ANNUAL REPORT 1993

BIOMETRY UNIT

For internal circulation
and discussion only

CIAT Internal Review 1993

CIAT
Centro Internacional de Agricultura Tropiocal

ANNUAL REPORT 1993

BIOMETRY UNIT

For internal circulation
and discussion only

CIAT Internal Review 1993

November 1993

Biometry Unit
Personnel 1993

Head: María Cristina Amézquita
Mathematical Statistician, MS and Adv. Dipl.

Statistical Consultants:

- Eloina Mesa
  Mathematical Statistician, MS
- Myriam Cristina Duque
  Mathematician, BS
- Germán Lema
  Industrial Eng. BS

Secretary: Marta Elena Carvajal B.

Content

BIOMETRY UNIT
Annual Report 1993

1. BIOMETRY UNIT: ROLE, ACTIVITIES, STATISTICAL SOFTWARE

Role

As a science-based institution, CIAT relies heavily on statistical and mathematical sciences and tools for research design and analysis of research results. The Biometry Unit is a research-support Unit of advisory and methodological nature in these areas. Its specific services and outputs are:

1) **Statistical/mathematical advice** in experimental design, data analysis methodology, interpretation of results, their forecasting ability, and final presentation.
2) Collaborative **methodological studies** and specific **data analysis projects** with CIAT scientists, aimed at responding relevant research questions. The biometrician contribution in this context is to evaluate and recommend appropriate experimental designs for a given research; identify and quantify sources of variation affecting specific response variables to support research planning; evaluate and recommend appropriate statistical analysis methodologies for a given research problem: their efficiency, accuracy and applicability.
3) Development of **software programs for end-users** to implement specific statistical analysis methodologies. These are called 'MACROS'.
4) **Training of personnel** from CIAT research programs and selected groups from NARD's in basic biometrical methods/research data analysis techniques.
5) Assistance to CIAT in defining centerwide standards for statistical/mathematical software. Present standards include SAS/BASICS, SAS/STATS, SAS/ETS, SAS/IML, SAS/GRAPH, SAS/OR, GENSTAT, MSTAT, GLMM, and AGROBASE/4.

Activities

An important activity of the biometrician is his/her involvement in **collaborative methodological studies and data analysis projects** with researchers, aimed at responding relevant questions of research. These projects utilize data generated by a given research project through the years, combine experimental results of a given research discipline, or combine data generated by various disciplines within a Program. The results of some of these projects have appeared as chapters of CIAT Programs publications, some as contributions to International Networks reports, some have been published as joint papers with the scientists, and some others are in progress. A brief summary of selected case studies are included in this report. They represent collaborative work between the Biometry Unit and CIAT research Programs/Units during 1993.

Basic **training** in statistical methods and data analysis was provided to 24 CIAT research associates/assistants during 1993: 7 from the Cassava Program, 8 from the Rice Program, 2 from Tropical Forages, 3 from Savanna, 3 from Hillsides and 1 from the Bean Program. It is hoped to renew the training activities for National Institution researchers from Latinamerican and African institutions, CIAT collaborators. The new Microcomputer Training Laboratory is expected to be used for this purpose. During the five years of existence of the old Laboratory, the Biometry Unit has offered a total of 35 one to two-week training courses, with a total number of 344 National Institution researchers trained from Latinamerica (260), Asia (24) and Africa (50). An approximate number of 105 participations from CIAT research associates/assistants have benefit from this effort during the last 6 years 1987-1992.

In the light of the new CIAT, new areas of biometrical expertise are foreseen in which the Biometry Unit complemented by invited Biometrician Consultants can add useful contributions. These are: 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-pastoral systems. c) Quantitative genetics/population dynamics. d) Econometric techniques in response to new expected demands from economists and Impact Assessment; and e) Geostatistics or Spatial Variability techniques.

Statistical Software

Statistical/data analysis software for the present IBM 4361 mainframe computer include: SAS/BASICS, SAS/STATS, SAS/GRAPH, SAS/ETS, SAS/IML and SAS/OR from SAS Institute Inc. Raleigh, North Carolina, USA; GENSTAT, 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, AGROBASE/4, from Agronomix Software, Manitoba, Canada; MATMODEL from Soil, Crop and Atmospheric Sciences, Cornell University, Ithaca, New York; Lotus 1-2-3 and Dbase III. All these software tools are expected to be installed under the new Unix-based computer network, and be made available to biometricians and scientists.

2. COLLABORATIVE METHODOLOGICAL STUDIES AND RESEARCH DATA ANALYSIS PROJECTS WITH CIAT RESEARCH PROGRAMS/UNITS

2.1 Collaboration with the Rice Program

Case Study 1:

A comparison between the Pedigree Method and Anther Culture in the generation of rice lines with stable resistance to blast: Use of Categorical Data Analysis Methods.

M.C. Amézquita, C. Martínez, F. Correa, G. Lema
(in progress)

Rice blast, caused by *Pyricularia grisea sacc.* is considered to be the single most important disease of rice on a world wide basis. Both the disease and the pathogen have been extensively studied. Development of resistant cultivars by the Pedigree Method (PM) has been the most extensively used method to control this disease. The production of doubled-haploids through Anther Culture (AC) has been proposed as an effective, efficient and economic breeding tool. This study compares the traditional breeding method, the PM, with AC method in their capacity to produce rice lines with stable resistance to blast.

Pedigree from 11 crosses of 3 types --a) Japonica/Japonica, Susceptible x Resistant, b) Japonica/Indica, Susceptible x Resistant, and c) Japonica/Indica, Susceptible x Susceptible-- made between a susceptible cultivar Fanny and other 11 rice varieties with varying degree of susceptibility to blast was used as genetic material for the study. (Table 1)

An initial number of 17,867 F_2 plants were submitted to blast selection by PM and advanced, obtaining 681 F_6 blast resistant lines. Using the AC method, 441 AC-F_2 and 740 AC-R_2 lines were generated using F_2 susceptible plants and F_1 plants respectively. After a field evaluation cycle of three semesters, 171 AC-F_2 and 178 AC-R_2 blast resistant lines were obtained. "Resistance stability" was defined as the percentage of resistant lines which remained resistant through the 3-semester evaluation cycle.

Table 1: Genetic Material

-Pedigree from 11 crosses-

Cross[1]	Parents
Japonica/Japonica SxR[2]	
. CT5782	Fanny/IRAT13
. CT8813	Fanny/TOX1011-4-1
. CT8816	Fanny/OS6
. CT8817	Fanny/LAC23
. CT8819	Fanny/IAC165
. CT8820	Fanny/ITA235
Japonica/Indica, SxR	
. CT8814	Fanny/Ceysvoni
. CT8815	Fanny/Tetep
Japonica/Indica, SxS	
. CT5780	Fanny/CICA4
. CT8821	Fanny/Colombia
. CT8818	Fanny/Carreon

[1] CIAT's designation

[2] Classification based on isoenzyme analysis

Statistical Analysis Methodology

As blast reaction is recorded under a 1-9 discrete scale with non-equally distant levels according to the International System for Rice Evaluation (IRRI, 1988), Categorical Data Analysis Methods were used for the statistical analysis. The statistical analysis methodology was divided in two steps:

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 blast resistance stability.

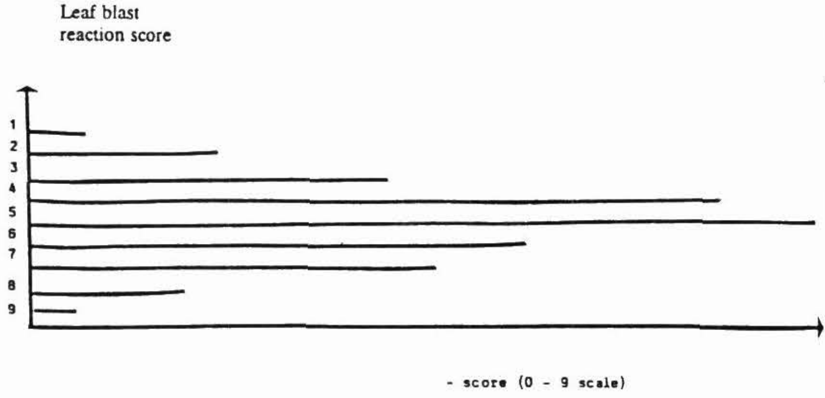
b) Inferential statistical analysis to assess the effect of 'method', 'cross type' and their interaction on blast resistance stability.

