CONTRIBUTIONS OF THE DATA SERVICES UNIT TO CIAT RESEARCH (1989 - 1991)

CIAT Internal Program Review 1991

December 2-6, 1991

CONTENT

DATA SERVICES UNIT
PERSONNEL 1991

a) Personnel budgeted to the Data Services Unit:

	NAME	POSITION	DEGREE
-	Unit Head (Acting)	Maria Cristina Amézquita	Ms. and Dipl. in Mathematical Statistics
-	Secretaries	Maria Eugenia Echeverri	
		Marta Elena Carvajal	
-	<u>BIOMETRY</u>		
	. Statistical Consultants:		
		Eduardo Granados	Ms. Mathematical Statistics
		James A. García	Ms. Industrial and Systems Eng.
		Pedro Pablo Perdomo	Ms. Mathematical Statistics
		Myriam Cristina Duque	Bs. Mathematics
		Germán Lema	Bs. Industrial Eng.
	. Statistical Programmer:		
		Rosalba López	Programming Technology
		Carlos Saa (½ time)	Industrial Eng. (Student)
-	<u>DATABASES</u>		
	. Analysts	Germán Serrano	Bs. Systems Eng.
		Fernando Rojas	Bs. Systems Eng.
		NN	-
	. Programmers	Norbey Marín	Systems Technology
		Carlos Saa (½ time)	Industrial Eng. (Student)
-	<u>IBM 4361 OPERATION</u>		
	. System's Programmer		
		Hugo Macias	Ms. Systems Eng.
	. Operators/Transcriptors		
		Jairo Ramirez	
		Carlos López	
		Elizabeth González	
		Amparo Rivadeneira	
		Eva Neuta	
		Fernando Arango	

b) Personnel budgeted to research Programs but technically responsible to the Data Services Unit:

-	<u>TROPICAL PASTURES</u>		
		Manuel Arturo Franco	Ms. Systems Eng.
		Eloina Mesa	Ms. Statistics
		Gerardo Ramirez	Bs. Statistics and Mathematics
		Carlos A. Hernandez	Systems Technology
-	<u>BEANS</u>	Javier Crespo	Systems Technology
-	<u>RICE</u>	Hector Fabio Ramirez	Bs. Statistics

CONTRIBUTIONS OF THE DATA SERVICES UNIT TO CIAT RESEARCH (1989-1991)

1. DATA SERVICES UNIT OVERVIEW: ROLE, FUNCTIONS AND RESOURCES

The Data Services Unit (DSU) is a research support unit responsible for providing advise and support to CIAT research Programs/Units in two main areas:

a) Biometry.

b) Research databases development.

Additionally, the DSU is responsible for the provision and maintenance of appropriate computer hardware and software to serve the scientific program's needs through the CIAT mainframe computer, IBM 4361, with its network of 70 terminals and connected microcomputers.

These three functions are fulfilled by three distinct groups in the Unit: the Biometry group composed by 5 statistical consultants and 1.5 statistical programmer; the Databases group, composed by 3 system's analysts and 1.5 programmers; and the IBM 4361 Operation group, composed by 1

Biometry support: Role definition:

Biometry collaboration to CIAT research Programs/Units includes the following functions:

a) To provide statistical/mathematical advice to researchers in project design, data analysis methodology, interpretation of results, their generalization capacity, and final presentation. Also, when required, Biometry provides advise to researchers in the selection of statistical summaries to store in databases needed by the scientist for their research process.

b) To get involved in collaborative methodological studies and in specific research data analysis projects with researchers.

c) To provide training in statistical/mathematical methods and data analysis to colleagues in other disciplines (both internal and external).

An important activity of the biometrician is his/her involvement in **collaborative data analysis projects** with researchers, aimed at responding relevant questions of research. These projects integrate data generated by a given research project through the years, by a given research discipline, or combine data generated by various disciplines within a Program. The results of some of these projects have already appeared as chapters of CIAT Programs publications, some as contributions to International Networks reports, some others have been published as joint papers with the scientists, and some others are in progress.

Another important activity that is expected to increase in the near future, is the biometrician involvement in collaborative **methodological studies** with the research Programs/Units. The biometrician contribution in this context is to offer orientation on: what designs are most appropriate for a given research; what sources of variation are relevant for research planning on a specific subject; suggestions on modifications on a given experimental design to provide better extrapolation capacity to a specific experiment, taking in consideration environmental conditions of the region over which the results are expected to be generalized; to evaluate what techniques have been the most effective research tools; the efficiency, accuracy and applicability of specific statistical methods to analyze a given research problem; evaluate the strategy of using multivariate methods for a large number of variables rather than ordinary univariate analysis, for example.

The **training** in statistical methods and data analysis is provided to CIAT research associates/assistants and to National Institutions researchers. The Microcomputer Training Laboratory is used for this purpose. During the four years of existence of the Laboratory, Biometry has offered a total of 35 one to two-weeks training courses, with a total number of 280 National Institutions researchers trained from Latinamerica (220), Asia (24) and Africa (36). An approximate number of 105 participations from CIAT research associates/assistants have benefit from this effort.

In the light of the new CIAT, new areas of biometrical expertise in which invited Biometrician Consultants can add useful contributions are foreseen. For example: a) Design and analysis of intercropping experimentation, combining multiple short-cycle crops or combining perennial and short-cycle crops. b) Design and analysis of agro-silvo-grazing systems. c) Use of operations research techniques (linear and quadratic programming, transportation problem, critical path method and simulation) in response to a new expected demand from the pool of economists/social scientists.

Statistical/data analysis software for the mainframe computer include: SAS/BASICS, SAS/STATS, SAS/GRAPH, SAS/FSP, SAS/ETS, SAS/IML and SAS/OR from SAS Institute Inc. Raleigh, North Carolina, USA; GENSTAT, GLIM and Fortran Library, from the NAG Algorithm Group, London, England. Microcomputer statistical/data analysis software include MSTAT, from Michigan State University; GLMM, from Louisiana State University; SYSTAT, from SYSTAT Inc. Chicago, Illinois, Lotus 1-2-3 and Dbase III.

Databases development:

The conceptualization, design and implementation of databases to store crops and pasture research results, require from the "database team" a clear understanding of the biological nature of the crop and its multiple components. Members of our "database team" include: a) the System's Analyst, who is the software expert, the designer of the how to efficiently store the datafiles with minimum redundance, and how to provide interactive access to the data in the most effective manner; b) the Researcher(s), who have a clear understanding of the problem and the purpose of the database; and c) the Biometrician, who has a clear understanding of the data and of the most suitable statistical summaries with which to characterize a given research process.

In terms of database management software technology an important decision was reached in September of this year: that of moving from the 'network' database technology, represented by our previous database management software IDMS/R from Computer Associates Inc., to the 'relational' database technology, represented by ORACLE, from Oracle Corporation. Starting October 1, 1991, ORACLE was acquired as the database management software for CIAT's mainframe and micro environment, as a replacement of IDMS/R. The Data Services Unit feels very proud of having successfully culminated this software evaluation process initiated in July 1990, devoted to solve the existing problems in the design, implementation and utilization of research databases developed on IDMS/R during the past 10 years. Certain characteristics of the IDMS/R software, such as lack of flexibility for modifying a database design, extremely long data loading times, lack of a user-friendly query tool, lack of a flexible and powerful development tool and lack of micro-mainframe interface have been greatly responsible for database implementation problems and for the very limited use of the existing database applications by the CIAT scientific community. Between September and December of this year, two important ORACLE database applications were designed, implemented and released to the end-user: a) The Genetic resources databases, including passport and characterization data on all CIAT's germplasm collections: beans (40,000 accessions approx.), cassava (4,700 varieties approx.), tropical forrages-legumes and grasses (25,000 accessions approx.); and b) The Cassava breeding database containing all descriptive information on material from the germplasm bank, parents and crosses and statistical summaries of preliminary yield trials, advanced yield trials and regional trials. This database contains at present research results between 1974 and 1991.

The re-design and implementation in ORACLE of existing IDMS/R databases, as is the case of the Tropical Pastures Program database, will be carried-out in 1992 in a very close collaboration with the Program's Leader and Scientists.

2. CONTRIBUTIONS TO TROPICAL PASTURES

2.1. BIOMETRY: METHODOLOGICAL CONTRIBUTIONS TO TROPICAL PASTURES RESEARCH

The special nature of tropical pastures research needs to be recognized. When compared with classic agricultural research carried-out with short-cycle crops whose final product is obtained at harvest in small plots, under standard experimental designs, tropical pastures research presents many additional complexities. Tropical pastures research is of long-term nature as it deals with a perennial crop. Since the final products of a pasture are milk, meat, wool or other animal products, the pasture researcher has to recognize that small-plot clipping trials and medium and large-scale grazing experiments are complementary.

Under the evaluation scheme used by the CIAT's Tropical Pastures Program, a grass or legume accession is first submitted to small-plot agronomic trials to evaluate its adaptation to soil, climate and biotic conditions and its biomass production potential; then, selected grass-legume associations are submitted to agronomic evaluations under grazing to study their compatibility and persistence under the animal influence; advanced materials are then submitted to large-scale grazing trials to measure animal productivity, the latter expressed in terms of weight gain of young steers, milk production capacity of a dairy herd, reproductive performance of breeding herds, or mixed beef and milk production under double-purpose production systems at farm level.

In small-plot agronomic experiments, standard experimental designs are utilized, in which the effect of one or more experimental factors at various levels can be studied under replicated factorials for example. However, response variables need to be analyzed as repeated measurements within season, and the statistical analysis may involve a response curve fitting by season and multivariate comparisons of regression parameters among treatments; or they could be expressed as summary indicators over the experimental period.

In pasture evaluation experiments under grazing, experimental designs tend to be simple, but additional sources of variation on the pasture response need to be considered for data analysis purposes. Besides 'soil', 'year', 'season within year' and 'pasture quality across time', 'animal variability' (sex, age, origin, condition) is of great importance.

Milk production trials, given the high cost of experimental animals, represent a very interesting research area for the utilization of change-over designs, which require less experimental units to attain similar levels of significance, when compared to standard continuous designs.

On the other extreme, large-scale pasture evaluation experiments with commercial breeding herds, conducted to measure reproductive efficiency in beef cattle, although of simple design (as RCB or CR) and oriented towards a direct adoption by producers, require the use of sophisticated and often complex statistical methods for efficient data manipulation and analysis.
These facts make tropical pastures research an extremely challenging field of work for biometricians.

Support to RIEPT in the management and statistical analysis of its information: Since its creation in 1979, the RIEPT* assigned the CIAT Tropical Pastures Program the responsibility to centralize and make available to network members all the information generated by the network. Since then, Biometry has collaborated very closely with the Tropical Pastures Program in the organization, storage and statistical analysis (by site, by country, by ecosystem or across-ecosystem

RIEPT = International Network for Tropical Pastures Evaluation data analysis) of RIEPT-generated research results. Up to now, 241 agronomic-trials (ERA and ERB) and some 10 grazing trials have been statistically analyzed and their results stored in the RIEPT database.

