range of native plant species of diverse genetic background. Farmers cultivated these species in small, scattered fields, which were often left fallow, thus, minimizing the build-up of plant pathogens. While some of these features still exist in isolated ecosystems, the modernization of tropical agriculture has followed an accelerated trend away from native species and genetic diversity. Today, a limited number of crops are found in virtual monoculture throughout the tropics, creating optimal conditions for the onset of severe disease epidemics.

In response to this challenging phytosanitary situation, the strategy followed by most national agricultural research institutions (NARIs) and international agricultural research centers (IARCs) in the tropics, has been breeding for disease resistance. Unfortunately, the complexity of new cropping systems in the tropics and increasing disease problems, have greatly overburdened the limited breeding capacity of NARIs and IARCs alike. Consequently, farmers continue to apply unnecessary amounts of agrochemicals as a risk-avoiding measure, causing pesticide resistance problems, human health hazards and a progressive contamination of the environment.

An example of a changing agricultural environment and its effect on the incidence of plant disease, yield loss and crop protection, is that of the expansion of soybean cultivation in agricultural regions of Latin America, previously devoted to the cultivation of traditional food crops. Soybean is a preferred host for the whitefly *Bemisia tabaci* (Gennadius). With soybean expansion, the whitefly population increased dramatically and subsequently migrated into other crops, including *Phaseolus* beans, causing an epidemic of bean golden mosaic virus that resulted in up to 85% yield loss in Brazilian bean production (Costa 1975, Costa & Cupertino 1976). In the absence of resistant cultivars and integrated pest and disease management practices, agrochemicals remain the most widely used method of crop protection. However, the use of insecticides for managing insect vectors such as *B. tabaci* is problematic. This whitefly species has become resistant to most insecticides, resulting in an increasing number of insecticide applications, and use of more costly and toxic systemic insecticides not registered for crop production. The recommendation of alternative strategies, i.e. in this case non-chemical management of the whitefly vector, is the work and outcome of basic research in botanical epidemiology.

## Botanical Epidemiology

Epidemiology is the science that studies disease in populations (Vanderplank 1963). Epidemiology can be classified according to the host of primary interest, that is, human, veterinary or botanical epidemiology (Zadoks 1974). Botanical epidemiology studies plant disease in populations.

The science of epidemiology can be divided into four research areas: 1) circumstantial, 2) etiological, 3) ecological and 4) mathematical epidemiology (MacDonald 1957, Anderson 1991). Circumstantial epidemiology describes the disease and the circumstances under whic:. it occurs. More specifically, circumstantial epidemiology describes the patterns of disease in time and in space. Etiological epidemiology deals with the identification of the causal agent and its mode of transmission or dispersal. Ecological epidemiology studies the biology of the pathogen and any other organism involved in disease transmission, e.g. the alternative hosts, insect vectors, as well as the relationship between the organisms in the transmission cycle and environmental factors that affect them.

meetings of the Brazilian and Latin American Phytopathological Societies, revealed that only 4/464 and 3/136 papers presented a those meetings, respectively, dealt with epidemiology.

The third detrimental outcome of the largely unsuccessful attempts at mathematical epidemiology in the 1970s, is a misunderstanding of the contribution of quantitative work to epidemiological research. Mathematical epidemiology relies fundamentally on mathematical models for the purposes of integration and analysis. The contribution of statistics to mathematical epidemiology is secondary. But, in the branches of circumstantial and ecological epidemiology, the reverse is true. The critical quantitative component is statistical analysis; mathematical analysis plays a secondary role. This is particularly true for circumstantial epidemiology. However, because mathematical epidemiology has been equated with epidemiology, circumstantial and ecological epidemiology have received less attention and, consequently, the development and utilization of modern statistical analysis as an aid in epidemiological research, has been slow.

## Examples of the Contribution of Statistical Analysis to the Development of Botanical Epidemiology

New plant diseases are emerging, world-wide, at an alarming rate (Anderson & Morales 1994). Description and analysis of spatial patterns of disease can provide important insight into the etiology and ecology of plant disease. Historically, spatial patterns have been analyzed by simple descriptive statistics, e.g. mean and variance. However, the application of more advanced statistical methods has enabled more sophisticated analysis of disease patterns.

Geostatistics, originally developed by geologists to analyze spatial distribution of soil types, has recently been applied to plant diseases (Lecoustre & Reffye 1986, Chemelli *et al. 1988*, Lecoustre *et al.* 1989, Larkin *et al.* 1995, Dandurand *et al.* 1995). Geostatistics goes beyond a *general* description of patterns to quantify and analyze spatial dependence, relationships and dynamics. Lecoustre *et al.* (1989) illustrate this concept in a simple example: "Two series of measurements

Mathematical epidemiology integrates an analyzes the complex sets of data generated by the other research areas, primarily through the use of mathematical models. One outcome of mathematical analysis is to suggest the most epidemiologically effective strategies for crop protection.

In 1963, J. E. Vanderplank, in his classic work "Plant Diseases: Epidemics and Control", introduced the first mathematical model for plant disease. Although he emphasized the need for botanical epidemiology to evolve from descriptive and experimental research to analytical and mathematical epidemiology, the role of mathematics and modelling did not become an academic concern until the 1970s.

The popularity of mathematical epidemiology resulted in an influx of mathematically-minded students into the field of plant pathology. This trend had several detrimental outcomes. First, the science of epidemiology became equated with the branch of mathematical (or "quantitative") epidemiology. The introductory virology text by Bos (1983) reflected the predominant belief that "quantitative aspects of the ecology of diseases...are dealt with in epidemiology". Consequently, the other, non-mathematical, branches of epidemiology were neglected. In particular, the understanding of complex biological and ecological phenomena underlying pathogen-host-vector interactions, made little progress.

Second, the flush of mathematically-oriented publications were unintelligible to most plant pathologists and crop protectionists who were responsible for linking the analytical conclusions to crop protection practices. There were several overwhelming success stories, such as the development and implementation of the EPIPRE model, which is still used throughout Europe, for predicting yield losses from six fungal diseases and aphid pests in wheat (Zadoks 1988). However, the successes were the exception, not the rule. The general view that epidemiology is quantitative and technique-oriented, has produced a distorted and partial view of the potential applications of epidemiology. Furthermore, the belief that this quantitative work is esoteric and not practical, has driven many plant pathology students away from the science of epidemiology. This has been particularly true of students from developing countries. A review of the papers presented at the 1994 plant pathology made of a hypothetical variable at regular intervals along a row in a field gave the following numerical sequences: A: 1-2-3-4-5-6-5-4-3-2-1 and B: 1-4-3-6-1-5-3-4-2-5-2. Sequence A has a clearly defined symmetry, whereas any structure for sequence B is irregular and difficult to define. Nevertheless, the two series of measurements have the same mean and variance; thus it is impossible to adequately describe the detailed spatial distribution of the variable by using only these two parameters".