Three response variables were analyzed: Neck blast (NBL), leaf blast (LBL) and General reaction to blast (GBL), this last expressed for each line as the maximum score between LBL and NBL. A line was considered 'resistant to blast' when its blast reaction score was $\leq$ 4 within the 1-9 scale.

Descriptive Statistical analysis. For this purpose, *blast reaction was analyzed* as a 9-level categorical variable following the multinomial distribution. For each population generated by each method (PM, AC-F_2 and AC-R_2), the following descriptive parameters were calculated:

1. Initial number of lines generated.
2. Number and percentage of blast resistant lines in the initial population.
3. Mean score, skewness, median, mode and range of blast reaction score in the initial population.
4. Number of blast resistant lines at the beginning of evaluation cycle for resistance stability.
5. Number and percentage of stable resistant lines.

Figure 1 illustrates the meaning of each parameter of the multinomial distribution, in F_6 lines generated by PM. Tables 2 and 3 show overall performance of populations generated by the 3 methods in terms of LBL, NBL, and GBL reaction. Table 2 shows data on LBL, NBL, and GBL across crosses while table 3 shows performance per cross, in terms of GBL only.



N	%	cumulative %
11	1.9	1.9
34	5.9	7.8
67	11.6	19.3
125	21.6	40.9
139	24.0	64.8
90	15.5	80.3
76	13.1	93.5
29	5.0	98.5
9	1.6	100.0
580	100.0	

- Number of lines produced	= 580	- Skewness	= 0.04
- Number of resistant lines	= 237	- Median	= 5
- % of resistant lines	= 40.9	- Mode	= 5
- Mean reaction score	= 4.9	- Range	= 1 to 9

Fig 1: Overall distribution of leaf-blast reaction scores in F_6 lines generated by the pedigree method.

Inferential Statistical Analysis: For this purpose, only GBL was analyzed. The 9 level response variable was transformed into a binary variable, whose 2 levels were: "Resistant', with GBL ≤ 4, and 'non-resistant', with GBL ≥ 5. An Stratified Analysis with Cochran-Mantel-Haenszel (CMH) statistic was performed to test the presence of a significant association between 'method' and 'resistance stability' across the 3 cross types. This analysis tested whether the observed favorable effect of PM on generating higher proportions of stable resistant lines in the first two cross-types --JxJ - SxR and JxI - SxS-- was statistically significant and was generalizable across the three cross-types. As the CMH statistic was highly significant (table 4), a second analysis was performed. A Logit Model to test the effect of 'method', 'cross type' and their interaction on the proportion of stable resistant lines generated.

Results. the Logit Analysis (Table 5) shows a highly significant difference between methods (in favor of PM) in the proportion of stable resistant lines generated (23.9% for PM, 7.8% for AC-F_2 and 19.4% for AC-R_2; a non-significant diffeerence between cross types in their capacity to generate stable resistant lines (28% for JxJ - SxR, 35% for JxI - SxS and 33% for JxI - SxR) and absence of interaction between method and cross type. This analysis also confirmed the hypothesis that blast susceptible parents (JxI - SxS crosses) could produce stable resistant progeny.

Table 5: Effect of "Method" and "Cross type" on stable resistance.

- Logit Model-

$$\log(\pi_{SR}/1-\pi_{SR}) = \mu + \text{Method} + \text{Cross type} + \text{MxC}$$

Source	df	Wald Chi-square statistic	prob.
Intercept	1	28.9	0.0001
Method	2	6.8	0.03
Cross type	2	9.9	0.07
Method x cross type	4	0.1	0.98

Table 2. **Overall performance of populations generated by each method across crosses in relation to Leaf Blast Reaction (LB), Neck Blast Reaction (NB) and General Blast Reaction (GB).**

	METHOD								
	Pedigree			AC-F2			AC-R2		
Parameter	**LB**	**NB**	**GB**	**LB**	**NB**	**GB**	**LB**	**NB**	**GB**
1. Initial no. of lines generated	17867[1]	-	17867	441	372	441	740	485	740
2. Descriptive statistics for initial population									
- Mean Score	5.96	-	5.96	5.4	5.8	6.6	7.2	4.3	7.3
- Skewness	-	-	-	-0.09	-0.02	-0.22	-0.91	0.10	-0.92
- Median	6	-	6	5	6	7	8	5	8
- Mode	6	-	6	4	9	7	9	7	9
- Range	0-9	-	0-9	1-9	1-9	2-9	1-9	1-9	1-9
- no. (and %) of resistant lines	3491 (19.5)	-	3491 (19.5)	14 (32.7)	114 (30.6)	64 (14.5)	85 (11.5)	228 (47.0)	67 (9.1)
3. No. of resistant lines at the begining of evaluation cycle for stable resistance	926[2]	724[2]	681[2]	144	114	64	85	228	67
4. No. (and %) of lines with stable resistance	192 (20.7)	394 (54.4)	163 (23.9)	13 (9.0)	46 (40.4)	5 (7.8)	25 (29.4)	65 (28.5)	13 (19.4)

[1] F2 lines

[2] F4 lines

Table 3. **Comparison between Pedigree Method (PM) and Anther Culture (AC-F2 and AC-R2) in the generation of stable resistant lines in 11 crosses [1]**

	Pedigree			AC-F2			AC-R2		
Cross no.	**N[2]**	**R[3]**	**SR (and %)**	**N**	**R**	**SR (and %)**	**N**	**R**	**SR (and %)**
2	1291	170	79 (46.5)	207	38	2 (5.3)	42	6	1 (16.7)
1	1922	34	15 (44.1)	42	5	- (0.0)	2	-	- (0.0)
3	1899	125	51 (40.8)	47	8	- (0.0)	55	12	6 (50.0)
11	2057	24	8 (33.3)	3	1	- (0.0)	30	2	- (0.0)
4	1404	6	1 (16.7)	104	5	1 (20.0)	101	7	4 (57.1)
10	1465	74	6 (8.1)	-	-	- (0.0)	238	10	1 (10.0)
7	1471	61	2 (3.3)	27	7	2 (28.6)	168	22	- (0.0)
9	1583	95	1 (1.1)	1	-	- (0.0)	30	-	- (0.0)
8	1480	50	- (0.0)	-	-	- (0.0)	42	6	1 (16.7)
6	1565	42	- (0.0)	-	-	- (0.0)	16	2	- (0.0)
5	1780	-	- (0.0)	10	-	- (0.0)	16	-	- (0.0)
Total	17867	681	163 (23.9)	441	64	5 (7.8)	740	65	13 (19.4)

CMH statistic for SR, 2df = 102.6 (p=0.0001)

N = Initial no. of lines generated
R = Number of resistant lines at the beginning of the evaluation cycle for stable resistance
SR = No. (and %) of stable resistant lines

[1] Variable reported: General Blast reaction (GB)

[2] F2 lines

[3] F4 lines resistant to blast

Table 4: Pedigree Method vs. Anther Culture in the generation of stable resistant lines

- Stratified Analysis Results-

	JxJ,SxR (4 crosses)			JxI,SxS (2 Crosses)			JxI, SxR (1 Cross)		
Method	**N**	**R**	**Stable R**	**N**	**R**	**Stable R**	**N**	**R**	**Stable R**
PM	6126	430	138 (.32)	3979	58	23 (.40)	1404	6	1 (.17)
AC-F_2	281	53	4 (.08)	45	6	- (.00)	104	5	1 (.20)
AC-R_2	366	50	8 (.16)	32	2	- (.00)	101	7	4 (.57)
Total	6773	533	150 (.28)	4056	66	23 (.35)	1609	18	6 (.33)

Pearson X^2	18.1 (p = 0.001)	4.9 (p = .09)	2.9 (p = .23)
CMH statistic	20.4 (p = .0001)		

[1] Four crosses were eliminated from the analysis, as they did not produce stable resistant lines. They were: 5 = Fanny/Tetep, 6 = Fanny/056, 8 = Fanny/Carreon, 9 = Fanny/IAC165

Case Study 2:

A methodology to determine the minimum evaluation period for disease-resistance characterization in rice

E. Guimaraes, M.C. Amézquita, G. Lema and F. Correa

This study which started during 1991 was completed this year.
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

The analysis has two main objectives:

a) Assuming seven seasons of continuous evaluations to be the most reliable experimental period length to characterize and select rice varieties for their stable resistance to rice leaf blast, the analysis aims to find out whether a **shorter** period of continuous evaluations would produce the same set of selected material, or would at least exhibit a high percentage of coincidence in selection and a low number of misclassified entries.

b) To illustrate that different (and possible wrong) conclusions might be reached when treating the ordinal scale as a continuous variable.