Examples 1 and 2 shown in the next pages, represent data analysis studies utilizing RIEPT information to answer important research questions. Additionally, some selected examples of methodological studies carried-out between the Biometry Section of the Data Services Unit and the Tropical Pastures Program will be summarized in this report.

Example 1:

RANGE OF ADAPTATION OF *Stylosanthes guianensis*, cv. Pucallpa IN THE AMERICAN TROPICAL RAIN FOREST ECOSYSTEM.

M.C. Amézquita, J.M. Toledo, and G. Keller-Grein
(Published by Tropical Grasslands, Sept. 1991).

The purpose of this study was to define the range of adaptation of *Stylosanthes guianensis* CIAT 184, released in 1985 as cv. Pucallpa by IVITA (Instituto Veterinario de Investigaciones Tropicales y de Altura) and INIPA (Instituto Nacional de Investigación y Promoción Agropecuaria) in Perú. Data from 32 RIEPT (International Network for Tropical Pastures Evaluation) type B trials conducted in the American Humid tropics between Mexico and Bolivia were used for this study (table 1, fig.1). Statistical methodology covered four stages:

a) The definition of agronomic indicators of rapidity in establishment and biomass productivity.
b) The identification of environmental parameters that would affect establishment and production of the cultivars. Stepwise regressions were carried-out with the agronomic indicator as the dependent variable in each regression, and a reduced-set of non-correlated environmental parameters (soil, climate and location) as independent variables.
Those environmental parameters found significant in the regression were considered important sources of variability on the agronomic performance of cv. Pucallpa.
c) The identification of groups of sites with similar environmental conditions for the cultivar performance. A hierarchical Cluster Analysis technique with Ward's minimum variance method was used for this purpose.
d) The description of agronomic performance of Cultivar Pucallpa in each group of environments.

This study shows that cv. Pucallpa is tolerant to anthracnose under a wide range of soil, climate and locations; the cultivar is better adapted to low altitudes (<850 m.a.s.l.), on soils that are acid (pH 5.0), which have low levels of organic matter (<3.4%), are moderately sandy (18-56% sand), and which have rainfall accumulated in 12 weeks > 800mm; at higher altitudes (>1000 m.a.s.l.), the cultivar appears to respond to higher levels of organic matter (tables 2,3 and 4).

Example 2:

AGRONOMIC PERFORMANCE OF THREE CULTIVARS RELEASED IN COLOMBIA

"Pasto Llanero" *(B. dictyoneura* 6133)
"Pasto Carimagua" (*A. gayanus* 621)
"Cultivar Capica" (*S. capitata* 10280)

E. Mesa, M.C. Amézquita, J.M. Toledo (1990)

The purpose of this study was to provide a quantitative description of the agronomic performance - in Colombia - of these forage cultivars recently released by ICA (Instituto Colombiano Agropecuario).

The study identifies contrasting zones in Colombia based on environmental parameters (altitude, precipitation, Index of bases content, Index of soil texture and organic matter content. The two latter ones corresponding to the first two Principal Components on soil parameters). Then describes performance of each cultivar in each zone making statistical comparisons between zones.

Data source: 22 Regional Trials B conducted in Colombia
from 1979-1987.

The results incorporated in the 1991 Tropical Pastures Program report (Dr. G. Keller-Grein) show the wide adaptability of both grasses - B. dictyoneura and A. gayanus and their high potential for productivity under the more humid environments with altitudes below 1500 m.a.s.l. (including coffee region and the Amazon). Their performance in the well-drained Llanos ecosystem is similar, with a drastic reduction in productivity during the dry season. Cultivar Capica, on the contrary, is more exclusively adapted to the well-drained sandier Llanos ecosystem, with some potential for dry season production in the coffee zone and the Amazona (Leticia zone). -

Example 3:

A DATA ANALYSIS METHODOLOGY FOR THE EVALUATION OF LARGE GERMPLASM COLLECTIONS CASE STUDY: EVALUATION OF THE CIAT *Brachiaria* COLLECTION IN BRAZIL

Cacilda do Valle*, M.C. Amézquita and P.P. Perdomo
(work in progress)

The agronomic evaluation of forage germplasm collections in the Tropics involves periodic measurements of plant responses that cover the most contrasting seasonal periods of the region of interest. In order to characterize an accession, summary indicators by season or dry-rainy

* *EMBRAPA researcher. CIAT Visiting Scientists during 1991*

season relations need to be computed. As the resulting number of plant response indicators is normally very large and significant correlations between them may exist, reduction-of-dimensionality techniques need to be applied to reduce them to a minimum number of non-correlated ones. The present study illustrates these aspects. It presents a methodology for data analysis of the agronomic evaluation of a large germplasm collection. 194 accessions of 9 species of *Brachiaria* were evaluated by EMBRAPA, in Campo Grande, Brazil, over 2 years, in small plots, under a split-plot design. Biomass production (total, leaf, stem) and regrowth capacity were periodically measured. Additionally, observations on resistance to insects and diseases, and plant vigor were made periodically . Early flowering capacity was recorded only once during the experimental period.

Methodology: A set of ten highest priority summary indicators were computed as functions of the original measurements. They were:

1. Annual accumulated total dry matter (kg/ha/year) (ATDM).
2. Accumulated total dry matter during the dry season, expressed as percentage of annual total dry matter ((TDM_{dry}/ATDM) x 100).
3. Annual accumulated leaf dry matter (kg/ha/year) (ALDM).
4. Accumulated leaf dry matter during the dry season expressed as percentage of annual leaf dry matter. ((LDM_{dry}/ALDM) x 100).

5,6. Percentage of leaf dry matter from total dry matter
- during the dry season ($PLDM_{dry}$)
- during the rainy season ($PLDM_{rainy}$)

7,8. Leaf-stem relation, based on dry matter
- during the dry season (LDM_{dry}/SDM_{dry} x 100)
- during the rainy season (LDM_{rainy}/SDM_{rainy} x 100)

9,10. Regrowth capacity (ordinal 0-6 scale)
- during the dry season (RC_{dry})
- during the rainy season (RC_{rainy})

A Factor Analysis, with varimax rotation method, was applied to these 10 indicators. Based on the resulting reduced number of factors, a Ward's minimum variance Cluster Analysis was performed to classify accessions with similar agronomic characteristics within species.

Results: As a result, the four first factors -explaining 87.8% of the total variation- were selected as a reduced set of non-correlated groups of indicators. One indicator from each one of the factors, was chosen to represent the factor. These were: a) Leaf-Stem relation, during the dry season (%); b) Annual accumulated leaf dry matter (kg/ha/year); c)Regrowth capacity during the rainy season (0-6 ordinal scale); d) Leaf dry matter during the dry season, expressed as percentage of annual leaf dry matter (%)

The Cluster Analysis helped identify 14 promising accessions, out of which 9 were selected to advance for grazing studies: 6 from *B. brizantha*, superior to the standard cultivar cv. "Marandú"; 1 from *B. decumbens*, superior to cv. "Basilisk"; 1 from *B. humidicola*, and 1 from *B. jubata*. (See Tables 1 and 2).

Example 4:

PASTURE EVALUATION UNDER GRAZING WITH BREEDING HERDS: A METHODOLOGY FOR DATA ANALYSIS

M.C. Amézquita, R. Vera and G. Lema

(Submitted for publication in October, 1991)

Pasture evaluation experiments with breeding herds use simple designs; however, efficiency in data manipulation and analysis requires sophisticated and often complex statistical methods. Results from a large grazing experiment, conducted in Carimagua research station, eastern Colombian savannas, for over 6 years, with 325 Zebu x Criollo cows, were used as data source to test the accuracy and applicability of different statistical methods for the analysis of reproductive performance. The "Herd Systems" experimental design corresponds to a non-replicated factorial with two factors: **production system** (at 3 levels: 1) savanna-based; 2) savanna-based plus 800m^2 per animal unit of an improved grass-legume association; and 3) savanna-based plus 1600m^2 per animal unit of an improved grass-legume association); and **site** (at 2 levels: 1) Yopare, loamy soil and 2) La Alegria, sandy soil (Vera 1982).

Methodology used for data analysis includes: a) an exploratory data analysis, to determine the minimum acceptable experimental period length for valid statistical inferences; b) the use of MANOVA to analyze continuous variables with repeated measures in time; and c) the use of three alternative statistical procedures to analyze categorical variables: Stratified Analysis using the Cochran-Mantel-Haenszel statistic expressed as a function of the 'traditional' chi-square test (CMH); an Stratified Analysis using the CMH_R, expressed as a function of a 'modified' chi-square test (X^2_R) proposed by Brown (1988); and a linear model fit on marginal probabilities.

Results of this study suggest that: a) 4 years is the minimum acceptable experimental period length for this type of experimental projects; b) there is a need to accept the use of non-replicated designs for large-scale grazing experiments with breeding herds, thereby using the between-animal variability as a proxy for experimental error; c) the selection of mixed breeding herds provides more generalization capacity to commercial situations although brings complications in data selection, data manipulation and statistical analysis; d) MANOVA is shown as a solid tool of practical use and easy interpretation for the analysis of continuous variables with repeated measures in time; e) The Stratified Analysis and the linear model fit on marginal probabilities represent a complementary set of tools to make integrated inferences on categorical variables.

This study also shows that the most sensitive indicators of treatment and site differences are: interval between parturitions, calf weaning weight, abortions/cow, total number of births/cow, total number of weaned calves/cow and the three selected summary parameters, ie. total production of weaned calves/cow (kg), total production of calves/cow (kg) and total beef production per cow (kg) during the experimental period.

Figures 1 and 2 illustrate that trends in animal performance parameters show consistency in stabilizing from the 4th year onwards. The resulting sub-sample of 178 experimental cows with complete reproductive records through the 4-year period was shown to maintain the same herd composition in terms of age, physiological stage and initial weight than the original experimental population. Tables 3 and 4 show the complementarity of results from the analysis of the categorical variables when using both methods: Stratified Analysis and a linear model fit on the marginal probabilities. We show, as an example, the results concerning the analysis of "number of births/cow".

2.2 DATABASES DEVELOPMENT FOR THE TROPICAL PASTURES PROGRAM

A. Franco, E. Mesa, M.C. Amézquita and C.A. Hernández, from the DSU
and
Luis H. Franco and Tropical Pastures Program Leader and Scientists, from the TPP

Genetic Resources Database:

The purpose of the genetic resources database is to store in an organized way, maintain and make interactively available, all the information related to tropical pastures genetic resources that have been generated, collected and handled by the CIAT Genetic Resources Unit or by the Germplasm Evaluation Section of the CIAT Tropical Pastures Program between 1978 and 1991.