Geostatistics allows for analysis of the structure of spatial pattern of disease. The resulting inference that disease is random or spatially dependent, gives insight to the etiology of disease and may lead to hypotheses on the direction and rates of spread, source and proximity of infection, vector type, vector mobility, etc.

Analyses of temporal disease patterns have also benefitted from applications of more advanced statistical methods. Traditional research in epidemiology has often involved the quantification of parameters, such as disease incidence or insect vector abundance, and the collection of weather data, usually temperature and precipitation. Attempts were made to accurately describe the changes in disease incidence or vector populations by simple correlation or regression analyses. However, because temporal patterns can be extremely complex, often simple statistical techniques are not sufficiently powerful as analytical tools. The application of more sophisticated statistics, even with limited temperature and precipitation data, can generate useful information.

For example, Madden et al. (1983) collected data for maize dwarf mosaic virus (MDMV) from experimental maize plots with varying disease incidence in Ohio. Ambient temperature and precipitation data were obtained from the local weather service They defined 18 environmental variables, which they thought might be influencing disease incidence. Applying stepwise discriminant analysis, they concluded that no one environmental variable could be correlated to the temporal disease patterns observed, but that three environmental variables could significantly describe low, medium or high disease intensity situations. This analysis permitted the development of a pre-planting predictive system to control maize dwarf mosaic in Ohio.

## The Importance of Developing Botanical Epidemiology for Crop Protection in the Tropics

The production of food in developing countries is a critical issue with long-lasting social and economic consequences. Global food security is a contemporary concern that must be addressed in light of the changing economic scenarios that favor export crops over staple food production. The deterioration of natural resources and the contamination of the environment are also key problems that need to be corrected in developing countries, where agrochemicals are abused due to the lack of integrated disease management practices.

Most developing countries are dedicated to the production of export crops. Some of the non-traditional export crops have been recently introduced from temperate countries together with exotic pathogens, and without a previous evaluation of their reaction to local pathogens and pests. The expected phytosanitary problems that have appeared as a result of these contaminated germplasm introductions, have diverted the limited human and financial resources of national research institutions away from the traditional crops. Moreover, the expansion of export crops in general, has greatly disturbed traditional cropping systems throughout the tropics, creating outbreaks of previously unrecorded diseases of food crops.

It is necessary to expect and to prepare for new disease outbreaks in disturbed agricultural environments, in response to changing economics that govern agricultural production policies in industrialized nations and developing countries alike (Morales 1992). The study of plant epidemics, including the monitoring, modeling and forecasting of epidemics of recognized plant pathogens as well as new plant pathogens, has been a .otally neglected science in Latin America and is only an emerging area of pursuit in the industrialized nations. The identification of disease-causing agents (etiological epidemiology), and knowledge of the distribution of plant pathogens, their vectors and their reservoirs in time and space (circumstantial epidemiology) is the *sine qua non* of a program for monitoring emerging plant diseases. Once the distribution of plant pathogens, their vectors and reservoirs is known, then, quarantine measures and monitoring efforts can be implemented to minimize the risks associated with the establishment and development of subsequent epidemics (Bos 1994).

However, the understanding of spatial and temporal patterns, i.e. where diseases occur and how and why they vary, will not improve without active collaboration from biometricians utilizing modern statistical methods. The analysis of biological, ecological and environmental phenomena and disease patterns in time and space, is a complex undertaking that will only be achieved through transdisciplinary collaboration of plant pathologists, entomologists, botanical epidemiologists, meteorologists, statisticians, and biomathematicians. Transdisciplinary collaborative research implies a basic knowledge of each other's discipline and, more importantly, the participation of the collaborators, including biometricians, in all stages of planning and execution of epidemiological research.

## Literature Cited

Anderson, P. K. 1991. Epidemiology of insect-transmitted plant pathogens. Doctoral Thesis. Harvard University, Boston, MA. 305 pp.

Anderson, P. K., and Morales, F. J. 1994. The emergence of new plant diseases: the case of insect-transmitted plant viruses. Annals of the New York Academy of Sciences 740:181-194.

Bos, L. 1983. Introduction to Plant Virology. Centre for Agricultural Publishing and Documentation, Wageningen, The Netherlands. 160 pp.

Bos, L. 1994. Environment and disease: ever-growing concern. Environmental Conservations 21:99-102.

Chellemi, D. O., Rohrback, K. G., Yost, R. S., and Sonoda, R. M. 1988. Analysis of the spatial pattern of plant pathogens and disease plants using geostatistics. Phytopathology 78:221-226.

Costa, A. S. 1975. Increase in the population density of *Bemisia tabaci*, a threat of widespread virus infection of legume crops in Brazil. Pages 27-41 in: Tropical Diseases of Legumes. J. Bird and K. Maramorosch, Eds. Academic Press, New York.

Costa, A. S., and Cupertino, F. P. 1976. Avaliacao das perdas na producao do fejoeira causada pelo virus do mosaico dourados. Fitopatologia Brasileira 1:18-25.

Dandurand, L. M., Knudsen, G. R., and Schotzco, D. J. 1995. Quantification of *Pythium ultimum* var. *sporangiiferum* zoospore encystment patterns using geostatistics. Phytopathology 85:186-190.

Larkin, R. P., Gumpertz, M. L., and Ristaino, J. B. 1995. Geostatistical analysis of *Phytophthora* epidemic development in commercial bell pepper fields. Phytopathology 85:191-202.

Lecoustre, R., and De Reffye, R. 1986. The regionalized variable theory: possible application to agronomical research, in particular to oil palm and coconut, with respect to epidemiology. Oleagineux 41:541-458.

Lecoustre, R., Fargette, D., Fauquet, C., and De Reffye, P. 1989. Analysis and mapping of the spatial spread of African cassava mosaic virus using geostatistics and the kriging technique. Phytopathology 79:913-920.

Madden, L. V., Knoke, K. J., and Louie, R. 1983. Classification and prediction of maize dwarf mosaic intensity. Pages 283-242 in: Proceedings, International Maize Virus Disease Colloquium and Workshop, 2-6 August 1982. D. T. Gordon, J. K. Knoke, L. R. Nault and R. M. Ritter, Eds. The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, MA, USA.

Morales, F. J. 1992. Viruses and the changing agricultural environment in the lowlands of Latin America. Pages 67-68 in: Fifth International Plant Virus Epidemiology Symposium on Viruses, Vectors and the Environment, Bari, Italy 27-31, July 1992.

Vanderplank, J. E. 1963. Plant Diseases: Epidemics and Control. Academic Press, New York. 349 pp.