For data analysis purposes in phase a, the disease-reactions on te 0-9 scale, were converted into 'disease-severity' according to the Standard Evaluation System for Rice (IRRI, 1988). Disease-pressure at SREE was estimated for each growing season (or semester), as the mean 'disease-severity' over the 70 varieties tested. The 70 varieties were characterized by their mean disease-severity (M) and by their response to increased levels of disease-pressure (b) using the 7-semester evaluations. Statistical comparison of varietal means and a test of homogeneity of slopes (b'_s) were performed using the model illustrated in table 2. In order to correct for lack of normally, 'disease-severity' (Y) was transformed into Y^T using the Box and Cox transformation for ratios (Johnson and Wichern, 1982).

$$Y_i^T = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda} & , \text{if } \lambda \neq 0 \\ \ln Y_i & , \text{if } \lambda = 0 \end{cases}$$

where λ value for this particular problem was estimated as $\lambda = 0.4$

Based on the 7-semester evaluations, a group of varieties was selected for their stable resistance. This group, the 'ideal' selections, included varieties with low mean disease-reaction and lack of response to increased levels of disease-pressure. That is, varieties whose M was not statistically different from that of the top variety, using the Waller-Ducan LSD test for mean comparisons, and whose b was not statistically different from 0. The same analytical procedure was applied to data sets simulating shorter **continuous** experimental period lengths. Two 6-semester periods, three 5-semester periods, four 4-semester periods and five 3-semester periods were simulated. For each one of the fourteen cases, a set of selected varieties was produced. Each set was compared to the 'ideal' set of selections produced by the 7-semester data analysis. The decision on the minimum number of continuous evaluations required to select promising varieties for their stable resistance was achieved based on the percentage of coincidence in selection, and the number of misclassified entries, when compared to the 'ideal' set.

In order to achieve objective b, the same data analysis methodology to identify the 'ideal set of selections' previously applied to 'disease-severity' was applied to the 0-9 scores using as dataset the 7-semester period length. Shorter simulated periods were not analyzed using 0-9 scores. The resulting set of selected varieties was compared with the previously identified *'ideal set of selections'* when using 'disease-severity' as the response variable.

Results:

Table 1 confirms the high, but variable levels of disease-pressure at SREE during the 4-year period considered. This supports the use of this 'hot spot' as experimental site to characterize and select rice varieties for their stable resistance to leaf blast, as was pointed out by Correa-Victoria and Zeigler (1992b).

As a result from the analysis on 'disease-severity', a group of 18 varieties, out of the 70 varieties tested, was selected as promising parental material in terms of their stable resistance to leaf blast disease (table 2). They are characterized by low mean disease-severity (M) --ranging between 4.71 and 20.43, when the overall mean was 40.8 and the maximum was 100.0--, and lack of response to increased levels of disease-pressure (b'_s not different from 0).

Selections resulting from shorter experimental period lengths were compared with this 'ideal' set of 18 promising material presented in table 2. Criteria used for comparison were coincidence in selection and number of misclassified entries. Resulting values are presented in table 3. These results indicate that 6 or even 5 continuous growing seasons would produce the same selections. However, when shorter periods are considered, the number of misclassified entries is high. Correlations between M's and between b'_s also support the decision of 5 semesters being an appropriate period length.

Selections resulting from the analysis of disease-reaction scores, expressed in a 0-9 ordinal scale, show only an 33.3% coincidence and a high number of misclassified entries, when compared to the set of 18 'ideal' selections resulting from the analysis of 'disease-severity'. This indicates that misleading results and conclusions can be drawn when appropriate statistical methodology is not applied.

Conclusions: The study allows the following conclusions:

a) The minimum evaluation period length to characterize and select rice varieties by their **stable resistance to leaf blast** is 5 semesters. This guarantees confidence in a proper selection of promising material with a more efficient use of research resources.

b) Very different, and possibly misleading results (selections) were attained when analyzing disease-reaction scores (in the 0-9 ordinal scale) as a continuous variable.

Table 1: Disease-pressure levels for leaf blast evaluation at Santa Rosa Exp. Station, Colombia, during a four-year period (1987 - 1990)

Year/Semester	Leaf Blast pressure (Mean disease severity)[1]	Range in varietal disease severity (min, Max)
1987 A	51.5	(6.0, 94.0)
1987 B	48.8	(6.0, 94.0)
1988 A	58.7	(0.0, 100.0)
1989 A	42.0	(6.0, 87.0)
1989 B	24.8	(3.0, 100.0)
1990 A	48.9	(3.0, 87.0)
1990 B	7.6	(3.0, 50.0)
Mean	40.8	
Standard deviation	32.1	
Mean standard error	1.4	
CV (%)	78.7	

[1] Expressed as the mean 'disease-severity' per semester over the 70 varieties tested.

Table 2: Group of selected varieties by their stable resistance to leaf blast,based on a 7-semester evaluation period (the most reliable)

Variety	Mean disease severity (M)[1]	Response to disease-pressure (b)	Standard error of b (S_b)	Prob of significance of T statistic (for $H_0 : b = 0$)
1. Amistad 82	4.71	0.061	0.4	0.879
2. Ceysvoni	4.71	0.093	0.4	0.879
3. Panamá 1537	5.50	0.100	0.4	0.821
4. Araure 2	6.00	0.100	0.4	0.801
5. IR 58	6.86	0.093	0.4	0.816
6. Panamá 1048	8.67	0.308	0.4	0.452
7. Centa A1	9.00	0.169	0.4	0.673
8. Dawn	9.67	0.313	0.4	0.436
9. Eloni	9.67	0.313	0.4	0.436
10. Colombia 1	10.43	0.349	0.4	0.384
11. Juma 58	11.33	0.363	0.4	0.376
12. Iniap 415	11.43	0.229	0.4	0.568
13. Tanaioka	12.71	0.358	0.4	0.373
14. Juma 62	13.57	0.387	0.4	0.334
15. IR 43	15.00	0.346	0.4	0.388
16. Iniap 7	16.29	0.315	0.4	0.432
17. Ciwini	16.50	0.521	0.4	0.193
18. Araure 4	20.43	0.596	0.4	0.138

[1] LSD Walter Duncan value on the transformed variable (=4.34), indicates no significant differences between M's for this group of varieties

Table 3: Coincidence in varietal selection, between the 7-semester evaluation period (the most reliable and shorter periods (using 'disease severity' as response variable

Period length/combination	Number of selected varieties	Coincidence in selection (%)	Misclassified entries	M's range for selected varities		b's range on selected varieties	
				(Min.)	(Max.)	(Min.)	(Max.)
7-semester	18	-	-	4.7	20.4	0.06	0.59
6-semester/combination 1	17	93	1	5.0	19.2	0.04	0.71
/combination 2	23	100	5	4.5	18.6	0.03	0.99
5-semester/combination 1	19	100	1	5.4	28.6	0.07	0.93
/combination 2	17	94	1	4.8	18.0	0.01	0.58
/combination 3	19	100	1	4.2	13.6	0.02	0.43
4-semester/combination 1	27	100	9	6.0	53.0	0.02	2.43
/combination 2	19	100	1	5.3	26.0	0.01	0.86
/combination 3	19	100	1	4.5	16.3	0.01	0.60
/combination 4	28	100	10	3.0	17.8	0.02	0.64
3-semester/combination 1	36	100	18	6.0	66.0	-3.28	3.43
/combination 2	22	100	4	6.0	43.6	-2.47	2.22
/combination 3	22	100	8	5.0	35.0	0.03	1.25
/combination 4	26	100	13	3.0	19.7	-0.10	0.54
/combination 5	31	100		3.0	34.3	0.00	0.92

Case Study 3

A mathematical model to describe Diffusion Patterns of Rice Commercial varieties in Colombia and Brazil during the period 1971-1990