The tropical pastures genetic resources database includes information on 22,818 accessions of grasses and legumes. The database is organized in seven sub-systems, according to the type of information being stored, as follows:

1. Passport data (22,818 accessions)
2. Morpho-agronomic characterization of germplasm (6,566 accessions, evaluated through 128 experimental projects)
3. Short-term seed inventory (20,220 accessions)
4. Long-term seed inventory (3,872 accessions)
5. Seed international shipments (200 shipments to Latinamerican, Asian and African countries)
6. Seed multiplication, at greenhouse and at field level (historical records of all material submitted to seed multiplication: 4,815 multiplication events)
7. Inventory of materials stored as "herbario" (9,699 accessions)

Germplasm evaluation Database:

The purpose of this database is to store, maintain and make interactively available to tropical pastures scientists, result summaries of the different research projects conducted between 1978-1991 on forage germplasm evaluation by the CIAT Tropical Pastures Program. These projects cover research work carried-out by the following groups of sections: a) Agronomy: small-plot agronomic evaluations (30 projects concerning 1050 accessions) and agronomic evaluation under grazing (40 projects); b) Plant-protection experiments: phytopathology (50 projects); fungal and bacterial collection (descriptive characteristics on 4,500 fungus and 200 bacteria); c) Soil-plant and soil microbiology research projects (30 soil-plant nutrition experiments, 50 microbiology research projects and the Rhizobium strains collection containing information on 4,200 strains); d) Pasture quality evaluations (8 projects on pasture nutritional characteristics and 8 palability trials at Quilichao Research Station); e) Pasture productivity evaluations in terms of animal production parameters (6 long-term projects conducted in Carimagua and Quilichao; f) Production systems: 3 long-term experiments on reproductive efficiency of beef cattle conducted in Carimagua (1972-1988);
1 experiment on early weaning and 1 on methodological research concerning animal categories.

The germplasm evaluation database also contains the information generated by three international networks in which the CIAT Tropical Pastures Program participates as an active member. These are: RIEPT (International Network for Tropical Pastures evaluation, 1979-1991), Centrosema International Network (1989-1991), and WECAFNET (West and Central Africa Forage evaluation Network, which started on October 1990).

The RIEPT database includes environmental and experimental information on 241 trials conducted (and reported to CIAT) by RIEPT between 1979 and 1991 in 18 countries of Tropical America between Mexico and Bolivia including Caribbean countries. The 241 trials are composed by 45 small-plot adaptability trials (ERA) and 196 small-plot agronomic trials (ERB). Information concerning grazing trials (ERC and ERD) is being collected and organized to be stored in the future. *Additionally, the* RIEPT database contains information on prices on beef production inputs/products at all RIEPT sites. The RIEPT database, resident in the CIAT IBM 4361 mainframe computer, is being now distributed to Latinamerican National Programs in diskettes to be consulted via DBASE III.

Additionally, the Germplasm Evaluation Database contains a sub-system called "Research proposals follow-up", that includes summaries of 680 research proposals and their follow-up, carried-out by *Tropical Pastures* Program researchers between 1978 and 1991. This very valuable historical information provides a feed-back to the Tropical Pastures Program Leader on the type of research topics being addressed.

Publications based on Tropical Pastures databases:

- "Catálogo de germoplasma de especies forrajeras", (1987) (3 Volumes)
- "Colección de Centrosema del CIAT", (1986).
- "Catálogo mundial de germoplasma de Centrosema", (1989).
- "Catalogue of Rhizobium strains for tropical forage legumes", (1985, 1986, 1987 and 1988)
- "RIEPT - Resultados 1979-1982", II Reunión.
- "RIEPT - Resultados 1982-1985", III Reunión.
- "RIEPT - Resultados RIEPT-Amazonía", (1990)
- "RIEPT - Análisis sobre localidades y evaluaciones de germoplasma en el Trópico Húmedo", (1990).
- "RIEPT - Análisis sobre localidades y evaluaciones de germoplasma en Centroamerica y Caribe" (in press.)
- "RIEPT - Recursos disponibles, demanda de servicios y logros en la RIEPT". Contribución de las pasturas mejoradas a la producción animal en el trópico", (1989).
- "RIEPT - Evaluación del comportamiento de ecotipos dentro y a través de ecosistemas", (1985).
- "RIEPT - Análisis de precios de productos e insumos ganaderos en localidades de la RIEPT", (1984, 1985, 1986, 1987, 1988, 1990).
- "Trends in CIAT commodities", (1982,.., 1991).
- "RIEPT - Base de datos estadística. Información y opciones para su utilización", (1987).
- "La colección de forrajeras tropicales del CIAT", (1991) (3 volumes).
    I. Catálogo de germoplasma de Asia Suroriental
    II. Catálogo de germoplasma de Venezuela
    III. Catálogo de germoplasma de Centroamerica, Mexico y el Caribe
- "Utilización de información de ensayos multilocacionales de evaluación de germoplasma. Organización de bases de datos", (1988).

Table 1: *Brachiaria* **species evaluated in Campo Grande, Brazil.**

OVERALL DESCRIPTIVE STATISTICS

Specie	No. of accesions	Accumulated Leaf Dry Matter (kg/ha/year)	Leaf Dry Matter during the dry season as % of annual Dry Matter (%)	Leaf-Stem relation during the dry season	Regrowth capacity in the rainy season (0-6 ordinal scale)
B. brizantha	97	9,161	27	1.73	3.13
B. decumbens	35	4,459	18	0.96	2.14
B. humidicola	21	5,794	16	1.37	2.81
B. jubata	11	4,292	25	1.42	2.82
B. ruzisiensis	20	3,744	14	1.83	2.1
B. arrecta	6	2,096	8	0.58	1.8
B. dyctioneura	2	9,391	8	-	4.0
B. negropedata	1	4,004	30	-	3.0
B. adspersa	1	2,743	14	0.7	3.0
TOTAL	194	6,837.9	21.9	1.5	2.76

Table 2: Evaluation of the CIAT Brachiaria collection in Campo Grande, Brazil.

14 SELECTED ACCESSIONS

ACCESSION IDENTIFICATION		Accumulated Leaf Dry Matter	Leaf Dry matter during the dry season, as % of annual dry Matter	Leaf-Steam relation during the dry season	Regrowth capacity in the rainy season
CIAT #	EMBRAPA #	(kg/ha/year)	(%)		(0-6 ordinal scale)
B. brizantha					
- 16288*	B132	20954	35%	1.55	3
- 16467*	B166	18380	35%	1.29	3
- 16316*	B144	17134	27%	1.58	3
- 16315*	B72	16905	23%	1.49	5
- 16306*	B138	15176	27%	1.15	4
- 16473*	B89	14415	30%	1.07	4
B. decumbens					
- 16488*	D1	11344	26%	1.08	4
- 606	D62	10171	31%	1.51	3
- 6699	D70	8814	36%	1.73	3
B. humidicola					
- 26155*	H18	9334	20%	2.19	3
- 16886	H13	8593	13%	1.66	4
- 116350	H19	7126	32%	1.79	3
B. jubata					
- 26237*	J13	8556	16%	1.20	3
- 16195	J1	4852	39%	2.47	3

1/ Out of these 14 accessions, the 9 accessions with an * were identified to advance for grazing studies.

Table 3.: Stratified Analysis results

Response variable: Number of births/cow (2,3 or 4)[1]

Site 1

Treat	2	3	4	N	R_t	R_a
1	13 (43.3)	13 (43.3)	4 (13.4)	30	2.7	0.68
2	3 (9.1)	19 (57.6)	11 (33.3)	33	3.2	0.81
3	4 (11.4)	23 (65.7)	8 (22.9)	35	3.1	0.78
	20 (20.4)	55 (56.1)	23 (23.5)	98	3.01	0.75

X^2 = 15.2 (prob = 0.004)
X^2_R = 0.72 (prob = 0.71)

Site 2

Treat	2	3	4	N	R_t	R_a
1	9 (37.5)	13 (54.2)	2 (8.3)	24	2.7	0.68
2	13 (46.5)	14 (50.0)	1 (3.5)	28	2.6	0.65
3	8 (33.3)	11 (45.8)	5 (20.9)	24	2.9	0.73
	30 (39.5)	38 (50.0)	8 (10.5)	76	2.73	0.68

X^2 = 7.26 (prob = 0.12)
X^2_R = 0.19 (prob = 0.91)

CMH = 6.47 (prob = 0.015)
CMH_R = 0.91 (prob = 0.86)

[1] 4 cows were deleted from the analysis: one with 5 births (in treat 2, site 1), and three with 1 birth (2 in treat 1 site 1, and 1 in treat 1 site 2)

[2] R_t = 4-year period birth rate (expressed as births/cow in 4 years)

[3] R_a = Mean annual birth rate (expressed as mean births/cow/year)

Table 4.: Linear model fit using marginal probabilities, for the analysis of "number of births/cow".

a) - Response frequencies and response probabilities for each population

	Response:	No. of births/cow		
Population	2	3	4	Total
1: Site 1 Treat 1	13 (43.3)	13 (43.3)	4 (13.4)	30
2: Site 1 Treat 2	3 (9.1)	19 (57.6)	11 (33.3)	33
3: Site 1 Treat 3	4 (11.4)	23 (65.7)	8 (22.9)	35
4: Site 2 Treat 1	9 (37.5)	13 (54.2)	2 (8.3)	24
5: Site 2 Treat 2	13 (46.5)	14 (50.0)	1 (3.5)	28
6: Site 2 Treat 3	8 (33.3)	11 (45.8)	5 (20.9)	24

Response functions/population: two marginal probabilities p_1, p_2,

where, p_1 = proportion of cows with 2 calves

p_2 = proportion of cows with 4 calves

3. CONTRIBUTIONS TO CASSAVA

3.1 BIOMETRY: METHODOLOGICAL CONTRIBUTIONS TO CASSAVA RESEARCH. SELECTED EXAMPLES.

Biometry collaboration with specific research disciplines of the Cassava Program will be highlighted in this report. It includes collaborative data analysis projects and methodological studies carried-out with the Breeding, Entomology, Phatology and Economics Sections.

Given the nature of the research carried-out by the Cassava breeding section, responsible for generating and handling large number of genetic material through various evaluation and selection stages, a very close collaboration with Biometry has always existed. Collaborative data analysis studies based on preliminary yield trials, advanced yield trials and regional trials information, have been carried-out to answer relevant research questions to the breeders. Examples # 1 and # 2 are illustrations of this. Example # 3 illustrates a collaborative study with the Entomology Section: A methodology for the statistical analysis of electrophoretic patterns of populations of *Amblyseius limonicus garman* and Mc. gregor, a beneficial mite that acts as biological control of *Mononychellus tanajoa*, a mite that heavily attacks the cassava plant causing serious yield losses. Example # 4 illustrates a methodology for the statistical analysis of physiology research projects, in which a large number of plant response variables as well as environmental indicators are periodically measured. Example # 5 illustrates an area of collaboration between Biometry and the Cassava Economics section; ie. Design and Analysis of Sample Surveys, to quantify technology adoption at farm level.