Zadoks, J. C. 1974. The role of epidemiology in modern plant phytopathology. Phytopathology 64:918-923.

Zadoks, J. C. 1988. EPIPRE: a computer-based decision support system for pest and disease control in wheat: Its development and implementation in Europe. Plant Disease Epidemiology 2:3-39.

# EPIDEMIOLOGIA AGRICOLA: UN ENFOQUE CUANTITATIVO DE LA FITOPATOLOGIA.

Aurelio Pedroza S. Unidad Regional de Zonas Aridas, UACH. Bermejillo, Dgo. CP. 35230.

## SUMMARY.

A new technological approach in crop production is required in order to increase a sustainable food production, it keeping the agroecosystem balance. A new approach to do agriculture needs taking adventage of the inter and multidisciplinary participation of different specialities in emergent areas, such as, System Theory, Biotechnology, Environment Impact, Geographical Information Systems, Sustainable Agriculture, Expert Systems, and Plant Disease Epidemiology. Epidemic triangle host-pathogen-environment has been recognized as an integrated factor in plant disease epidemiology, which, currently is a useful phytopathological tool to quantify the plant diseases through temporal and spatial analysis, and it contributing in the design of forecaster systems to prevent the damage caused by plant pathogens. Important surveys about Comparative Epidemiology were conducted as a result of development in plant disease epidemiology in order to achieve better strategies in plant disease management.

**KEY WORDS:** Plant Disease Epidemiology, Temporal and Spatial Analysis.

## RESUMEN.

La producción sostenida de cultivos requiere de un nuevo enfoque tecnológico, tal que, permita incrementar la producción de alimentos sin alterar sustancialmente el agroecosistema. Este nuevo enfoque de hacer agricultura debe estar basado en la participación inter y multidisciplinaria de diferentes especialidades en areas amergentes, como: Teoria de Sistemas, Biotecnología, Impacto Ambiental, Sistemas de Información Geográfica, Agricultura Sustentable, Sistemas Expertos y la propia Epidemiología Agrícola. El triángulo epidémico patógeno-hospedante-ambiente se ha reconocido como la enfermedad en forma integrada, a traves de la Epidemiología Agrícola , la cual es actualmente una base importante para el análisis temporal y espacial de las epidemias, asi como para el diseño de sistemas de predicción para prevenir el impacto nocivo de las enfermedades. En los últimos años, se han hecho importantes contribuciones en estudios de Epidemiología Comparada, lo cual ha contribuido al diseño de mejores estrategias en el manejo de enfermedades en la agricultura.

**PALABRAS CLAVE:** Epidemiología Agrícola, Análisis Temporal y Espacial.

# PLANT DISEASE EPIDEMIOLOGY: A QUANTITATIVE APPROACH ON PHYTOPATHOLOGY

Aurelio Pedroza Sandoval, Unidad Regional Univeristaria de Zonas Aridas, UACH. Bermejillo,
Dgo. CP 35230.

## I. INTRODUCTION.

Current world population is over 5,000 millions people, with a proyection to up 12,000 millions for the next 50 years,therefore increasing the world food deficit (l2). A few decades ago, during the "green revolution", food production was increased exponentially; with higher cuantities of inputs (pesticides, fertilizers, etc.), the harvesting of several crops was increased. However, in the last two decades, the relationships between inputs-food production is asyntotical, thus to higher cuantities of inputs corresponded minimum or nules increments in crop production (2).

Crop production needs a new technological approach in order to achieve an increase in sustainable crop production, and in keeping the agroecosystems balance. This new approach in agriculture, requires of joining efforts and knowledge, and taking adventage of the inter and multi-disciplinary participation of different specialities in emergent areas, such as System Theory, Biotechnology, Environment Impact, Geographical Information Systems, Sustainable Agriculture, Expert Systems, and Plant Disease Epidemiology.

## II. PLANT DISEASE EPIDEMIOLOGY.

Phytopathology story is divided in three steps. First,Phytopathology was referred to plant disease etiology; later, it was focused to the plant as host of pathogen and; finally it was related to host-pathogen relationships influenced by the environment, as an epidemiological triangle host-pathogen-environment (fig. 1). Environment is recognized as an inductor and predispositional factor in plant pathosystems (7). Almost, parallel to the "Environmental Epidemiology" approach, started "Quantitative Epidemiology" (13). rate, Plant Disease Epidemiology is a useful phytopathological tool to quantify plant deseases through temporal and spatial analysis (2). Important surveys about Comparative Epidemiology were conducted as a result of development in Plant Disease Epidemiology (3), supporting the design of new and more useful strategies in plant disease management in crop production (11).

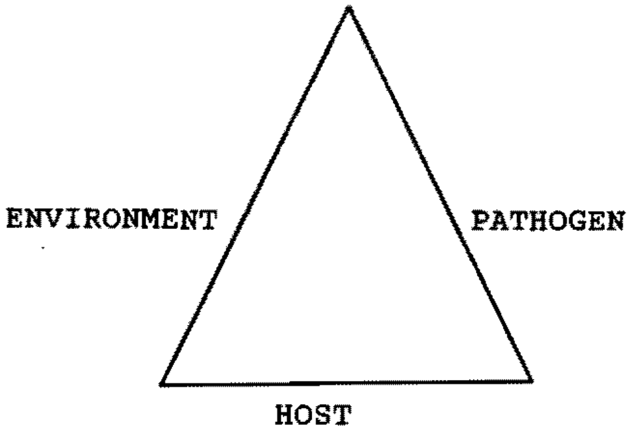


Fig. 1. Triangle of plant disease epidemic.

## III. EPIDEMIOLOGICAL ANALYSIS.

In the evaluation of epidemics, one of the most important variable to rate is the plant disease intensity (incidence or severity) either for plant disease management, or plant disease research. Incidence as a proportion of diseased plants, related to total of evaluated plants; severity as a proportion of diseased tissue, related to the total of plant tissue. With preliminary sampling we can to estimate the, accuracy and reliability of the evaluators. Sample size depends on the host-pathogen relationships; for instance, whether a pathogen is in an aggregated pattern, a sampling in X or W (fig.2), and a bigger sample size, could be are required.