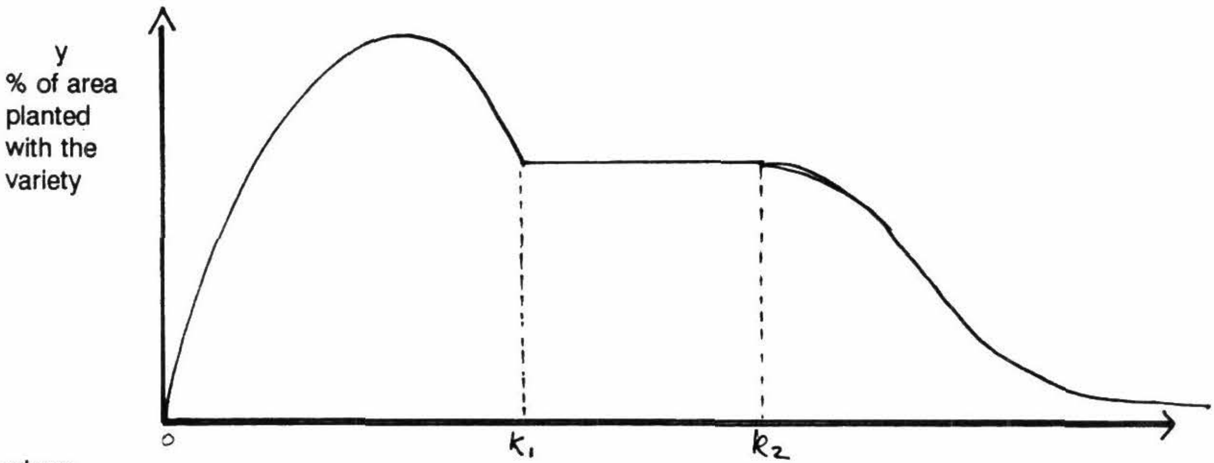
F. Cuevas, M.C. Amézquita, G. Lema
(in progress)

A tree-phased *segmented regression* model was found appropriate to describe the observed patterns in share of area planted by 9 individual irrigated rice varieties, commercially released in Colombia and Brazil during a 10-year period: 1971-1981 and observed from the year of their release up to 1990.

The cycle starts with a parabolic phase followed by stability and ends with a period of decrease.

The mathematical model used was as follows:

$$y = \begin{cases} a + b(x - k_1) + c(x - k_1)^2 & \text{, if } x \le k_1 \\ a & \text{, if } k_1 < x \le k_2 \\ ae^{-k(x-k_2)} & \text{, if } x > k_2 \end{cases}$$



where,

y	% of total area planted by the individual variety
x	time (in years) from varietal release
a,b,c	represent the parameters of the parabola
k_1	indicates the time (in years) from release to stabilization
k_2	indicates the time (in years) from release to the beginning of the decline period
k	represents the rate of decreases during the decline period

The model parameters served to calculate two other useful indicator of varietal performance.

y_{max} = maximum % of area planted with the variety, and

x_{max} = time (in years from varietal release) to reach that stage of maximum adoption

$$x_{max} = k_1 - b/2c$$

$$y_{max} = a - b^2/4c$$

The NLIN procedure of SAS version 6 was used to estimate the model parameters. Convergence was achieved in all cases.

Five of the six Colombian varieties completed the diffusion cycle, while all the 3 from Brazil only reached the stability phase. The general pattern suggests an early interest in new varieties followed by the identification of the niches under which commercial potential is maximized. The final stage --the decrease phase-- indicates that improved varieties with similar adaptation became available. The observed dynamism in varietal use could be maintained with the commercial exploitation of a new germplasm base.

Case Study 4

A categorical data analysis model to quantify progress in blast resistance through recurrent selection

E. Guimaraes, M.C. Amézquita, G. Lema
(in progress)

A Logit Model with orthogonal contrasts for the analysis of categorical response variables, was utilized to quantify progress made in blast resistance between two cycles of recurrent selection.

This research deals with the construction of a gene pool aiming at improving resistance to blast. 30 rice lines of diverse origin were selected as parents because of their stability of reaction to a range of leaf and panicle blast races. This group of parents represents the first selection cycle (C_o) and is called the C_oP_o population. As a product of combining the 30 parents, 417 double crosses (second selection cycle) were obtained. Three populations were generated from them according to three different selection criteria, resulting in C_1P_1, C_1P_2, and C_1P_3 populations, as follows:

C_1P_1 : 58 parents with blast reaction score ≤ 3
C_1P_1 : 51 parents with blast reaction score $= 4$ or 5
C_1P_1 : 58 parents with blast reaction score ≤ 5

Each group of lines, selected as parents for the next cycle (C_2) will in turn be combined to generate C_2P_1, C_2P_2 and C_3P_3 populations respectively.

We here report on the progress made between cycle o and cycle 1. Most important comparisons are C_1P_1 vs. C_o, C_1P_2 vs. C_o and C_1P_3 vs. C_o.

For data analysis purposes both categorical response variables of interest --Leaf Blast reaction (LBL) and Neck Blast reaction (NBL)-- measured on a 1-9 scale, were transformed into a 3-level categorical variable whose levels were 'Resistant' (R), when blast reaction score ≤ 3, 'Intermediate' (I), when score = 4 or 5, and 'susceptible' (S), when blast reaction score ≥ 6. As the central concern was to quantify progress between C_o and C_1 populations, the logit model used (shown below), splitted-up the *source of variation* 'Population', with 3 degrees of freedom, into 6 orthogonal contrasts: C_1P_1 vs. C_o, C_1P_2 vs. C_o and C_1P_3 vs. C_o, as follows:

Logit model:

$$\begin{bmatrix} \log\ (\pi_R/\pi_S) \\ \\ \log\ (\pi_I/\pi_S) \end{bmatrix} = \vec{\mu} + \overrightarrow{\text{Block}} + \overrightarrow{\text{Population}} + \vec{\text{B}} \times \text{P}$$

Source of variation	df	LBI		NBI	
		Wald Statistic (Chi-square distributed)	prob	Wald Statistic (Chi-square distributed)	prob
Intercept	2	85.0	0.00001	211.6	0.00001
Block (B)	4	3.0	0.4629	2.2	0.6945
Polulation (P)	6	111.5	0.00001	44.2	0.00001
C_1P_1 vs. C_o	2	25.0	0.00001	27.3	0.00001
C_1P_2 vs. C_o	2	19.3	0.0001	0.2	0.9025
C_1P_3 vs. C_o	2	10.2	0.0061	13.9	0.0009
B x P	12	4.3	0.9779	9.5	0.6565

Results show a significant progress in blast resistance between the first two selection cycles, indicating significant increase in the proportion of resistant lines between C_1P_1 and C_1P_3 vs. C_o.

2.2 Collaboration with the Tropical Forages Program

Support to RIEPT (Red Internacional de Evaluación de Pasturas Tropicales) in the management and statistical analysis of its information:

Since its creation in 1979, the RIEPT assigned the CIAT's Tropical Pastures Program --and now the CIAT's Tropical Forages Program-- the responsibility to centralize and make available to network members all the information generated by the network. Since then, the CIAT's Biometry Unit has collaborated very closely with the Program, in the organization, storage and statistical analysis (by site, by country, by ecosystem or across-ecosystem data analysis) of RIEPT-generated research results. Up to now, 251 agronomic-trials (ERA and ERB) and some 10 grazing trials have been statistically analyzed and their results stored in the RIEPT database. Multilocational analysis to identify promising germplasm by agro-ecosystem have been performed using RIEPT information generated between 1979 and 1993. Recent studies include: "The analysis of germplasm evaluated in the Humid Tropics" presented at the 1990 RIEPT meeting in Pucallpa, Perú in November 1990 and published in its memories; "The analysis of forage germplasm for Central America" published in the document "RIEPT results for Central America in 1991"; "The analysis of germplasm evaluated in the savanna ecosystem" presented at the 1992 RIEPT meeting in Brasilia, Brazil (Nov. 22-26), and "Use of information generated by RIEPT: 1979-1992" presented at the 1992 RIEPT meeting in Brasilia. During this meeting, the CIAT Biometry Unit presented an overview talk on the use of the information generated by RIEPT (1979-1992) and offered to RIEPT members the microcomputer version of the RIEPT database, (for a more detailed description refer to the section entitled the RIEPT database in the Biometry Unit. (Annual Report 1992)

Support to AFRNET (African Network for forage evaluation in the management and statistical analysis of its information)

AFRNET was created in 1988 as a collaborative research effort between IEMVT of CIRAD[1], ILCA[2] and CIAT[3] to conduct adaptative research on forage species for cattle feeding in the humid and sub-humid areas of West and Central Africa. National agricultural research institutions from 11 countries were invited to participate and 15 trials were initially established in these countries with seed received from CIAT (table 0). IEMVT finances the project and is responsible for its overall coordination; CIAT provides forage germplasm, provides definition of experimental design/implementation, scientific monitoring of the trials and is also responsible for the management

[1] IEMVT = Institute de l'Evage et Medicine Veterinaire Tropicale, France. IEMVT is one of the research Centers of CIRAD (Centre Internationale pour la Recherche Agricole et Development), for France and its ex-colonial territories.