Example # 1:

ADAPTABILITY ANALYSIS WITH UNBALANCED SETS
CASE: CASSAVA YIELD TRIALS IN 5 ECOZONES IN COLOMBIA

E. Granados and C. Hershey (1990)

34 varieties
35 regional trials
(1979-1986)

The adaptability of a genotype is defined as "its physiological response to improvement in environmental quality". For environmental quality we understand the combination of soil conditions, climate, pests, diseases, weeds, and aspects of the management of vegetable material (planting, evaluation techniques, sampling errors, among others). To quantify "environmental quality", several alternatives have been proposed. The most accepted is to express it by means of the overall mean of crop yield in that particular environment. The yield--as a resultant factor of the interaction of soil-climate and biotics-plant factors--expresses the potential quality of that environment for the growth of a given genotype of the crop.

In general, statistical methods to analyze the adaptability of genotypes through a wide range of environmental conditions, assume that the entire set to genotypes is evaluated in the same set of locations during the same years. That is, these methods are valid for balanced datasets. Such methods are: a) Regression methods (Yates and Cochran, 1938; Finlay and Wilkinson, 1963;

Eberhart and Russel, 1966); b) Multivariate methods, *like Principal Component Analysis* - a reduction of dimensionality technique - (Pearson, 1901; Hotelling, 1933) and Cluster Analysis - a classification technique - (Abou-el-Fittough, 1967; Rawling and Miller, 1969; Mungomery et al, 1974; Byth et al, 1976; Fox and Rosiete, 1982); and c) Geometric methods like Principal Coordinates Analysis (Schoenberg, 1935), Multi- dimensional Scaling and Correspondence Analysis.

The british statistician, P.G.N. Digby, (Unit of Statistics, Agricultural Research Council, University of Edinburg), developed a method for adaptability analysis for unbalanced datasets: "Modified regression analysis for incomplete variety x environment data" - (J. of Agricultural Sciences, Cambridge, 1979, q3, p. 81-83). We applied this method for the adaptability analysis of 34 cassava varieties evaluated in 53 cassava yield trials conducted between *1979 and 1986 in five ecozones* of Colombia.

The method is based on the model (non-linear in the parameters)

$$\boxed{Y_{ij} = \mu_i + b_i A_j + e_{ij}} \qquad \sum_{j=1}^{53} A_j = 0$$

yield of variety i in environment j

expected mean for variety i

Adaptability coeff. for variety i

Relative environmental
performance w.r.t. an "average" environment.

error

It permits statistical comparisons of **varieties** tested in different sets of locations and years (as μ_i estimates the excepted mean of variety i in an <u>average</u> environment). Also the method allows the comparison of **sites** where not the same set of varieties were tested (as A_j estimates an "*environmental* index" expressed as the environment yield potential with respect to the <u>average</u> environment). Based on this method, adaptability indexes were estimated and *a classification of* the 34 cassava varieties was made, using as classification criteria: (μ_i/st. error of μ_i) and (b_i/st. error of b_i). Figure 1 shows 10 of the varieties with their adaptability index (horizontal axis) and the environmental range where they were tested (vertical axis). Figure 2 shows the 8 variety groups obtained (Ward's minimum distance Cluster Analysis).

Example # 2:

A METHODOLOGY FOR THE STATISTICAL ANALYSIS OF ELECTROPHORETIC PATTERNS
Case study: Biochemical differentiation of populations of the mite *Amblyseius limonicus* Garman and Mc. Gregor (Acarina: Phytoseiidae)

M.C. Duque, Ma. E. Cuellar and Ann Braun
(Work in progress)

In order to determine an effective strategy for the biological control of a serious cassava pest -the mite *Mononychellus tanajoa* (Bondar) (Acarina: Tetranychidae) ("acaro verde de la yuca")- it is necessary to clearly characterize its natural enemies, both in terms of their ecologic and biological behavior. Among them, the mite *Amblyseius limonicus* Garman and Mc. Gregor (Acarina: Phytoseiidae) is known as its most important predator.

The present study was carried-out to make a biochemical differentiation of populations of the mite *A. limonicus* and to test the hypothesis that variability observed between populations of distinct geographic origin may be associated with differences biochemical patterns between them.

222 samples of *A. limonicus* collected in 16 distinct sites of Tropical America were submitted to electrophoretic analysis utilizing the isoenzymes GOT and MDH. The presence or absence of 70 electrophoretic bands (representing 70 distinct proteins or protein fractions in the *A. limonicus* DNA) were recorded for each one of the samples. In this way, the resulting data set was constituted by 222 rows (samples) and 70 binary (0,1) response variables.

For the statistical analysis of the electrophoretic binary results, a Correspondence Analysis was applied. This technique, a reduction-of-dimensionality technique for categorical variables, similar to the Principal Components Analysis, finds a low-dimensional graphical representation of the 222 samples. In this way, visual groups of samples are formed, being these groups interpreted as possible distinct populations of the mite *A. limonicus*.

As a result, six distinct groups were identified in a 3-dimensional graphical representation (a reduction of the 70 initial binary response variables) as illustrated in figure 1. The six groups corresponded to samples of *A. limonicus* of similar geographic origin. The hypothesis of association between geographic origin of *A. limonicus* and their distinct biochemical composition was then accepted.

Example # 3:

CHARACTERIZATION OF GENETIC AND ENVIRONMENTAL FACTORS INFLUENCING YIELD AND PHYSIOLOGICAL PROCESSES OF CASSAVA GENOTYPES

M. El-Sharkawy, M.C. Amézquita, H.F. Ramirez and G. Lema
(in progress)

In order to establish a methodology for the screening of cassava varieties well adapted to low P levels and with excellent physiological response to the environment, three experiments were conducted in CIAT Quilichao Research Station, between 1988 and 1991, in which 99 cassava varieties were evaluated by their *physiological response to a wide range* of environmental conditions, under two P levels: 0 and 75 kg/ha.

Seven plant architecture parameters (total leaf number, total leaf area per plant, mean leaf area, total leaf dry weight per plant, single leaf dry weight, number of nodes/plant and "leaf specific weight"), and six plant physiology parameters (photosynthesis rate, intercellular CO_2, transpiration rate, water use efficiency, stomatal conductance, mesophyll conductance) were measured under a very wide *range of environmental* conditions (light intensity, air temperature, leaf temperature, relative humidity and vapor pressure deficit). Fresh root yield was recorded at harvest.

Statistical analysis methodology covered four stages:

a) Reduction of the dimensionality of the problem via Principal Components Analysis. Given that some response variables were found to be highly correlated among themselves, the purpose of using this technique was to reduce the number of plant physiology, plant architecture and environmental variables to sub-sets of non-correlated ones.

b) The identification of cassava varieties with good adaptation to low P *and with good* physiological response to increments on P was done through standard Anova models on root yield. High yielding varieties with a statistically significant response to P were selected.

c) Mathematical models were constructed to explain the varieties physiological response to changes in environmental conditions. Models were fitted for two groups of varieties: those adaptated to low P with significant response to increments in P (identified in (b)), and the rest.

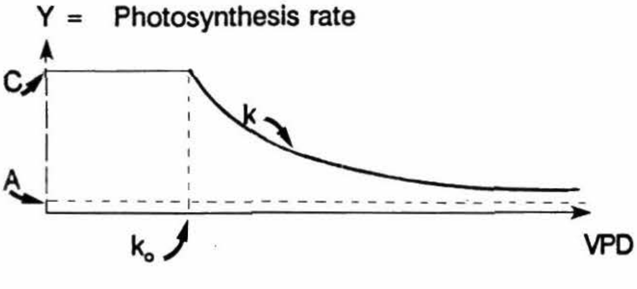
d) Using the model parameter *estimates, as classification criteria*, the 99 cassava varieties were grouped in homogeneous classes. The Cluster Analysis technique with Ward's minimum variance method was utilized for this purpose.

Results:

As a result of the Principal Components Analysis, plant physiology response variables were reduced *to* **two** *independent ones*: Photosynthesis rate and water use efficiency; plant architecture variables were reduced to **three**: total leaf number, mean leaf area and "leaf specific weight"; and vapor pressure deficit (VPD) was considered the most relevant environmental indicator.

The relationship between photosynthesis rate = f(VPD) was expressed by the mathematical model

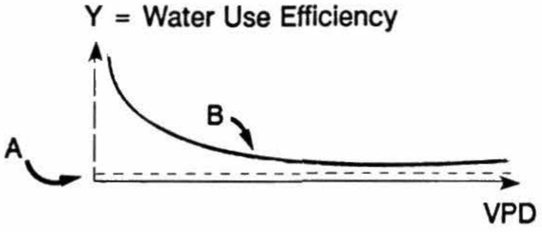
$$Y = \begin{cases} C & \text{,if } x \le k_o \\ A + Be^{-kVPD} & \text{, if } x > k_o \end{cases}$$



where, C = maximum photosynthesis rate
A = Photosynthetic rate at stabilization
k_o= VPD level for maximum photosynthesis
k = rate of decline of photosynthesis capacity

The relationship between Water use Efficiency = f(VPD) was expressed by the model

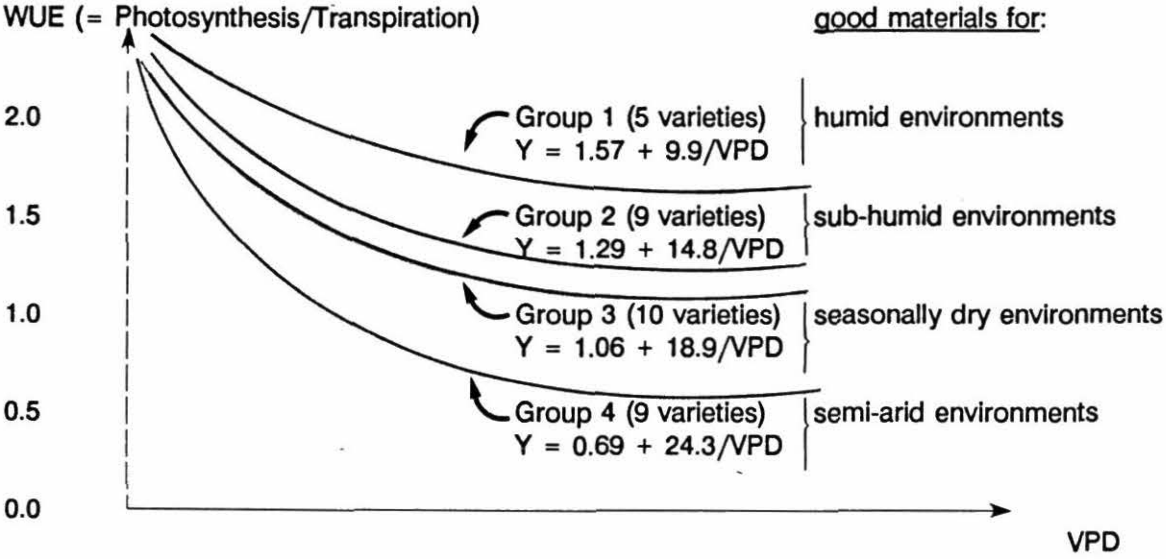
$$Y = \quad A + B/x$$



These two models were found effective to classify cassava varieties according to their physiological response to environmental conditions. The grouping of the 33 varieties tested on the first year, using as classification criteria the model parameters, A and B, from the relationship.