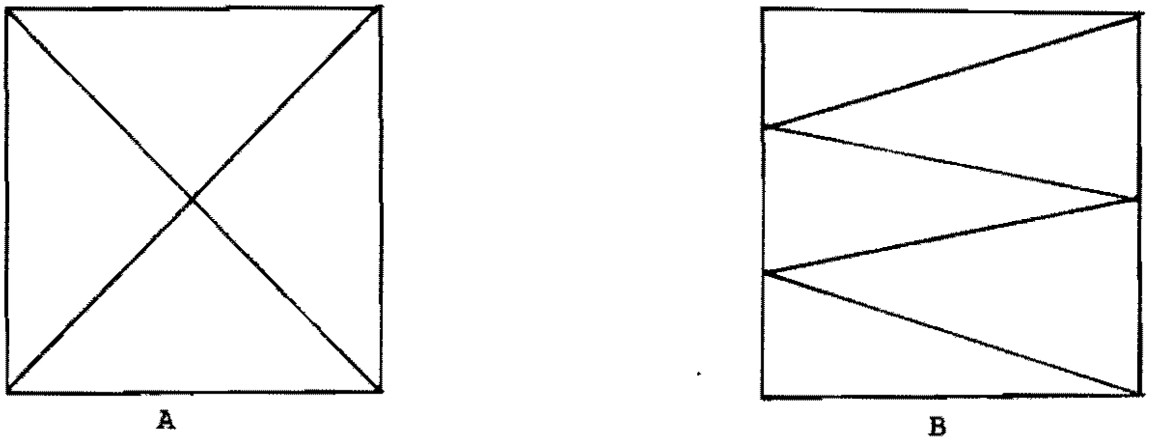


Fig. 2. Sampling pattern in X (A) and W (B).

Initial time of plant disease intensity (Xo=BD); progressive time of plant disease (Xp=DT); maximum time of plant disease (Xa=P1P2); decreasing time of plant disease (Xd=P2D; time of plant disease (Xt=DF); slop (b); initial plant disease (Yo); maximum plant disease (Ymax); final plant disease (Yf) and area under the disease progress curve (AUDPC) (Fig. 3), are other epidemiological variables, which are either independent or intercorrelated, depending of the host-pathogen relatioships. Principal Components (4) and Factor Analysis (5,9) are used to identify the relationships among variables.

In addition, stages of growth (Fig. 4) and micrometeorological variables, such as atmosphere humidity in the canopy, foliar humidity, maximum and minimum temperatures in the canopy, are important when one wants to know the cause and effect relationships in plant disease epidemiology (10).

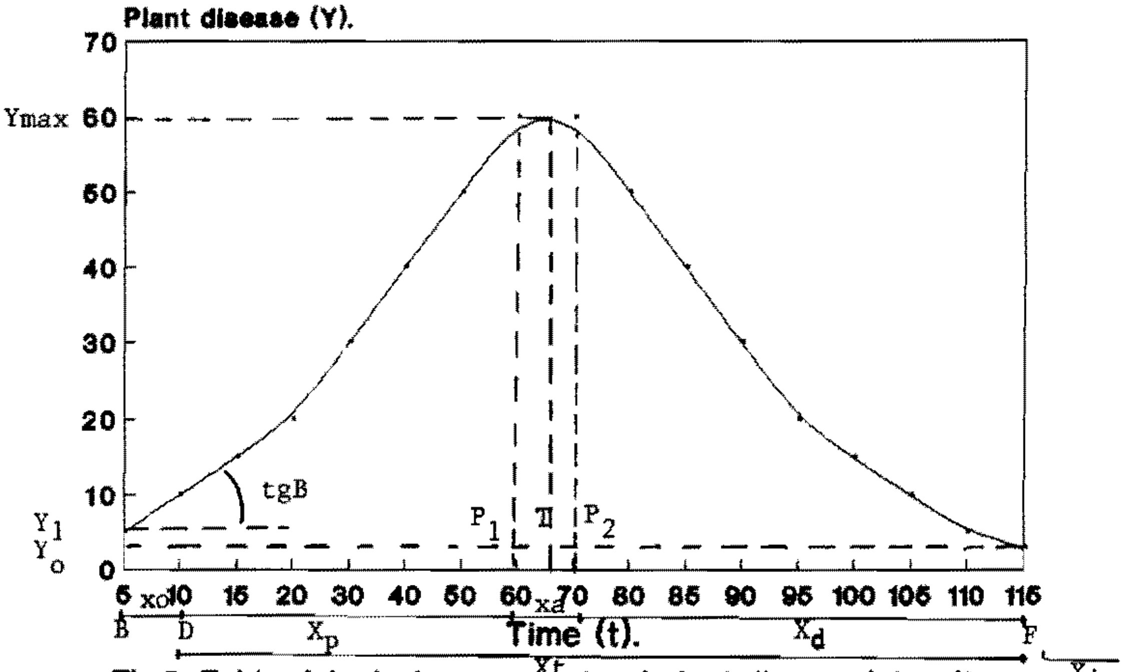


Fig.3. Epidemiological components of plant disease intensity

## IV. PLANT DISEASE MODELING.

Temporal dynamics of the plant disease is one of the most important epidemiological analysis, where plant disease modeling is common. Frequently, plant disease curve trends to take an S form, and the intensity of epidemic (incidence or severity) over time is analyzed with the fitted model (Fig.5).

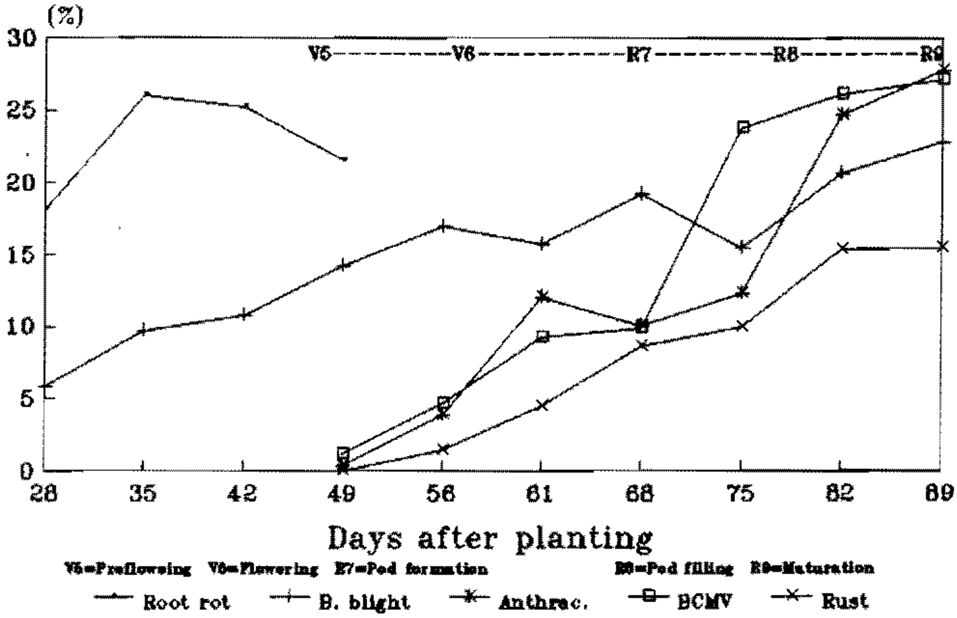


Fig.4. Mean of bean disease severities at different growth stages. Puebla, Mexico. 1991.