[2] ILCA = International Livestock Centre for Africa, Addis Ababa, Ethiopia

[3] CIAT = International Center for Tropical Agriculture, Cali, Colombia

and statistical analysis of the information generated by AFRNET; ILCA provides scientific guidance. The "project identification meeting" took place in Togo in November 1989; the first annual workshop took place in Togo, in April 1990, to clearly identify experimental sites; the second annual workshop was conducted in April 1991 "to balance and reactivate the project". The third annual workshop took place in Bouaké, Ivory Coast in March 1992 in which partial results were presented (de Fabregues, 1991). This 4rd annual meeting took place in Bamako, Mali, from March 29 to April 2, 1993. During that meeting results attained during the 2-year period April 1991 - March 1993 were analyzed.

As part of CIAT's contribution to AFRNET's 4rd annual meeting, a collaborative paper between the CIAT's Biometry Unit and the CIAT's Tropical Forages Program, entitled "Analysis of performance of herbaceous and woody forage species in Central and West Africa", was presented. We would like to present here a brief summary.

Analysis of performance of herbaceous and woody forage species in Central and West Africa

M.C. Amézquita, C. Lascano, G. Ramírez, L.H. Franco (1993)

This paper presents a methodology for the across locations statistical analysis of information generated by the Network. Although 7 sites, out of the initial 15, had reported information to CIAT, only 4 of them had reported information on all experimental periods: establishment, biomass production during maximum rainfall and biomass production during minimum rainfall. They are Kurmin Bire (Nigeria), Bouaké (Ivory Coast), Avetonou (Togo) and Kovie (Togo). Information generated by these four sites was used as data source for the present study. They represent humid or subhumid savanna environments, with a relatively long dry season. Ecotypes considered for the analysis included 8 grasses, 21 herbaceous legumes and 6 tree legumes. Out of them 8, 16 and 3 respectively were evaluated by the four experimental sites -and were therefore included in the across location analysis. Independent analysis for grasses, herbaceous legumes and tree legumes were performed.

Statistical Analysis Methodology

The following indicators were selected to best characterize performance of an ecotype.

For herbaceous grasses and legumes

- Establishment indicators
  1. % cover at 12 weeks
  2. Ratio of $\frac{\%\text{ cover at 4 weeks}}{\%\text{ cover at 12 weeks}}$, indicating rapidity of establishment

- Production indicators
  3. Dry Matter (DM) at 12 weeks during maximum rainfall (kg/ha)
  4. DM-12 weeks during minimum rainfall (kg/ha)
  5. Dry-rainy season ratio in terms of DM-12 weeks

For tree legumes

- Establishment indicators
  1. Plant height at 12 weeks (cm)
  2. Ratio of $\frac{\text{plant height at 4 weeks}}{\text{plant height at 12 weeks}}$

- Production indicators

  3,4,5 as for herbaceous species above.
  6. Plant height at 12 weeks during maximum rainfall (cm)
  7. Plant height at 12 weeks during minimum rainfall (cm)
  8. Dry-rainy season ratio in terms of plant height at 12 weeks.

As significant correlations were found between some indicators, suggesting therefore the need for a reduction in the number of response variables. For this purpose, a Principal Component Analysis on establishment and production indicators -previously standardized to 0 mean and variance of 1 - was performed. A reduced number of principal components explaining a high percentage of the total variance, was selected as the new set of response variables. As a 'principal component' is a linear combination of the original variables, and normally distributed, the selected principal components were interpreted and analyzed through analysis of variance, under the model shown below:

	df		
Sources of variation	**Grasses**	**Herbaceous legumes**	**Tree legumes**
Location	3	2	2
Rep (Location)	8	6	6
Ecotype	7	15	2
Ecotype x location	21	30	4
Error	56	90	12
TOTAL	95	143	26

As the dataset presented unequal subclass numbers, the analysis of variance for each principal component was performed using the PROC GLM of SAS (version 6,07), using SS type III. Least-square means, to adjust for missing values, were reported instead of arithmetic means. To facilitate a visual interpretation of ecotype performance, the ecotypes were placed in a two-dimensional graph, where axes corresponded to the first two resulting principal components.

Results

Table 1 shows overall performance of indicators. Tables 2, 3 and figure 1 illustrate --for the case of grasses-- results from the Principal Components Analysis on indicators (table 2) ANOVA on resulting principal components (Table 3) and classification of ecotypes accroding to the first two principal components (Figure 1).

The first two principal components explained 78% of the total variance and were therefore selected as the new set of response variables. By observing their score, the first principal component, explaining 44%, can be interpreted as 'high dry matter production potential'. The second one, explaining 34%, can be interpreted as 'high dry-rainy season ratio, with low cover % during establishment' (table 2). Least-square means per ecotype for the first two principal components as well as for original indicators were presented. Significant differences between ecotypes were detected in both principal components. (See table 3). When the 8 grass ecotypes are classified in terms of both principal components, the best ecotypes--those with high DM production potential and high dry-rainy season ratio-- can be easily identified. They are (see figure 1) *B. dictyoneura* 6133, *B. brizantha* 26646, *B. decumbens* 606, and *P. maximum* 673. Figures 3 to 6 show DM production performance during maximum and minimum rainfall periods, at each individual location, of an outstanding ecotype -*B. brizantha* 26646- vs. a non-adapted ecotype -*P. maximum* 16031-.

Another selected example of collaborative studies with the Tropical Forages Program is summarized below.

Case Study

A data analysis methodology for the evaluation of large germplasm collections Case study: Evaluation of the CIAT Brachiaria collection in Brazil

Cacilda do Valle [4] and M.C. Amézquita
(Presented at the XVII International Grassland Congress in New Zealand, March 1993)

This study, initiated during 1991, concluded this year. It used as data source 3-year experimental results of the agronomic evaluation of 194 accessions of *Brachiaria* species, carried-out by EMBRAPA, in Campo Grande, Brazil in small plots, under a split-plot design. During these 3 years 18 evaluations were performed: 14 during the rainy season and 4 during the dry season. Let us present a short summary.

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 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

[4] EMBRAPA researcher. CIAT Visiting Scientists during 1991

evaluation of a large germplasm collection. Biomass production (total, leaf, stem) and regrowth capacity were periodically measured. Additionally, observations on resistance to insects Spittle bug, diseases, and plant vigor were made periodically . Early flowering capacity was recorded only once during the experimental period.

Methodology: A set of eleven 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})
11 "Index of Spittle bug resistance", calculated as the percentage of a score ((0='the plant was resistant' 1='the plant was not resistant') assigned to a given accession among the 14 rainy season scores.

A Factor Analysis, with varimax rotation method, was applied to these 11 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 three 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) Annual accumulated leaf dry matter (kg/ha/year); b) Leaf-Stem relation, during the dry season (%); and c) Index of resistance to spittle bug (expressed as a % of **zeros** among 14 evaluations).

The Cluster Analysis helped identify 22 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).

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

OVERALL DESCRIPTIVE STATISTICS

Specie	No. of accessions	Accumulated Leaf Dry Matter (kg/ha/year)	Leaf-Stem relation during the dry season	Index of resistance to Spittle bug (% of zero score among 14 wet season eval.)
B. brizantha	96	9324	1.44	63.2
B. decumbens	35	4229	0.81	67.9
B. humidicola	21	5843	0.93	76.9
B. jubata	11	4085	1.19	67.5
B. ruzisiensis	20	3809	1.43	55.0
B. arrecta	6	2096	0.58	77.4
B. dyctioneura	2	9391	-	82.2
B. negropedata	1	4004	-	78.6
B. adspersa	1	2743	0.7	100.0
Total	193	5058	1.0	74.3

Table 2: Multivariate evaluation of the CIAT *Brachiaria* collection (193 accessions) in Campo Grande, Brazil for a 3-years period.