Water Use Efficiency = A + B/VPD,

allowed the identification of four distinct groups as shown in the figure:



Example # 4:

DESIGN AND ANALYSIS OF SAMPLE SURVEYS TO QUANTIFY TECHNOLOGY ADOPTION AT FARM LEVEL Case study: Adoption and impact of dry cassava technology in the north coast of Colombia

P. P. Perdomo and G. Henry

(Work in progress)

The broadening of traditional cassava demand for introducing dried cassava for the animal feed industry has stabilized cassava prices and increased overall cassava demand. This has served as a strong incentive for cassava farmers to increase their demand for cassava production technologies. Hence it was hypothesized that as a result, farmers would increase cassava area and intensify production.

In order to test the above hypotheses, a survey was conducted during the first semester of 1991 in one important cassava ecozone -the north coast of Colombia- to quantify adoption and impact of dry cassava technology.

Biometry methodological support for this type of adoption studies concerns: a) the definition of the sample frame (coverage, sample size); b) the identification of stratification criteria that highly influence the desired response (yield estimation, for example); c) *the definition of the sampling* design; d) the methodology for statistical data analysis consistent with the sampling design; and e) based on the initial results, the identification of new possible stratification criteria, including agroecological as well as socio-economic factors, to increase precision in the estimations obtained from the survey.

The target population for the present survey were small cassava production farms, located across 91 municipalities within five Departments in the North Coast of Colombia (Cordoba, Sucre, Bolivar, Atlantico and Magdalena). The 91 municipalities were stratified according to "level of technology influence" (low, medium and high), depending on the concentration of cassava-drying plants and the presence or absence of institutional support in the area (presence of ICA, CIAT, NGO'S activities). The sampling unit was the farm. Sampling design corresponded to a "Three-stage stratified sample in clusters". The clusters corresponded to the 5 departments (geographic division), the stratification criteria for municipalities was the "level of technology influence", *and farm selection* was carried-out in three-stages: 1) selection of municipalities within Department, according to an stratified random sampling procedure, with a proportional allocation of municipalities according to the Department size; 2) selection of a fixed number of "veredas" within each selected municipality (3 "veredas" per municipality); and 3) selection of 4 farms per "vereda". A total sample size of 444 farms was selected from 38 municipalities within the five Departments, with a significance level of $\alpha = 0.05$.

Preliminary results are reported in the Cassava Program Annual Report, 1991 (chapter 25).

3.2 DATABASES DEVELOPMENT FOR THE CASSAVA PROGRAM CASE: CASSAVA BREEDING DATABASE

F. Rojas, E. Granados, N. Marín, from the Data Services Unit
C. Hershey, C. Iglesias and team, from the Cassava Program

The purpose of the Cassava Breeding Database is to store in an organized way, maintain and make interactively available to cassava researchers, all the information related to the collection, generation, and testing of cassava germplasm, including research results attained by the Cassava Breeding Section and the CIAT's Genetic Resources Unit during a 12-years period: 1978-1991.

The Cassava Breeding Database includes at present the following information:

1. Germplasm Bank: 4,600 varieties with
   - passport data (collection site descriptors, collection date, origin, local names, etc.)
   - morpho-agronomic characteristics (19 descriptors on 4081 varieties)
   - electrophoretic characterization (presence or absence of electrophoretic bands).
2. Crosses: 12,648 crosses with their names and their parent's names.
3. Progeny Evaluation trials: F_1 evaluations (non-replicated trials. Results correspond to original observations per material).
4. Advanced trials: Data stored correspond to statistical summaries of replicated experiments for the evaluation of advanced cassava material, conducted by the Cassava Breeding Section between 1978 and 1990, with an average of 60 experiments/year, as follows:
   - Observational trials: 98 experiments
   - Preliminary yield trials: 76 experiments
   - Advanced yield trials: 116 experiments

   For each one of these trials, the following information is stored:
   - experimental site descriptors (10 variables)
   - morphological response variables (20 variables)
   - agronomic response variables (6 variables)
   - pests/diseases scores (variable number of response depending on the trial)
5. Regional Trials: Statistical Summaries of 421 replicated trials conducted in 5 cassava ecozones in Colombia between 1978 and 1990. For each regional trial the following information is stored in the database:
   - experimental site descriptors
   - agronomic responses per variety
   - pest/diseases scores per variety
   - Statistical Summaries of the trial for the most important response variables.
6. Elite Clones: 284 clones are stored, with the following descriptors:
   - clone code
   - parent's code
   - principal (and second principal) adaptation zone
   - yield potential (mean yield in the principal adaptation zone)
   - quality descriptors
   - resistance to diseases score (5 diseases)
   - morphological descriptors (6 descriptors)
7. Seed inventory and international shipments according to seed type: (50 shipments/year approx.)

   a) dry stakes b) immature stakes
   c) in-vitro d) sexual seed

The initial design and partial implementation of the Cassava Breeding Database started in 1984, using the software product IDMS/R (initially supported by Cullinet Corporation of America, and now supported by Computer Associates Inc.). However, given the technical limitations of this software - previously explained in this report-, which have been greatly responsible for the unsuccessful implementation of the various database applications in the past, the Data Services Unit started in July 1990 a series of software evaluation projects whose purpose was to analyze different software alternatives to IDMS/R. Three database management software products were studied as a complement or replacements to IDMS/R: SYSTEM 2000, from SAS Institute Inc., Raleigh, North Carolina, as a complement; SQL/DS, from IBM and ORACLE, from Oracle Corporation, as possible replacements. A sub-sample of the Cassava Breeding Database was used as the "pilot database" for the last two software evaluation projects. In September 1991 the CIAT Information Management Committee accepted the DSU proposal to acquire ORACLE as the database management software for CIAT's mainframe computer and microcomputers environment.
As a sub-sample of the Cassava Breeding Database had been used as the "pilot database" for the software evaluation projects, it was the first one to be ready for release when ORACLE was accepted.

Now, the Cassava Breeding Database is complete, fully operational and ready to be made available to the Cassava Breeding Section and other cassava researchers by the end of November 1991. It can be accessed through a terminal (or connected micro) of the IBM 4361 mainframe computer or through a PC with 4Mb of memory/40Mb hard disk and equipped with ORACLE tools.