Linear and non-linear mathematical models are used in plant disease epidemiology. Linear models are identified with the general equation $Y=b_0 + b_1 X$ and non-linear models are identified with exponential terms as $Y=b_o+b_1X_1+b_2X_2+b_3X_3+...+b_nX_n$, which is represented by $Y=ae^{bx}$. Most of the epidemiological models are exponential models, which are based in the general equation $d_y/d_t=(Y_t-(Y_{t-1}))/(t-(t-1))$, where $d_y/d_t$ is an increasing relative rate of plant disease (Y) in time (t). In fact, plant disease modeling is based in the following assuptions: relative rate of plant disease ($d_y/d_t$) depends on current plant disease quantity (Y); $d_y/d_t$ depends of the available health plant tissue (1-Y) and; $d_y/d_t$ depends on plant disease quantity (Y) and available plant tissue to be infected (1-Y), where maximum plant disease quantity ($Y_{max}$) ranks from 1 to 100 % (2).

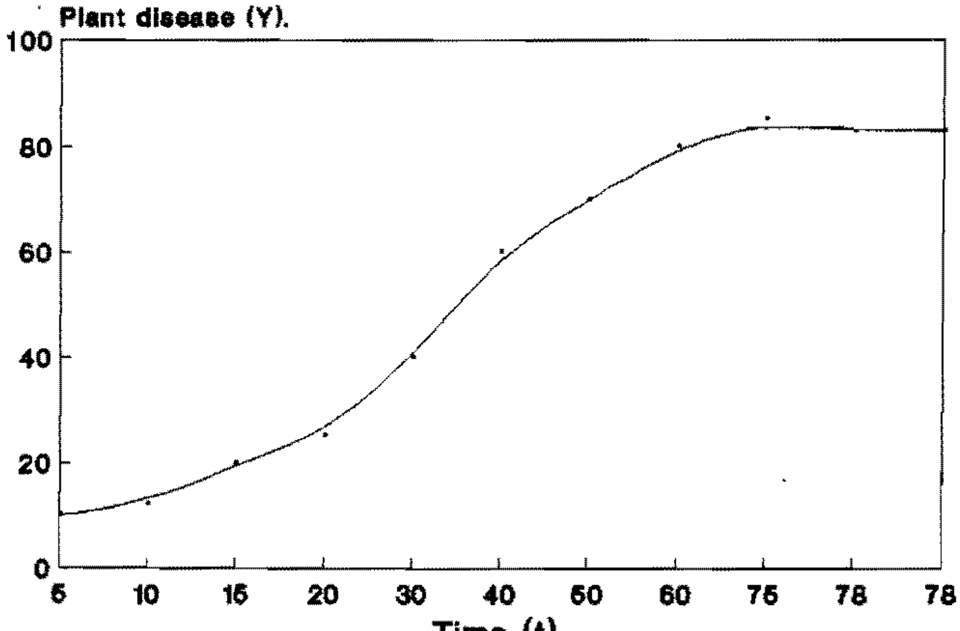


Fig. 5. Common curve in a plant disease epidemic.

**Epidemiological Models.**

Temporal dynamic of plant diseases are described using several models, but the most commonly used are:

a). Exponential model. General equation: $Y=Y_oe^{r_et}$. This model assumes that increasing plant disease has no limit (Fig. 6).

b). Monomolecular model. General equation: $Y=1-be^{-r_mt}$. In this model, relative rate of plant disease increase ($r_m$) depends on the available health plant tissue to be infected ((1-Y) (Fig. 7).

c). Logistic model. General equation: $Y=1/(1+be^{r_lt})$. This model is, in reality, the exponential model, but the plant disease quantity has a limit (Fig. 8).

d). Gompertz model. General equation: $Y=(e^{-bc})^{-r_g}$. In this model, relative rate of plant disease increase is limited by plant disease quantity (Fig. 9).

Some criteria are used to identify the model fitting:

1. Coefficient of determination ($R^2$) in regression analysis: Y=t;

2. Comparison of residual vs predicted values and;

3. Error Mean Squares (EMS).

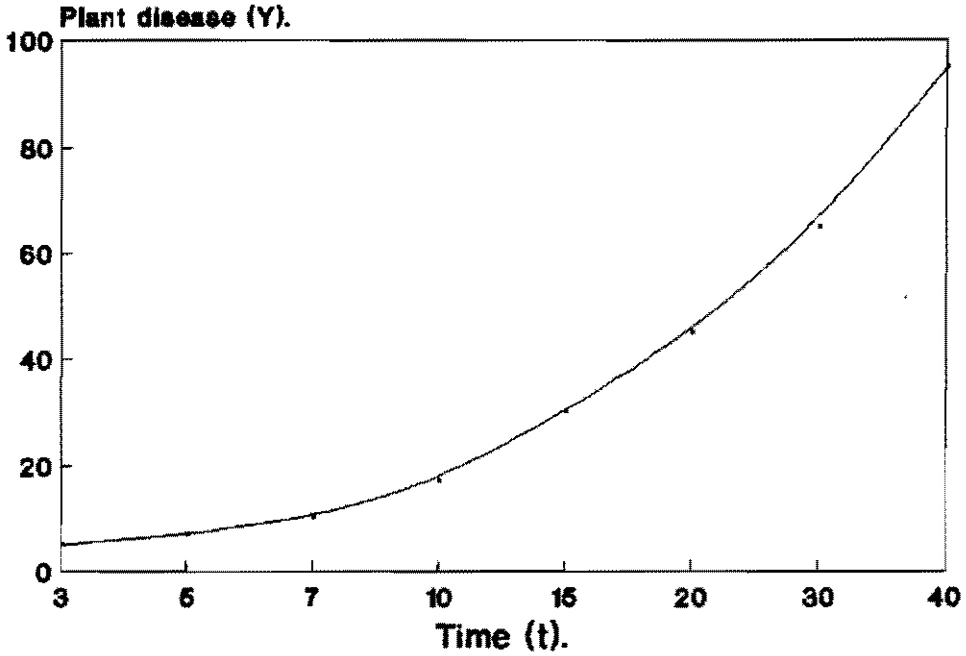


Fig.6. Plant disease curve fitted to Exponential Model.