PROMISING ACCESSIONS[5]

Accession Identification		Accumulated Leaf Dry Matter	Leaf-Steam relation during the dry season	Index of resistance to Spittle bug
CIAT #	EMBRAPA #	(Kg/Ha/year)		(% of zero score among 14 wet season evaluation)
B. brizantha				
16288	B132*	19234	1.36	85.7
16467	B166*	17021	1.18	71.4
16306	B138*	16542	1.10	85.7
16316	B144*	14971	1.48	71.4
16473	B89 *	14268	1.06	71.4
	B163	13864	1.19	57.1
	B73	13823	1.41	57.1
	B65	11838	1.65	50.0
	B52	10648	1.51	42.9
	B51	10351	1.58	42.9
	B137	10289	1.41	42.9
	B59	10252	1.30	35.7
	B136	10127	1.35	42.9
B. decumbens				
16488	D1*	12892	1.02	57.1
606	D62	9157	1.32	50.0
6699	D70	8206	1.46	71.4
B. humidicola				
16886	H13	8226	1.40	78.6
26155	H18*	8027	1.19	85.7
	H25	7318	1.18	85.7
B. juvata				
26237	J13*	7325	1.10	64.3
	J3	4748	1.51	57.1
B. ruzisiensis				
	R103	5105	2.67	42.9

[3] High Annual Leaf Dry Matter, High Leaf-Steams relation during the dry season and low or medium incidence of spittle bug.

* Out of these 22 accessions, the 8 accessions with an * were identified to advance for grazing studies.

2.3 Collaboration with the Cassava Program

Case Study 1:

A Methodology for the Statistical Analysis of electrophoretic patterns.
Case: Biochemical differentiation of mite populations.
Amblyseius limonicus Garman and Mc. Gregor
(Acarina: Phytoseiidae)

M.C. Duque, M.E. Cuéllar, A. Braun; 1993

This study started during 1991 and was completed this year.

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.

In order to verify the results, two experiments were conducted: a) the first to evaluated whether the dietary composition of the various populations of *A. limonicus* was similar, and b) the second one, to evaluate whether their reproductive performance.

Case Study 2

The use of Stepwise Regression, Factor Analysis and Heritability coefficients of various traits in the definition of a Selection Index for cassava in four different agroecozons

E. Mesa, C. Iglesias
(in progress)

Objetivo: Determinar índice de selección para cuatro zonas de evaluación en Colombia: Palmira, Costa Atlántica, Popayán y Llanos Orientales.

Información: Se consideró correspondiente a ensayos de campo de observación (CO), preliminares de rendimiento (EPR) y de rendimiento (ER) para Palmira, Media Luna, Carimagua, La Libertad y Popayán, durante los años 1980 a 1990.

Variables de respuesta:

Producción de raíces (Ton ha)
Número de raíces comerciales
Indice de cosecha
Indice de ramificación
Altura de planta (cms)
Longitud de tallo (cms)
Contenido de materia seca (%)
Contenido de cianuro (escala de 1 a 5)

Metodología:

a) Determinar variables a incluir en el índice de selección para maximizar rendimiento de materia seca de yuca. Se logró mediante:
   1) Análisis de regresión por pasos (stepwise) de rendimiento en función de las restantes variables, el cual introduce caracteres a la ecuación de Regresión Múltiple en el orden en el cual ellas contribuyen al rendimiento de materia seca.
   2) Análisis de factores que complementa la información anterior con base en el análisis de la estructura de covarianza.

b) Estimación del coeficiente de Heredabilidad. Se logró mediante lo siguiente:
   1) Para cada zona, se hizo análisis de regresión utilizando el siguiente modelo:

$$y(t) = B_0 + B_1 \times y(t-1) + e(t)$$
$$t = 1981, ..., 1990$$

La estimación del coeficiente de regresión corresponde a h-cuadrado, coeficiente de heredabilidad.

c) Cálculo de índices de selección, considerando el Indice base Modificado (Smith et all, 1981)

Este índice pondera los valores fenotípicos por la heredabilidad estimada para cada carácter, además de los pesos económicos relativos. Para cada carácter la ponderación es de la forma: $W_i = a_i \times H_i^2$, donde H_i^2 es la heredabilidad asociada a la variable i; luego el índice se calcula de la forma:

$$I = W_1 \times P_1 + W_2 \times P_2 + ... + W_n \times P_n$$

Una vez determinadas las variables a considerar en el índice de selección, éste se calculará utilizando el Indice Base Modificado de Smith, y luego se compara con la selección resultante del Análisis de Factores y de Componentes principales. A partir de la matriz de correlaciones se hizo

Análisis de factores sin rotación, análisis de factores con rotación (VARIMAX) y análisis de componentes principales; para cada uno de los métodos, se calcularon coeficientes de los "scores" para cada variable con el vector de coeficientes correspondientes a cada factor en el modelo específico. Estos coeficientes fueron sumados a través de los factores, resultando en un vector de suma de coeficientes de "scores", lo cual es factible por la ortogonalidad de los mismos. Finalmente estos "scores" se utilizan de igual forma que los "scores" generados con el índice base modificado, calculado a partir del coeficiente de heredabilidad. Las alternativas propuestas (métodos multivariados) con comparadas con el índice base modificado con base en el coeficiente de correlación de Spearman y el número de materiales comunes seleccionados por cada uno de los métodos.

Se consideró que las variables que maximizan el rendimiento de materia seca son: índice de cosecha, número de raíces comerciales, índice de ramificación, altura de planta y longitud de tallo. (Estas variables fueron las consideradas para el cálculo del índice).

Después de calcular los "scores" con base en el índice de selección (Indice Base Modificado) y con los "scores" obtenidos del análisis multivariado, éstos se correlacionaron con el Rendimiento de materia seca. Se observa el rendimiento de materia seca altamente correlacionado con los "scores" obtenidos del análisis multivariado.

En cuanto al número de materiales comunes seleccionados en los primeros 20, el análisis de factores con o sin rotación seleccionan el mayor número de materiales en común con el índice base modificado calculado con base en el coeficiente de heredabilidad.

Case Study 3

A Logistic Regression Model on a 3-level categorical variable. Case: Cassava variety 'Venezolana' adoption in the North Coast of Colombia

M.V. Gottret, G. Henry, M.C. Duque, 1993

Economic studies of technology adoption very often built production functions which are regression models on a continuous dependent variable (yield), regressed against a set of explanatory variables of diverse nature. However, there are circumstances in which the dependent variable cannot be continuous but categorical, being either a binary variable, such as 'presence or absence of adoption' of a given technology, or a multi-level categorical response. In these cases, a logistic regression model is appropriate. A logistic regression model is a regression model in which the dependent variable is categorical, while explanatory variables can be a mixture of continuous and class variables.

The general form for a logistic regression model on a binary variable is

$$\text{Log}\,(\pi/1-\pi) = \alpha - \Sigma \beta_i X_i$$

where π = proportion of 'successes'; ie proportion of $1'_s$
$1-\pi$ = proportion of 'failures', 1e proportion of $0'_s$
α = regression intercept
X_i = explanatory variable i
β_i =
regression coefficient associated with explanatory variable X_i

The general form for a logistic regression on a multi-level categorical variable, (K levels say) is a multivariate model of the form

$$\begin{bmatrix} \log(\pi_1/\pi_k) \\ \log(\pi_2/\pi_k) \\ \log(\pi_{k-1}/\pi_k) \end{bmatrix} = \overrightarrow{\alpha} + \overrightarrow{\Sigma \beta_i X_i}$$

where $\pi_1, \pi_2, \ldots \pi_k$ = proportions of responses 1,2,...k
α = vector of intercepts
β_i = vector of regression coefficients
X_i = explanatory variable i

Logistic regression models on a multi-level categorical response are more difficult to interpret than the corresponding model on a binary response. However statistical software now offers the tools for parameter estimation and hypothesis testing on this type of models.

This study is an example of the use of a logistic regression model on a 3-level categorical variable 'adoption of variety "Venezolana"', recorded in a 3-level code: 1) No adoption, 2) partial adoption, 3) total adoption. PROC LOGISTIC of SAS version 6.08 was used for parameter estimation and hypothesis testing.