Fig. 1: ENVIRONMENT RANGE FOR VARIETAL EVALUATION

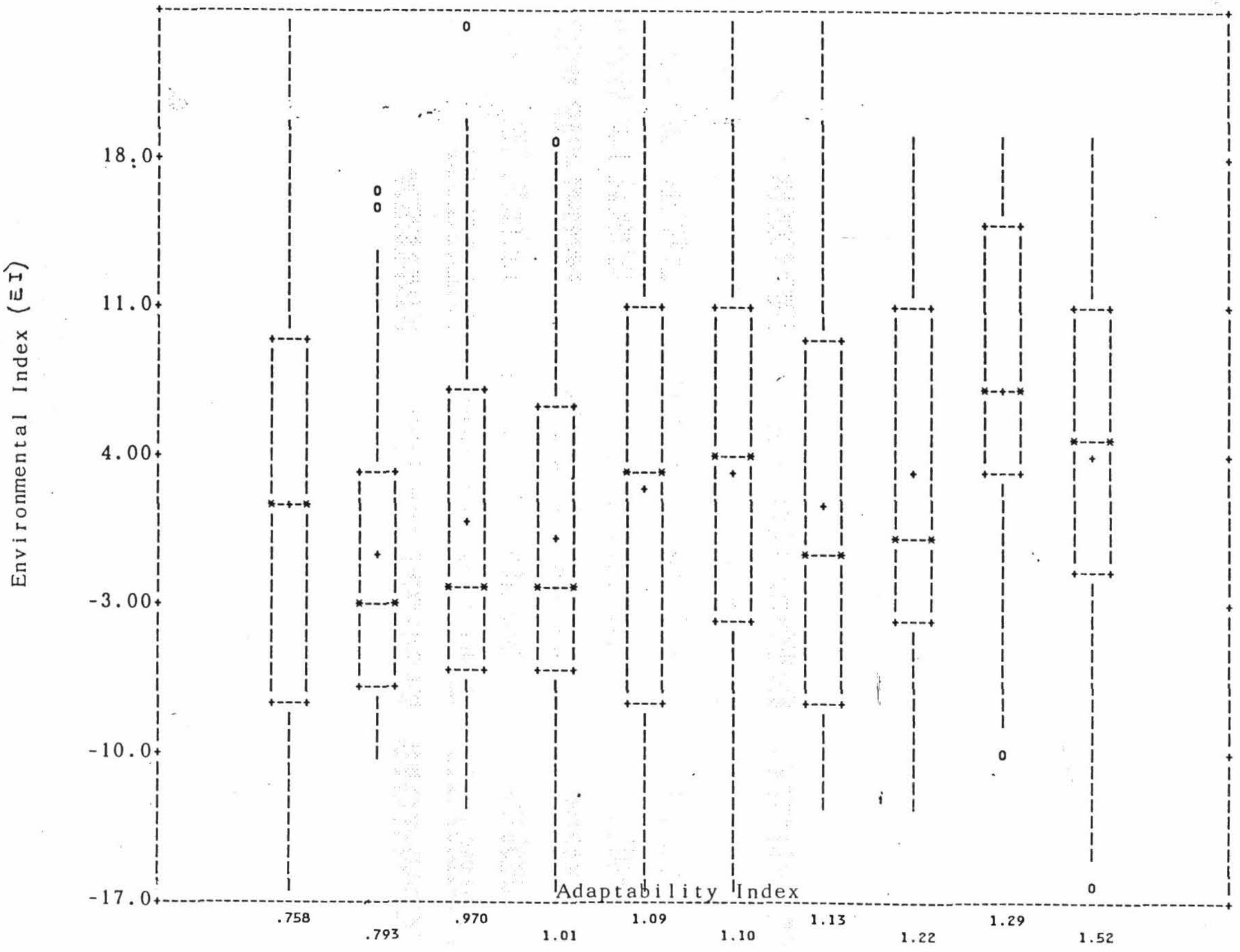


Fig. 2 CLASSIFICATION OF CASSAVA VARIETIES ACCORDING TO THEIR PRODUCTION AND ADAPTABILITY

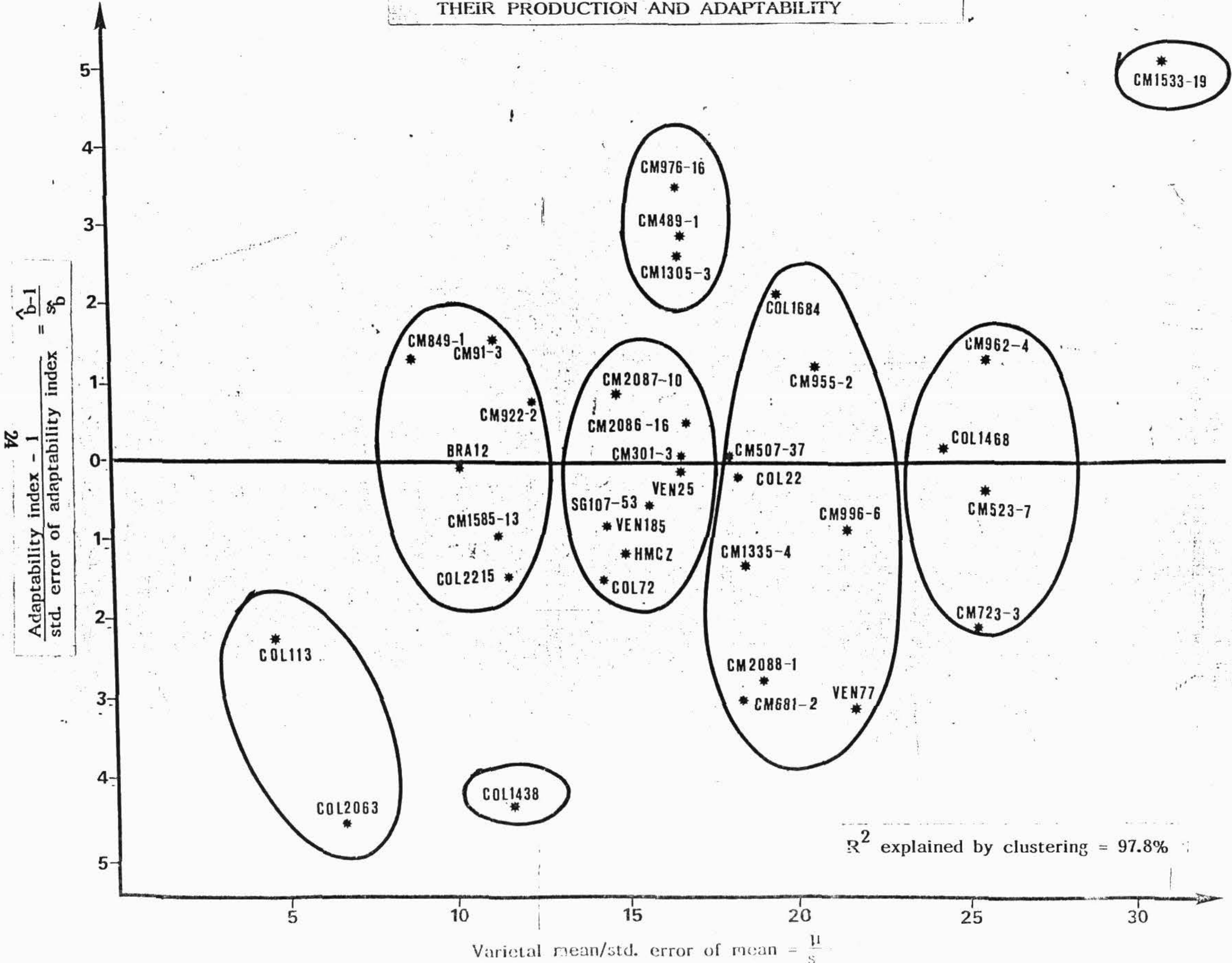
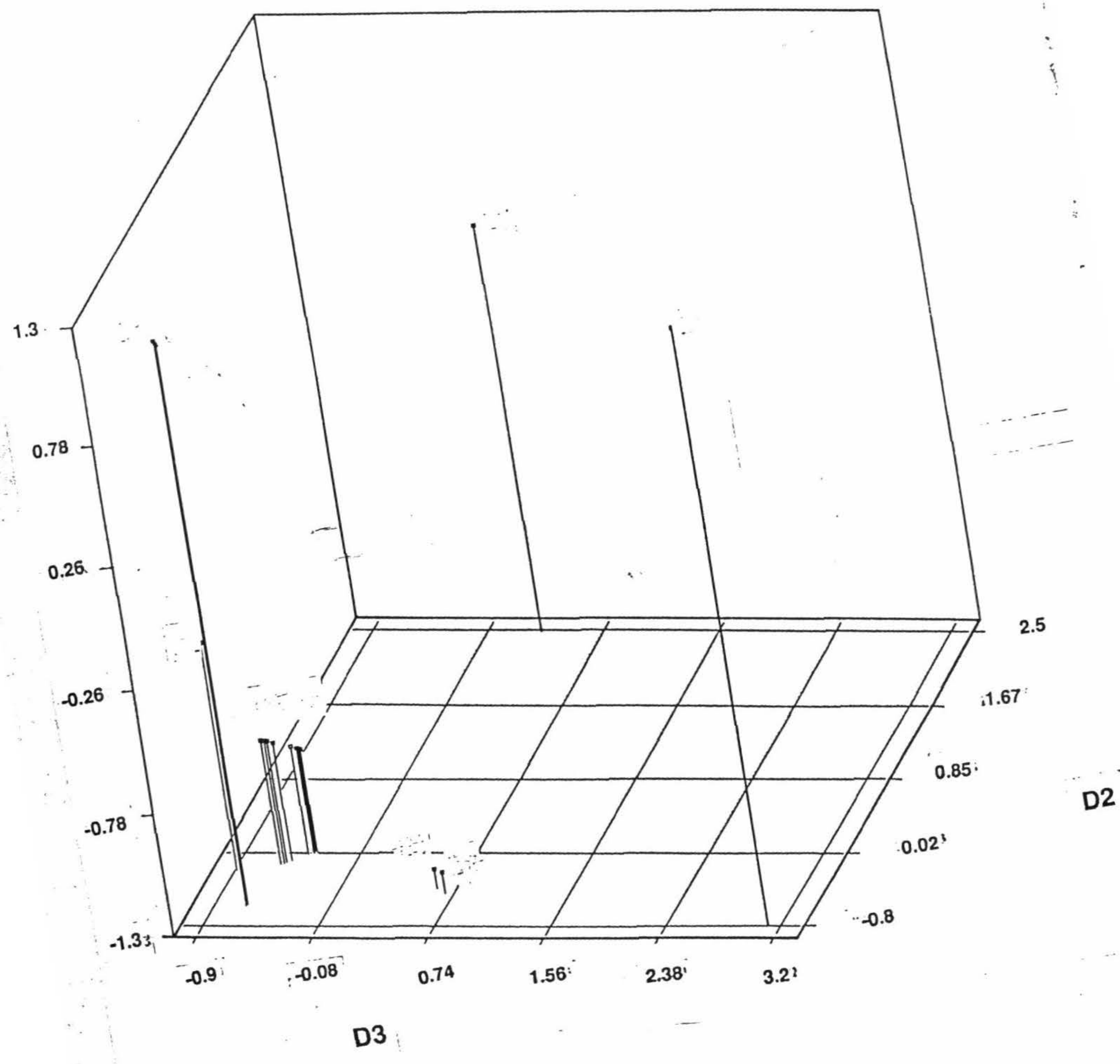


Fig. 3: Graphic representation of the six different groups (populations) of the mite *Amblyseius limonicus* in terms of the 3 principal components resulting from Correspondence Analysis.



4. CONTRIBUTIONS TO RICE

4.1 BIOMETRY: METHODOLOGICAL CONTRIBUTIONS TO RICE RESEARCH. SELECTED EXAMPLES.

Example # 1:

A METHODOLOGY TO DETERMINE THE MINIMUM EVALUATION PERIOD FOR DISEASE-RESISTANCE CHARACTERIZATION IN RICE

E. Guimaraes, M.C. Amézquita, G. Lema and F. Correa
(work in progress)

Santa Rosa Experimental Station, located at the eastern Colombian savannas (at 333 m.a.s.l., 25°C, 66-87% relative humidity) is used by the CIAT Rice Program as a hot spot site for screening breeding lines for the prevalent diseases in Latin America. Given the high variability in disease pressure, even at this hot spot, varietal characterization scores may vary from one semester to the next. An objective criteria to decide on the minimum evaluation period required to characterize rice varieties by their disease reaction in Santa Rosa supports an efficient use of research resources and represents a methodological contribution to partner institutions.

Results on disease-evaluation trials conducted at Santa Rosa Station during a 4-year period were used to accomplish this objective. Data source selected for this study corresponds to disease-reaction scores on 70 varieties commercially grown in Latinamerica, evaluated through 7 consecutive semesters (4 semesters "A", under high rainfall (242 to 460mm/month) and 3 semesters "B", under lower rainfall (25 to 36mm/month)) between 1987 and 1990. Disease evaluations include: 1) leaf blast (LBl), at 42 days after sowing; 2) leaf scald (LSc), at flowering time; 3) neck blast (NBl), 30 days after flowering and 4) grain discoloration (GD), 30 days after flowering. Disease reaction was recorded using the 0-9 ordinal scale from the "Standard Evaluation System for Rice".

Data analysis methodology consisted on:

a) Simulate 5 different experimental period lengths: of 7, 6, 5, 4 and 3 consecutive semesters.

b) For each period length, and for each disease separately, characterize the 70 varieties according to three criteria:

   M = overall mean disease reaction for the variety across the experimental period.

   b = varietal response to increments in the site disease pressure, (b is the slope of the regression of variety disease reaction (Y) vs. site disease pressure (X).

   S_b = standard error of b.

   In this way, the 70 varieties were characterized in five different ways, associated with the five different experimental period lengths.

c) Correlation matrixes between the M's, the b's and the S'_bs obtained from the different experimental periods, (longest period was considered as the most reliable one) and between semesters "A" vs. semesters "B" were calculated.

d) Based on the correlation pattern, a decision on the minimum acceptable period was made.