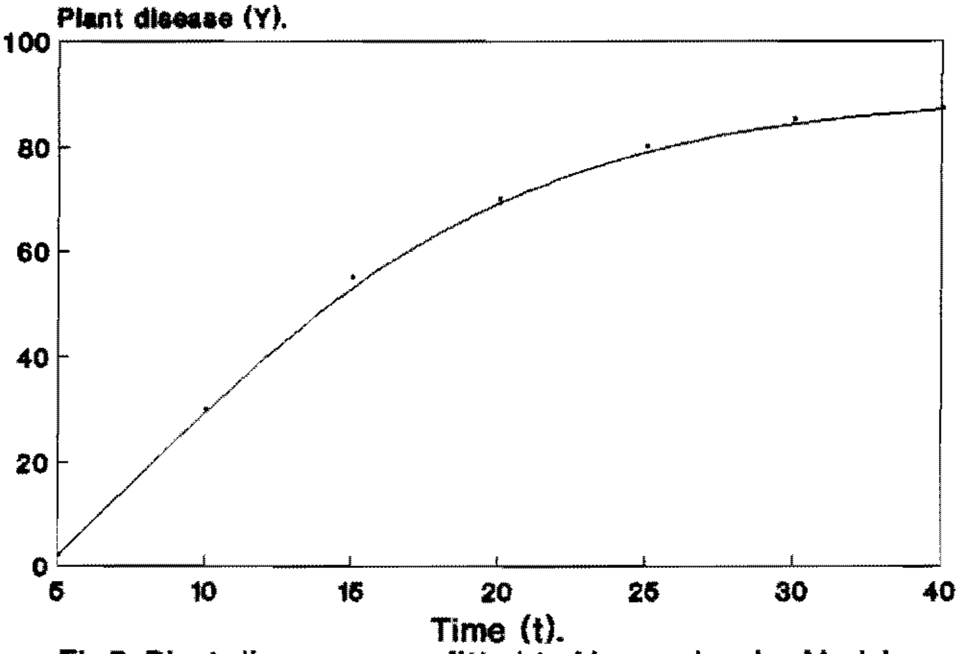


Fig.7. Plant disease curve fitted to Monomolecular Model.

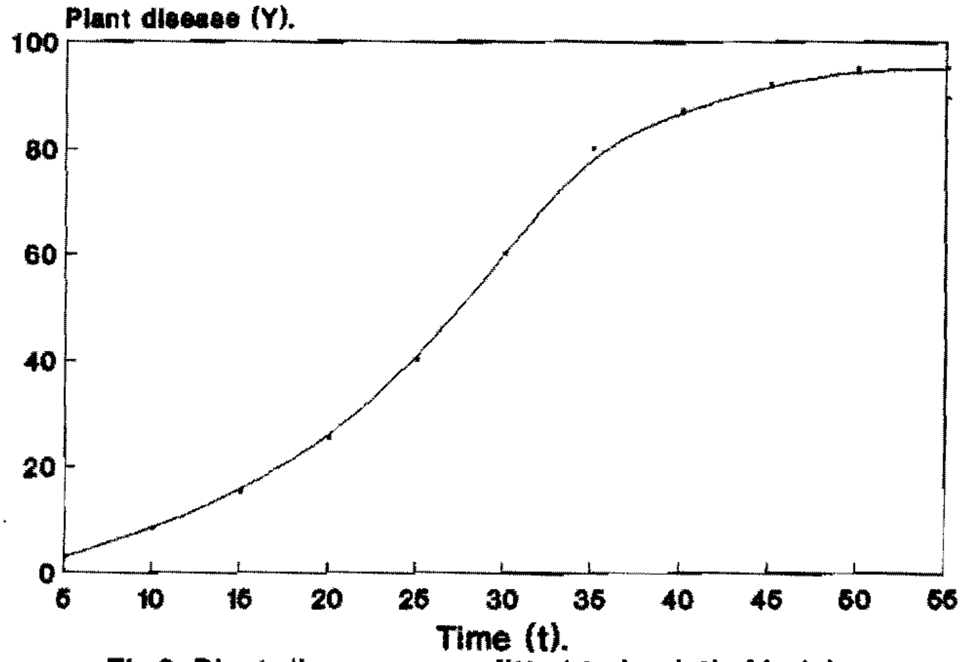


Fig.8. Plant disease curve fitted to Logistic Model.

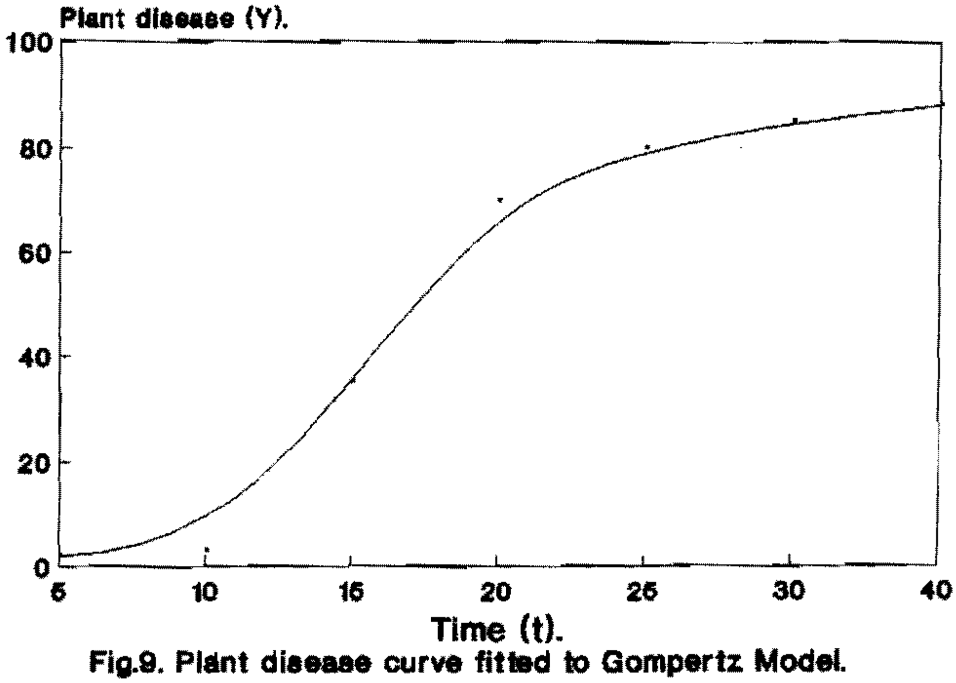


Fig.9. Plant disease curve fitted to Gompertz Model.

## V. STATISTICAL ANALYSIS.

Several statistical options are available to anlyze the epidemiological variables, and sometimes it is not easy to choose which statistic method is better.

a). **Plant disease intensity analysis using mean of plant disease intensity values**. This method is common when the epidemic is changing over time, and it is not possible to fit it to some epidemiological model. Square root arcsine or other equivalent transformation, is suggested when the plant disease is in percentage and in scale values, in order to decrease the variance between treatments (Table 1). This analysis is not recommended, since it keeps a high coefficent of variation, and the changes of plant disease over time is not considered.