Data source corresponds to a survey conducted in the North Coast of Colombia, using 544 cassava farmers. 11 explanatory variables were sonsidered in the logistic regression model - 11 continuous and 5 dummy as follows:

Continuous variables:

X_1 = distance to the nearest urban area (km)
X_2 = distance to the nearest drying plant (km)
X_3 = farm size (ha)
X_4 = area planted with cassava (ha)
X_5 = relative importance of cassava (% of land under crops planted with cassava)
X_6 = land tenancy (% land owned by the farmer)
X_7 = formal education (yr)
X_8 = experience (yr planting cassava)
X_9 = age (yr)
X_{10} = family size (no. of family members)
X_{11} = availability labor (No. of family members who work on the farm)

Dummy variables:

X_{12} = topography (0 = flat land)
(1 = rolling land)
X_{13} = credit (0 = farmer recieves no technical assistance for cassava cropping)
(1 = farmer receives credit for cassava cropping)
X_{14} = tech. assistance (0 = farmer receives no technical assistance for cassava cropping)
(1 = farmer receives technical assistance for cassava cropping)
X_{15} = cropping system (0 = intercropped)
(1 = monoculture)
X_{16} = membership in farmer association (0 = farmer is not a member)
(1 = farmer is a member)

This model produced as output the regression coefficients associated with each explanatory variable, from which, probabilities of total and partial adoption of cassava variety 'Venezolana' by diverse type of 'clients' were calculated: Typical farmer, cooperative member, farmer with access to credit, farms on rolling land and farms under cassava monoculture.

Results of this study, conducted by the Cassava Economic Section, appear in the Cassava program Annual Report 1993.

2.4 Collaboration with Bean Program

Case Study 1:

The use of the 'Coefficient of Parentage' to estimate genetic diversity among Andean and Mesoamerican Common Bean Cultivars

Oswaldo Voysest, M.C. Valencia, M.C. Amèzquita (1993)

This study was undertaken to analyze the genetic base of common bean cultivars released in Latin america from the onset of breeding activities in 1934 up to 1992. Each of the 184 cultivars of hybrid origin and their 187 ancestors were classified into one of the six racial groups: Mesoamerica, Nueva Granada, Durango, Chile, Jalisco and Peru. Genetic diversity was assessed, based on the pedigree, through the Coefficient of Parentage.

The coefficient of parentage(r), defined as the probability that a random allele at any locus in one cultivar is identical by descent to a random allele at the same locus in the second cultivar (16), was used to estimate genetic diversity. Actually, r is a measure of the minimum degree of relationship between 2 genotypes since it is based on genes that are identical by descent and does not include genes that identical by state; r=0 if the genotypes have no common ancestors and r=1 if they are identical. The coefficient of parentage was computed for all pairwise combinations of the 184 cultivars from pedigree information. for the calculations of r it was assumed that a cultivar received half of its genes from each parent. It was also assumed that parents used in crosses were homozygous and homogenous. A relationship matrix, numerically estimated using the INBREED procedure of SAS version 5 Supplemental Library, included ancestral parents, surveyed cultivars and all other cultivar intermediate to the parentage cultivars. Mean genetic contributions of ancestral lines in each bean type as well as in the Andean and Mesoamerican gene pools were calculated by computing the average r value of each ancestor with the cultivars making up the different bean groups of types and races. The mean coefficient of parentage(r) among cultivars within each gene pool or country was interpreted as an estimate of genetic diversity.

For the first two racial groups --122 cultivars of race Mesoamerica and 43 cultivars of race Nueva Granada-- both received the majority of their genes (82% and 79% respectively) from ancestors of their same race, and a substantial part of their genetic contribution (53% and 40% respectively) was attributed to no more than 10 ancestors of their same race. For the next two racial groups, the genetic contribution from ancestors of their same race was much as 40% for Durango and 50% for Chile. Races Jalisco and Peru have received little or no improvement and have been used little in the improvement of germplasm of other races as well. This study also showed the recent increased use of interracial hybridizations, indicating the trend to broaden genetic variability of cultivars within the various bean races.

Case Study 2

Identificación de resistencia en el germoplasma de Phaseolus acutifolius al patógeno de la bacteriosis común del frijol.

Dacier Mosquera, M. Pastor-Corrales, M.C. Duque

El banco de germoplasma de frijol contiene 303 entradas de *Phaseolus acutifolius*, en estado cultivado o silvestre, las cuales se quieren analizar por su reacción a Xanthomonas (XCP), agente causal de la bacteriosis común, teniendo en cuenta esta variable de estado.

Básicamente la necesidad de este trabajo se plantea al identificar la resistencia genética como el método más práctico y económico de manejo del problema, pero bajo las siguientes hipótesis:

- En *P. vulgaris* los niveles de resistencia son intermedios o bajos y las fuentes son pocas.
- En *P. acutifolius* los niveles son altos, se cree que la especie es resistente pero las cruzas interespecíficas son difíciles.

Para el ensayo el material experimental se complementó con algunas entradas de las especies *vulgaris* y *tenuifolium*. Se utilizaron diferentes técnicas para inocular la bacteriosis y solo en algunas pruebas se hicieron repeticiones debido a la escasez de semillas. Con los datos se pretende hacer una descripción del material disponible, en lo posible asociando su origen y estado, con la susceptibilidad a XCP.

Se descartó la hipótesis de independencia entre el estado y la reacción a XCP, asociando la mayoría de los resistentes a los cultivados, de los silvestres a susceptibles y de los tenuifolium a intermedios.

Tomando como variables activas la descripción de la respuesta a la enfermedad y como suplementarias las de origen y estado, se realizó un Análisis de Correspondencia Múltiple. A partir de la selección de las 3 primeras componentes se hicieron las gráficas respectivas para variables suplementarias y activas donde se apreciaron grupos cuya diferencia significativa pudo establecerse a partir del análisis de varianza multivariado sobre las nuevas coordenadas de cada entrada (las definidas en el sistema de las 3 componentes principales mencionadas previamente). Estos grupos se generaron de tal manera que se conformó un gradiente completo en la respuesta: "siempre resistentes", "siempre susceptibles", "intermedios" y el gran grupo de los "cambiantes", que puede subdividirse según el sentido del cambio.

Este trabajo fue presentado en el congreso de ASCOLFI, versión 1993.
(Ver Informe Anual)

Case Study 3

Comportamiento de diferentes variedades de frijol frente a aislamientos de P. griseola

Carlos Jara, M. Pastor-Corrales, M.C. Duque

Se pretende fijar la metodología para la evaluación de la reacción de las variedades de frijol a diferentes aislamientos del hongo causante de la mancha angular.

Un grupo de variedades de frijol definido previamente con orígenes andino y meso-americano será sometido al ataque de un conjunto de aislamiento del hongo sobre los cuales se desea obtener información. Se pretende conocer la patogenicidad de los aislamientos en cada variedad y tratar de generalizar al grupo de origen.

Para tal fin, se evalúa cada planta en 6 fechas por sus síntomas de mancha angular, constituyendo ésta la variable de respuesta.

Se define como espacio muestral para este ensayo:

$\Omega = \{ (X_1, ..., X_6) / \text{ Si } i \leq j \quad i, j = 1 ... 6 \quad x_i \leq x_j \}$ x_i = calificación de enfermedad en la evaluación i

Para buscar una estandarización en la clasificación se generan sistemáticamente puntos del espacio muestral los cuales según criterios técnicos de la sección de Patología de frijol permitirán la clasificación de resistente, intermedio o susceptible.

Tomando el anterior como un archivo de calibración se estima una Función Discriminante para ser aplicada sobre los datos obtenidos en invernadero. Los resultados: una tabla de variedades vs. calificación de resistencia son corroborados con los obtenidos a partir de análisis del área bajo la curva de progreso de infección y daño.

Los resultados permitieron identificar aislamientos de baja patogenicidad y variedades con diferentes niveles de resistencia, lo mismo que agrupar los aislamientos en Andinos y Meso-Americanos. El trabajo se desarrolla por fases y está en curso, por lo tanto no hay resultados finales debido al amplio número de aislamientos a evaluar y a la lentitud propia del proceso que se estudia. (Ver Informe Anual Patología de Frijol).

2.5 Collaboration with the Savanna Program

Case Study 1

Descripción de vegetación de sabana nativa en Carimagua

G. Rippstein y E. Mesa
(en proceso)

Objetivo: Hacer una descripción y analisis de la vegetación de sabana nativa en Carimagua.

Metodología de análisis estadístico:
1. Descripción general de la composición de la vegetación por: familia, género y especie.
2. Cálculo de frecuencias (número de veces que se observó la especie en los diferentes conteos realizados), para cada especie en cada sitio (comunidad).
3. Análisis de Correspondencia con el fin de establecer la relación entre las diferentes especies con cada comunidad. Se hizo Análisis de Correspondencia dado que los datos vienen de una tabla de doble entrada (Especie x Comunidad) de dos variables categóricas. El análisis localiza las categorías de la tabla de doble entrada en un espacio Euclidiano mostrando la relación entre la presencia de las especies en cada comunidad (sitio).
4. Agrupación de comunidades con base en características de suelo y de clima (uso de componentes principales).