Results: Table 1 shows the high variability in disease pressure at Santa Rosa Station. Table 2 shows values of M, b and S_b for groups with similar Leaf Blast reaction among the 70 varieties evaluated, based on a 7-semester evaluation period. Fig. 1 illustrates various patterns of varietal response to Leaf Blast pressure. Tables 3 and 4 show the correlations between varietal disease-reaction scores (M's), and between varietal response to disease-pressure (b'_s) based on a 7-semester evaluation period (the most reliable one), vs. shorter periods.

Conclusion:

a) According to the disease reaction (M), the high correlations found between M estimates suggest that a shorther period (3-semesters) would be appropriate to rank and group the rice varieties.

b) Characterization of rice varieties according to their **response** to disease-pressure (b), based upon the correlations found between b estimates, suggest that a long period (7-semesters) with **contrasting** disease-pressure levels is necessary.

Example # 2:

CHARACTERIZATION OF RICE VARIETIES FOR THEIR REACTION TO THE "Hoja Blanca" Virus: A call for the need of further replication and adjustment by a known check

F. Cuevas and G. Lema (1991)

With the objective of characterizing the reaction of Latin American rice (*Oryza sativa L.*) varieties to the Hoja Blanca virus, 107 varieties and 5 known checks were exposed to a colony of the insect *Sogatodes oryzicola* (Muir), in one greenhouse and three replicated field experiments, in CIAT Palmira Experimental Station, Colombia during 1987-1988. In the field experiments, the varieties were planted using 1 row/variety. Check varieties: Bluebonnet 50 (Resistant(R)), Cica 8 (moderately susceptible (MS)), Metica 1 (S), Oryzica 1 (moderately resistant (MR)), and Colombia 1 (R), were planted every 40 rows in the field and at random in the greenhouse. Response variables considered were the percentage of diseased plants per variety/row, non-adjusted and adjusted by disease-reaction in adjacent plots of each one of the checks. In this way, the 107 varieties were scored under 6 different criteria in the greenhouse and in each field experiment. T exhibiting the highest consistence (highest correlation) in the varieties score within and between experiments, was considered the most effective characterization criteria for reaction to Hoja Blanca.

Results and conclusions:

1) Both, between and within experiment correlations were low and sometimes non-significant, suggesting the need to replicate even more this type of trials.

2) When the percentage of diseased plants per variety was adjusted by disease reaction in adjacent plots of Blubonnet 50, the resistant check, the correlation coefficients improved.

3) Based on this last criteria, the 107 varieties evaluated were classified as R, MR, MS and S, according to their deviation from Blubonet 50. At least 50% of the varieties evaluated were classified as S. Commercial varieties from countries having the risk of Hoja Blanca epidemics were in class MS or better.

(Results are reported in the Rice Program Annual Report 1990).

Example # 3:

THE USE OF CATEGORICAL DATA ANALYSIS METHODS TO ANALYZE A COMPLEX GENETIC MODEL
Case: Genetics of Rice Hoja Blanca Virus resistance

E. Granados, F. Cuevas and H.F. Ramirez (1991)

Within the CIAT's Rice Program Strategy concerning research work on Hoja Blanca genetic resistance, an experiment was conducted at CIAT-Palmira, Colombia between 1989 and 1991 in order to understand the genetic base of different resistance sources. Parents, F_1 and 100 F_3 families from 6 crosses (between six resistance sources with the susceptible cultivar Blubonet 50) and 15 single-cross (combinations among the six resistance sources), were evaluated in the greenhouse for their reaction to Hoja Blanca.

The response variable considered was a binary variable "presence or absence of symptoms of disease", (0,1), recorded on each seeding. F_3 families were grouped based on this binary response, using the GSK (Grizzle, Starmer and Kock, 1969) methodology for the statistical analysis of binary data. The GSK method fits a log-linear model to a function of cell proportions $-\log(p_1/p_2)$, where

p_1 = proportion of plants with presence of symptoms
and p_2 = proportion of plants with absence of symptoms
to study difference in Hoja Blanca reaction between F_3 populations.

The complex nature of this problem concerns the experimental design: a CR design with 103 "treatments" (2 parents, 1 F_1 and 100 F_3 populations) with between 50 and 100 replications per treatment (individual plants). So, the total number of experimental units was 5,300, each reporting presence/absence of sumptoms.

Results of this experiment are reported in the 1991 Rice Program Annual Report.

Example # 4:

A METHODOLOGY FOR THE ESTIMATION OF "PRODUCTION EFFICIENCY" INDEXES. Case: Economic efficiency in the use of rice inputs

A. Ramirez and M.C. Duque
(Work in progress)

Economic analysis of production functions involves the fitting of equations of the form $Y=f(X_1, X_2,...,X_n)$, where Y represents the production of a given crop and the $X_i{}'_s$ represent production inputs. These multiple regression equations very often exhibit three main statistical problems:

a) Multicolinearity, or lack of independence, between the $X_i{}'_s$ variables.

b) Heteroscedasticity, or variance heterogeneity in the response variable (Y) among the levels of one or more of the input factors ($X_i{}'_s$).

c) Non-normally distributed errors:

Problem a) can be corrected by reducing the $X_i{}'_s$ to a non-correlated set of variables -through the use of Principal Components Analysis or through the use of Ridge Regression, an algorithm that corrects the over-estimation of variance associated with multi-colinearity.

Problem b) can be corrected through the use of transformations, the most commonly applied being the log transformation on the Y and the $X_i{}'_s$. In this case, the resulting function is a Cobb-Douglas.

It has been referred in the economics literature (Aigner, Lovell and Schmidt, 1977) that problem c) may indicate the presence of **inefficiency** in the use of some production inputs. If this is the case, it is necessary to calculate the so called "production efficiency indexes" (**technical**, **productive** and **distribute** efficiency indexes) for each of the inputs factos in order to precisely identify where the inefficiency occurs.

The purpose of this work is to contribute with a methodology for calculating such production efficiency indexes.

Methodology covers four stages:

a) The estimation of the production function, correcting problems of multicolinearity and heteroscedasticity.

b) Analisis of residuals, to investigate their departure from simetry and normality.

c) If the errors (e) are not normally distributed, it is then necessary to estimate two error components, U and V, such $e=U+V$. Depending on the distributions of U and V error components, the production function is classified as a "stochastic frontier function" or as a "deterministic frontier function".

d) Once identified and estimated the frontier function, the "production efficiency indexes" (global and for individual inputs) are calculated (following the methodology of Farrell Generalized Model).

Partial results of these analysis are reported in the 1991 Rice Program Annual Report.

Table 1: Variability in disease-reaction scores for rice evaluation at Santa Rosa Experimental Station, Colombia, during a 4-years period (1987-1990)

	Disease-reaction[1/] to			
Year/Semester	*Leaf Blast*	Neck Blast	Leaf Scald	Grain Discoloration
1987 A	5.0	4.3	4.1	3.0
1987 B	4.9	4.8	3.5	3.7
1988 A	5.4	5.2	3.2	3.5
1989 A	4.4	4.1	3.0	2.9
1989 B	3.2	5.0	3.8	3.3
1990 A	4.8	4.7	4.6	3.1
1990 B	1.8	2.2	0.8	2.2
- Overall mean and standard deviation	4.3 (2.1)	4.3 (2.3)	3.3 (1.7)	3.1 (1.5)
- CV (%)	49.0	53.0	52.0	48.4

[1/] Scores are recorded on a 0-9 ordinal scale.

Table 2: Values of M, b and S_b for groups with similar Leaf Blast-reaction, based on 7-semester evaluation period.

GROUP (no. of varieties)	M	b	S_b	INTERPRETATION
1 (n = 10)	2.36	0.54	0.17	Resistant and stable
2 (n = 7)	2.34	0.68	0.37	
3 (n = 14)	3.59	1.21	0.33	Intermediate resistance Respond to increased pressure
4 (n = 8)				
5 (n = 13)	5.45	0.81	0.29	Susceptible and respond to increased pressure
6 (n = 12)	5.34	1.39	0.20	
7 (n = 6)	5.22	1.43	0.55	

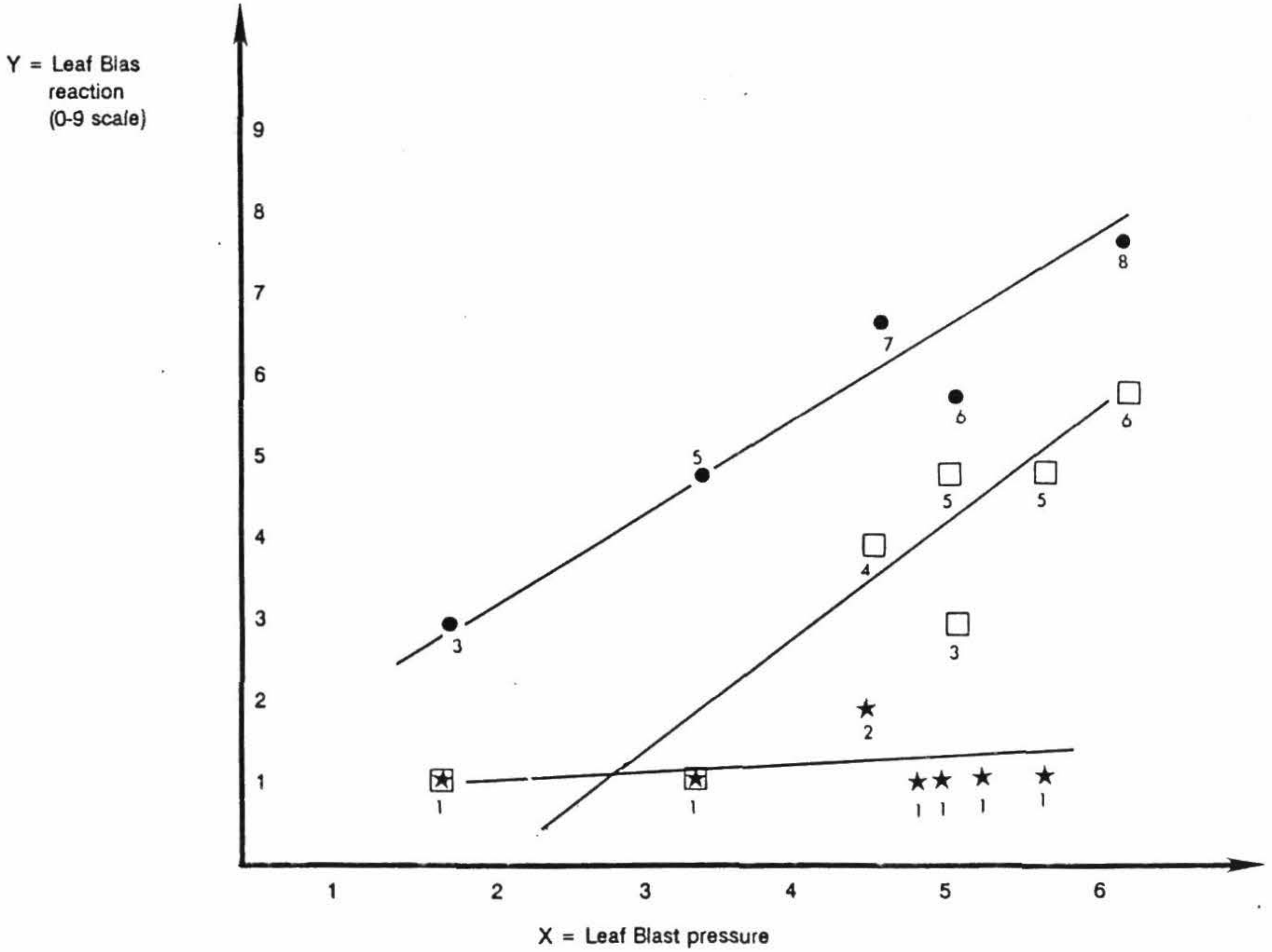
Table 3: Correlations between varietal disease-reaction scores (M's) based on a 7-semester evaluation period (the most reliable), vs. shorter evaluation periods.