**Table 1. Foliar disease intensity (incidence and severity) on three bean varieties. Puebla, Mexico.**

| VARIETY | ANTHRACNOSE | | RUST | | BCMV | |
|---|---|---|---|---|---|---|
| | INC | SEV. | INC. | SEV. | INC. | SEV. |
| AMARILLO 153 | 1.10a<br>(67) | 0.32a<br>(13) | 1.97a<br>(65) | 0.29a<br>(11) | 0.86a<br>(57) | 0.33b<br>(14) |
| BAYO ZARAGOZA | 1.11a<br>(71) | 0.32a<br>(13) | 0.46b<br>(26) | 0.15b<br>(4) | 1.10a<br>(72) | 0.40a<br>(18) |
| MANTEQUILLA DE CALPAN | 1.11a<br>(72) | 0.33a<br>(13) | 0.88a<br>(57) | 0.27a<br>(10) | 0.74b<br>(48) | 0.28b<br>(11) |

(Pedroza, 1991)

Values within a column followed by the same letter do not differ according Tukey test (P=0.05).

Values without parenthesis are arcsine transformated values.

Values inside of parenthesis are percentage values.

b). **Plant disease increasing relative rate analysis**. Plant disease increasing relative rate (PDIRR) is an important epidemiological variable, since it considers the influence of the environment in the host-pathogen relationships. PDIRR is the regression coefficient. In other words, PDIRR is the consequence of the host-pathogen-environment interaction in plant pathosystem. Fitted models in the group of epidemics to be analyzed is required for this kind of analysis, and if the fitted model is different among epedimics, a standarization of PDIRR to the same fitted model is required. Richard's method is used to homogenize the PDIRR to the more frequently fitted model, which is identified by $R^{2,}$ and Error Mean Square in the regression. Richard's method is based in the general equation $Y=(1\pm be^{-rt})^{1/(1-m)}$, where (+) is when $m > 1$, and (-) when $m < 1$. m is a form parameter to allow flexibility in the model, then the equivalence among PDIRR derived of different fitted models is possible by using a common pondered change parameter (Rho), where Rho=r/(2m+2). In this equation, r is the PDIRR of the best fitted model and m=0, m=1, and m=2, whether the best fitted model is the Monomolecular, Gompertz, and Logistic model, respectively. Using the Rho value, it is possible to calculate the new r value, by using the equation Rt=((mx2)+2)(Rho). After homogenizing the PDIRR in a group of plant diseases, we can used this variable in different statistical analysis related to Comparative Epidemiology, such as Analysis of variance and the Tukey Test (Table 2), and this comparison is useful, since the PDIRR considers the change of the epidemic over time.

**Table 2. Increasing relative rate of some plant disease severity on different bean varieties. Barmejillo, Dgo.**

| VARIETY | ROOT ROT | BMCV | BACTERIAL BLIGHT |
|---|---|---|---|
| PINTO AMERICANO | 0.004 a | 0.018b | 0.014b |
| PINTO LAGUNA | 0.008a | 0.025a | 0.020a |
| PINTO ZACATECAS | 0.006a | 0.024a | 0.022a |

(Pedroza, 1994)

Values within a column followed by the same letter do not differ according Tukey test (P=0.05).

c). **Area Under Disease Progress Curve (AUDPC) Analysis**. When a group of epidemics to be analyzed is changing over time and it is not possible to get a fitted model, the AUDPC analysis is a good option. This analysis allows to decrease the variance between treatments, mainly when the plant disease is in percentage, and in scale values. AUDPC analysis reduce the coefficient of variation and consequently increases the reliability of results. General equation $AUDPC=E((Y_i+(Y_{i+1})/2)(t_{i+1})+t_i))$ is used to calculate AUDPC values, where Yi is the plant disease intensity, and t is the interval time (days, weeks, years, etc) among samplings to evaluate plant disease. Units will be in %-days, whether plant disease was rated in percentage values over days. The AUDPC equation is an integration of whole rectangles, using the mean values of the plant disease intensity raised between two dates, during the evaluation (Fig. 10). AUDPC method is an important statistical anlaysis, since it takes advantage of the PDIRR analysis, but in addition, the AUDPC analysis is possible to do it whenever the epidemics were not fitted to some epidemiological model (8) (Table 3).

**Table 3. Area under disease progress curve for incidence and severity of foliar diseases on three bean varieties and two nitrogen sources. Puebla, Mexico.**

| Factor of Variation | DISEASES (%-days) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variety | Bact. blight | | Rust | | Anthracnose | | Mosaic | |
| | Inc. | Sev. | Inc. | Sev. | Inc. | Sev. | Inc. | Sev. |
| Amarillo 153 | 5450a | 901a | 2737a | 450a | 2954a | 511a | 2406b | 587b |
| Bayo Zaragoza | 5606a | 930a | 1089c | 147b | 3002a | 524a | 3023a | 773a |
| Mantequilla de Calpan | 5503a | 954a | 2413b | 417a | 3037a | 540a | 2033c | 464c |
| Nitrogen source (N-P-K)(kg/ha) | | | | | | | | |
| Organic[1] (120-30-00) | 5514a | 922a | 2272a | 385a | 3052a | 515a | 2429a | 601a |
| Inorganic[2] (60-30-00) | 5447a | 921a | 1978a | 315a | 3020a | 540a | 2571a | 639a |
| Control (00-30-00) | 5598a | 942a | 1989a | 315a | 2921a | 520a | 2461a | 584a |

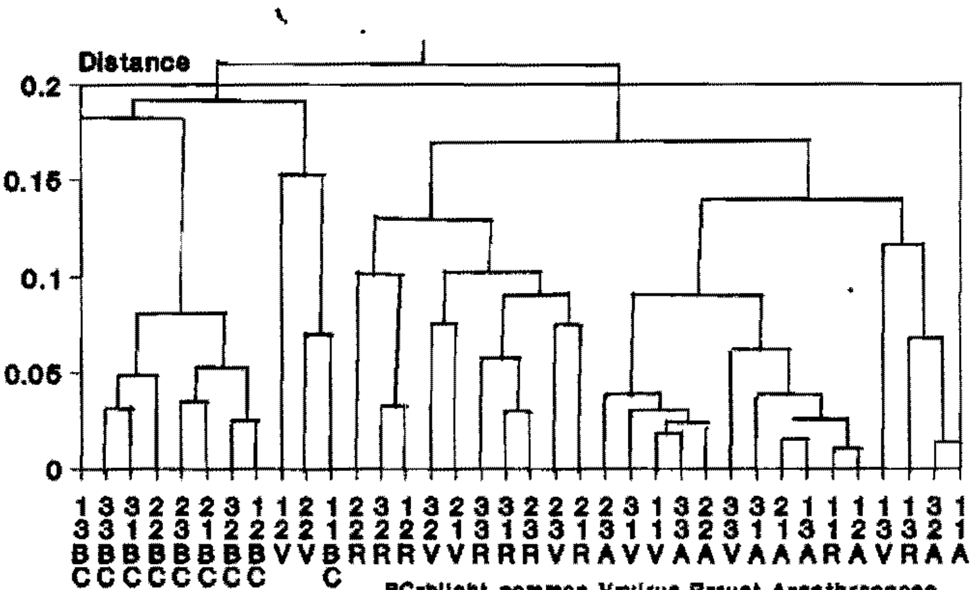
(Pedroza, 1993)

Values within a column in each factor of variation followed by the same letter do not differ according Tukey test (P=0.05).