	Period length						
DISEASE	6-Semester	5-Semester	4-Semester	3-Semester	Only "A" Semesters	Only "B" Semesters	Between "A" and "B"
				correlation coeff. and its probability of significance			
Leaf Blast	0.99 (0.0001)	0.98 (0.0001)	0.97 (0.0001)	0.92 (0.0001)	0.97 (0.0001)	0.94 (0.0001)	0.85 (0.0001)
Neck Blast	0.99 (0.0001)	0.98 (0.0001)	0.94 (0.0001)	0.91 (0.0001)	0.92 (0.00001)	0.89 (0.0001)	0.64 (0.0001)
Leaf Scald	0.97 (0.0001)	0.94 (0.0001)	0.90 (0.0001)	0.89 (0.0001)	0.90 (0.0001)	0.86 (0.0001)	0.58 (0.0001)
Grain Discoloration	0.98 (0.0001)	0.95 (0.0001)	0.89 (0.0001)	0.86 (0.0001)	0.89 (0.0001)	0.86 (0.0001)	0.54 (0.0001)

Table 4: Correlations between varietal response to disease pressure (b'_s), based on a 7-semester evaluation period (the most reliable), vs. shorter evaluation periods.

	Period length						
DISEASE	6-Semester	5-Semester	4-Semester	3-Semester	Only "A" Semesters	Only "B" Semesters	Between "A" and "B"
				correlation coeff. and its probability of significance			
Leaf Blast	0.57 (0.0001)	0.56 (0.0001)	0.45 (0.0001)	0.25 (0.03)	0.34 (0.004)	0.77 (0.0001)	0.19 (0.11)
Neck Blast	0.38 (0.01)	0.31 (0.08)	0.23 (0.05)	0.22 (0.07)	0.28 (0.02)	0.80 (0.0001)	-0.21 (0.08)
Leaf Scald	0.68 (0.0001)	0.48 (0.0001)	0.37 (0.001)	0.21 (0.07)	0.68 (0.0001)	0.67 (0.0001)	0.09 (0.46)
Grain Discoloration	0.58 (0.0001)	0.44 (0.0001)	0.54 (0.0001)	0.25 (0.04)	0.10 (0.39)	0.92 (0.0001)	-0.12 (0.32)

Fig. 1: Three varietal response patterns to Leaf Blast pressure



★ "Amistad 82"	: Mean Score = 1.57,	Y = 0.32 + 0.30 X ,	S_b = 0.13 , R^2 = 49%
□ "Bluebonnet 50"	: Mean Score = 4.96	Y = 2.30 + 1.03 X .	S_b = 0.31 , R^2 = 78%
● "Metica 1"	: Mean Score = 6.0	Y = 0.83 + 1.23 X ,	S_b = 0.20 , R^2 = 91%

5. CONTRIBUTIONS TO BEANS

5.1 BIOMETRY: METHODOLOGICAL CONTRIBUTIONS TO BEAN RESEARCH. AN EXAMPLE.

THE APPLICATION OF CORRESPONDENCE ANALYSIS TO INTERNATIONAL NURSERY DATA, Case: Analysis of AFBYAN data (The African Bean Yield and Adaptation Nursery)

J. García and B. Smithson (1991)

Statistical analysis methods for data generated by International Nurseries very commonly involve an **adaptability analysis**, a **Principal Components Analysis** and **clustering** of varieties and/or locations. The first methodology is applied to characterize varieties according to their yield and their physiological response to improvement in environmental quality; the second method, to reduce the dimensionality of the problem, and the third to group varieties with similar performance across locations, or to group locations according to their similarity in varietal performance. These methods have proved useful in the identification of good genetic material. However, they can only be applied to **continuous** response variables (such as yield (kg/ha), plant height (cm), days to maturity, etc.).

The Correspondence Analysis technique is a weighted Principal Components Analysis on binary (0,1) or categorical response variables. Then, if the varietal response variables for each location are (0,1) responses (for example, the variety i "was selected or not" at the location), this technique is applicable. It reduces the number of response variables to a reduced-set of independent ones (the "Principal Components") and then, through graphical representation of the locations in terms of the new axis, groups of locations that would select the same variety(ies) are identified.

The purpose of this study was to evaluate the usefulness of the Correspondence Analysis in the identification of promising varieties, by comparing its results with those obtained through Adaptability Analysis and Clustering. The results of 14 AFBYAN trials in which 21 bean varieties were evaluated in terms of yield, were utilized as source of information.

Data analysis methodology used covered four stages:

a) Adaptability Analysis for the 21 varieties across 14 locations.

b) Cluster analysis to group varieties according to their yield at **each** location.

c) Correspondence Analysis applied to a newly generated data set, with location as row variable, and binary response variables "the variety i was selected or not at the location". For this purpose, a variety was considered **selected** within a given location when its yield was greater or equal to the 85% of the maximum location yield.

d) Cluster Analysis to group locations based on the 3 first Principal Components resulting from Correspondence Analysis.

Results: The three types of analysis used identified the same group of varieties as the best ones. Groups of locations that would select the same varieties were identified by using the Correspondence Analysis. This study shows that the analysis of binary or categorical responses in multilocational trials is feasible.

5.2 PRESENT STATE OF THE BEAN BREEDING DATABASE.

G. Serrano, N. Marín, J. García, from the DSU
J. White, from the Bean Program

The initial design of the "Bean Database" started in 1983 using the software product IDMS/R (initially supported by Cullinet Corporation of America, and now supported by Computer Associates Inc.). It was initially conceived to contain all the information related to the collection, storage, generation, evaluation, multilocational testing, complementary-disciplines evaluation and international distribution of bean germplasm. However, given the technical limitations of the IDMS/R software -lack of flexibility for modifying a database design, lack of a powerful and user-friendly development and query tools, extremely long data-loading times, and lack of a flexible micro-mainframe interface-, the practical implementation of the conceived design was not successfully achieved.

Between 1983 and 1990, partial implementations of the Bean Database were completed using IDMS/R: a sub-set of the bean genetic resource data, information on crosses and advanced lines generated by the CIAT's Bean Program up to 1990 and VEF-EP-IBYAN nurseries were stored. However, the use of this database by the CIAT bean scientists has been limited, for lack of user-friendliness of the software.

With the decision made in September 1991, of replacing IDMS/R by ORACLE as the database management software for CIAT's mainframe, micro and future network's environment, all existing database applications needed to be re-designed and re-implemented in ORACLE.

The Data Services Unit first priorities for 1991 included the "Genetic Resources Database" and the "Cassava Breeding Database", which are now fully operational in ORACLE. So, the re-design and full implementation of the **bean genetic resource database** is already completed. However, only a very small portion of the "Bean Breeding Database" has been re-designed and implemented in ORACLE. This includes:

a) **Crosses:** **42,188** with their code (which includes crossing criteria) and their parent's names.

b) **Advanced lines:** Includes information on 15,000 advanced lines generated by the Bean Program breeders between 1978 and 1990. Descriptors include: line code, genealogy, and statistical summaries of agronomic and disease-resistance evaluations.

c) **Seed inventory and national/international shipments** (partial data is stored up to now)

The rest, including information on VEF-EP and all International Nurseries conducted by bean scientists, will be gradually added to this database, under a very close collaboration and advise from the Bean Program Leader and scientists. The addition of bean research results generated by other disciplines within the Bean Program will require a very careful planning during 1992.

6. CONTRIBUTIONS TO THE GENETIC RESOURCES UNIT
Case: The Genetic Resources Database

F. Rojas, G. Serrano, A. Franco, C. Saa and M.C. Amézquita, from DSU
M. Iwanaga, B. Maas, R. Hidalgo and team, from GRU

The purpose of the Genetic Resources Database is to store in an organized way, maintain and make interactively available to the GRU personnel, to CIAT Scientists and to NARI's partners, all the information related to the **collection, characterization, storage** and distribution of germplasm collections handled by CIAT between 1972 and 1990.

The Genetic Resources Database is composed at present by three sub-systems corresponding to:

a) The "Tropical Pastures Genetic Resources Database" (already described in section 2.2 of this report).

b) The "Cassava Genetic Resources Database" (already described as a sub-system of the Cassava Breeding Database in section 3.2 of this report), containing information on the cassava collection (4,600 accessions) with:
- passport data
- field characterization, and
- in-vitro characterization

c) The "Bean Genetic Resources Database" which has not been described in this report.

The "Rice Genetic Resource database" is at present handled directly by the CIAT Rice Program and is not yet included in this central database.

The "Bean Genetic Resource Database"

The Bean Genetic Resources Database includes at present the following information:

1) **Bean Germplasm Bank**: 25,000 *phaseolus accessions belonging to vulgaris, acutifolius, coccineus, and lunatus species. Out of the 32,500 accessions in stock (with "S" identification code), these 25,000 have passed through quarantine, have been characterized and are given the "G" code. They constitute the Germplasm Bank.*
Information on each "G" accession includes:
   . *passport data (collection site descriptors, collection data, origin, local names, etc.)*
   . *morpho-agronomic descriptors (60 descriptors on 25,000 accessions, product of experimental evaluations carried-out by the GRU and the CIAT's Bean Program).*

2) ***Bean Germplasm in Stock:*** (accessions with *"S" code but not in the germplasm bank): 15,400 phaseolus accessions with passport data.*

3) ***Seed inventory and international shipments:*** *24,000 out of the 25,000 germplasm bank accession have seed inventory information. Also, information related to seed distribution to research institutions and universities all around the world is included.*

Between September and the end of November 1991, the "Genetic Resources Database" has been re-designedand implemented in ORACLE software. It is fully operational now, and can be accessed through IBM 4361 terminals or through a PC with 4Mb of memory/40Mb hard disk and equipped with ORACLE tools.

More detailed information on this database appears in the 1991 GRU Annual Report.