[1] Bovine manure as a nitrogen source, only about 50% of nitrogen is available at the first year.

[2] Ammonium sulphate as nitrogen source.

d). **Multivariate analysis.** Multivariate Analysis is an important statistic tool, since it allows us to analyze the epidemiological process in an hollistic approach. Last statistical methods cited beforing, statistical analysis is an univariable analysis, which is done by using each separated plant disease and, taking one variable only; however, in the field plant diseases ocurre simultanously over time and space, as an epidemiological complex, keeping an interesting relationships, which sometimes it is not easy to understand them. In fact, Multivariate Analysis as Principal Components, Factor Analysis (Table 4) and Cluster Analysis (Fig.10), allows us to analyze the plant diseases as Epidemiological Complex, and also, it allows remove some redundant variables getting epidemiological indexes. This hollistic approach in plant disease management to control plant diseases, is required to keep a better ecological balance in plant pathosystems.



BC=blight common V=virus R=rust A=anthracnose
1st digit Nitrogen source (1=chemical 2=organic 3=test)
2nd digit= beans cultivar (1=A-153 2=B-Za 3=M-C).

Fig.10. Groups of epidemics in different treatments on bean varieties (Pedroza, 1993).

**Table 4. Factor analysis of some epidemiological parameters on a plant disease complex on beans. Puebla, Mexico.**

| PARAMETER | NUMBER OF FACTOR | | | | | COMUNALITY |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | FACTOR $hi^2$ |
| AUDPCS | 0.85* | 0.34 | 0.31 | 0.13 | 0.16 | 0.982 |
| AUDPCI | 0.83* | 0.41 | 0.33 | 0.11 | -0.07 | 0.981 |
| MEAN OF SEV. | 0.83* | 0.35 | 0.35 | 0.15 | 0.18 | 0.988 |
| MEAN OF INC. | 0.80* | 0.45* | 0.36 | 0.10 | -0.08 | 0.988 |
| INITIAL SEV. | 0.33 | 0.91* | 0.09 | -0.06 | 0.14 | 0.970 |
| INITIAL INC. | 0.37 | 0.91* | 0.04 | 0.05 | -0.07 | 0.971 |
| FINAL INC. | 0.37 | 0.03 | 0.92 | 0.06 | -0.05 | 0.987 |
| FINAL SEV. | 0.45* | 0.18 | 0.69* | 0.33 | 0.40 | 0.991 |
| SIRR | 0.13 | -0.02 | 0.12 | 0.98* | 0.03 | 0.997 |
| EXPLAINED VAR. | 3.34 | 2.30 | 1.79 | 1.14 | 0.25 | 8.83[3] |
| PERCENTAGE | 37.1 | 25.5 | 19.8 | 12.6 | 2.7 | 98.1 |

(Pedroza, 1993)

Values identified with * withing a column means intercorrelated variables.

AUDPCS = Area Under Disease Progress Curve of Severity.

ACBI = Area Under Disease Progress Curve of Incidence.

SIRR = Severity Increasing Relative Rate.

[3] $\Sigma hi^2 = 8.83$

## VI. LITERATURE CITED.

1.Berger, R.D. 1989. Description and application of some general models for plant disease epidemics. Plant Disease Epidemiology. Vol. 2. Genetics, Resistance, adn Management (K.I. Leonard and W.E. Fry, Eds). McGrow-Hill, New York. p:125-149.

2.Campbell, C.L., and L.V. Madden, 1990. Introduction to Plant disease Epidemiology. John Wiley & Sons, New York. 532 pp.

3.Kranz, J. 1974. Comparison fo Epidemics. Ann. Rev. Phatopathology 12:355-374.

4.Mora-Aguilera, G. Kiltie, R., Nieto-Angel, D. and Téliz, D. 1993. Outliers and redundant variables in principal component and factor analysis. South East SAS Users Group. First Annual Conference, Florida. p. 185-194.

5.Pedroza Sandoval, A. 1994. Disease Index, Redundant Variables and Grouping of Diseases in Cluster and Factor Analysis. III Network Mecting of the Biometric Society. Caracas, Venezuela p. 106.

6.Pedroza Sandoval, A. 1993. Efecto de variedades, densidades de plantas, dósis y fuentes de nitrógeno en la incidencia y severidad de las enfermedades del frijol (Phaseolus vulgaris L.). Tésis de Doctor en Ciencias. Colegio de Postgraduados, Montecillo, Méx. 166 pp.

7.Pedroza Sandoval, A. 1995. El Déficit Hídrico en Patosistemas Agrícolas. Rev. Mexicana de Fitopatologia (En Prensa).

8.Pedroza Sandoval, A. Téliz, O.D., De La Torre, A.R., and Campbell, C.L. 1995. Cultural practices as a tool in plant disease management on beans (Phaseolus vulgaris L.) Rev. Mex. de Fitopatología (A ser enviado).

9.Pedroza Sandoval, A. y Téliz Ortíz, D. 1993. Dinámica temporal de un complejo epidémico foliar en el cultivo del frijol (Phaseolus vulgaris L.) usando diferentes prácaticas de manejo. Rev. Mex. de Fitopatología. 11:118-123.

10.Pedroza Sandoval, A. y Téliz, O.D. 1992. Relaciones microclimáticas en patosistemas agrícolas. Revista Mexicana de Fitopatología. 10:11-14.

11.Robinson, R. 1976. Plant Pathosystems. Springer Verlag, Berlin, 184 pp.

12.Schumann, L.G. 1991. Plant Diseases: Their Biology and Social Impact. APS Press. The American Phytopathological Society. St. Paul Minnesota, USA 397 pp.

13.Van der Plank, J.E. 1963. Plant Diseases: Epidemics and Control. Academic Press, New York. 349 pp.