Case Study 2

Statistical Analysis of reproductive performance in beef cattle under extensive production systems

M.C. Amézquita and R. Vera (1993)

Results from a large grazing experiment conducted in Carimagua research station, eastern Colombian savannas, for over 4 years, with 178 Zebu x criollo cows, were used as data source to test and recommend an integrated statistical methodology for the analysis of beef reproductive performance. As a complementary product or the analysis, the most sensitive indicators of treatment differences were identified. ANOVA and MANOVA were compared in terms of hypothesis testing and estimation in the analysis of continuous variables with repeated measurements. ANOVA on raw data (ANOVA-raw data), a generalized linear model using mean scores (GENLIN-mean scores) and the Stratified Analysis (STRAT) were compared for the analysis of categorical variables. Table 1 shows the overall descriptive statistics for animal performance parameters with 178 cows observed during a 4-year period. Results indicate a close agreement between ANOVA and MANOVA in the first case, although MANOVA showed more powerful in detecting significant effects. Table 2 shows the comparison between MANOVA and ANOVA in the analysis of continuous variables with repeated measurements. In the analysis of categorical variables, ANOVA-raw data and GENLIN-mean scores--theoretically considered the best method--presented very consistent results; STRAT did not work when missing cells were present. Table 3 shows the comparison between ANOVA and LOGIT-mean scores in the analysis of ordinal data. Case: Beef reproductive traits. Sensitive indicators of treatment differences were: 'number of conceptions during lactation/cow', 'number of births/cow', 'number of weaned calves/cow', 'interval between parturitions', 'calf weaning weight', 'total production of weaned calves/cow' and 'total beef production/cow'. The other animal performance indicators were relatively insensitive.

Table 1: Overall descriptive statistics for animal performance parameters. 178 cows observed during a 4-year period.

Animal performance parameter	N[1]	Mean	SD[2]	CV[2] (%)
CONTINUOUS VARIABLES				
Cow liveweight, adjusted at non-lactanting-non pregnant (kg)	3355	325	39.8	12.2
Cow weight at conception (kg)	534	326	38.0	11.6
Interval between parturitions (months)	314	21	5.3	25.5
Cow culling weight (kg)	15	291	38.2	12.8
Calf birth weight (kg)	275 [3]	26	3.1	11.9
Calf weaning weight (kg)	360	143	20.2	14.1
Total production of weaned calves/cow (kg)	178	317	107.1	33.7
Total beef production/cow (kg)	177	640	110.9	17.3
CATEGORICAL VARIABLES	**TOTAL**	**Overall Mean Score/Cow**	**Range/Cow (Min-Max)**	**Mean annual rate (%)**
Number of conceptions	534	3.00	1-5	76
Number of conceptions during lactation	92	0.52	0-3	13
Number of abortions	25	0.14	0-2	3
Number of births	509	2.86	1-5	72
Number of perinathal deaths	49	0.28	0-3	8
Number of weanded calves	360	2.02	0-4	50

[1] N= Number of observations entered in the analysis
[2] Calculated after removing effect of sources of variation
[3] 234 calves did not have information on birth weight

Table 2: Comparison between MANOVA and ANOVA in the analysis of continuous variables with repeated measurements

Source of Variation [1]	(1) Cow liveweight adjusted		(2) Cow liveweight at conception		(3) Interval between parturitions	
	prob (F)[2]	prob(Wilk's Lambda)[3]	prob(F)	prob(Wilk's Lambda)	prob(F)	prob(Wilk's Lambda)
Site (S)	0.001	0.0001	0.0001	0.0171	0.0001	0.0001
Prod. System (P)	0.3221 (ns)[4]	0.0001	ns	ns	0.0092	0.007
S x P	0.2554 (ns)	0.005	ns	ns	0.002	0.001
Year	0.002	0.0001	ns	ns	ns	ns
Year x S	0.001	0.0001	ns	ns	ns	ns
Year x P	0.001	0.0005	ns	ns	ns	ns
Season	0.001	non-applic.	ns	non-applic.	ns	non-applic.
Lactation stage	-	-	0.0529	0.2541	-	-
No. of significant terms 4 MAD [5] = 0.0724 r [6] = -0.002 (prob = 0.95)		6	2	2	3	3

[1] Found significant at 0.10 level

[2] Prob of significance of ANOVA F-test

[3] Prob of significance of MANOVA Wilk's Lambda statistic

[4] ns = non-significant term ($p \geq 0.10$)

[5] Mean Absolute Departure between MANOVA and ANOVA significant levels

[6] Correlation coefficient between MAÑOVA and ANOVA significant levels

Ex 3: Comparison between ANOVA and LOGIT-mean scores in the analysis of ordinal data
Case: Beef reproductive traits[1]

Response variable/effect	prob (ANOVA F-test)	prob (Wald statistic)
1. Conceptions during lactation/cow		
Site	0.001	0.00001
Prod. System	0.0001	0.00001
SxP	0.003	0.00001
2. Abortions/cow		
Site	0.012	0.006
Prod. System	0.033	0.024
SxP	0.4	0.3
3. Births/cow		
Site	0.001	0.001
Prod. System	0.009	0.001
SxP	0.003	0.001
4. Weaned calves/cow		
Site	0.0001	0.00001
Prod. System	0.01	0.006
SxP	0.04	0.03

MAD[2] = 0.011
r = 0.98 (p = 0.001)

[1] Source: Amézquita, M.C. and Vera. R (1993)

[2] Mean absolute departure between LOGIT and ANOVA significant levels.

2.6 Collaboration with the Hillsides Program

Support in the design and analysis of the census of 'Microcuenca del Río Ovejas'

J. Ashby, G. Lema, M.C. Amézquita

In order to characterize one of the major research sites within the hillsides agro-ecosystem -- Microcuenca del Río Ovejas, Departamento del Cauca, Colombia -- the CIAT's Hillsides Program, together with two other institutions (CVC and CETEC), initiated a socio-economic census of a representative community of farmers established in the chosen area. The community consists of 1300 families, farmers by nature, distributed in 21 'veredas'. the census main objective is to characterize the relationship between socio-economic variables and variables quantifying the degradation of the natural resource base such as erosion, deforestation, soil degradation and water quality. Additionally in order to satisfy needs of the participating institutions, the following sub-objectives are addressed: a) Socio-economic description of the family: family composition; age, sex, education and occupation of its members; work type and distribution of responsibilities within the family. b) Farmers participation in farmer's associations. c) Availability and type of agricultural tools access to credits and other financial resources, public services and house conditions for each family. d) Agricultural production and its most important limiting factors. e) Land use patterns, f) Agro-silvo systems and g) type of domestic animals in the farm.

The Biometry Unit collaboration in this project initiated during 1993 and is expected to continue during 1994 in the following topics:

- Formulary design
- Datafiles design
- Methodology for the census statistical data analysis

Formulary design: 11 independent sections conform the formulary, whose content will constitute the 11 independent datafiles for data analysis purposes. Criteria considered for the definition of each formulary section and subsequent datafile we census objectives and information unit, as follows.

	Datafile Identification	Unit of information
1.	Family description	1 member within the family
2.	Work activity within the family	the family
3.	Family participation in farmer's associations	the family
4.	Housing conditions	the family
5.	Public services	the family
6.	Agricultural Production	1 plot within a farm belonging to the family
7.	Limiting factors for agricultural production	the family
8.	Availability of agricultural tools	the family
9.	Access to credit/financial resources	the family
10.	Present land use	1 plot within a farm belonging to the family
11.	Tree cultivation	the family
12.	Presence of domestic animals	the family

Methodology for the Statistical Analysis

At present, 830 families have been surveyed, and 630 formularies have been submitted to initial data processing. the Data analysis phase has not started yet. The methodology for the census statistical data analysis includes: a) A data validation phase, b) a descriptive analysis of the population concerning all the different census objectives, c) an inferential analysis to test hypothetical relationships between land use variables, socio-economic conditions, agricultural technology and degradation of the natural resource base.

3. Training Activities

Basic **training** in statistical methods and data analysis was provided to 24 CIAT research associates/assistants during 1993: 7 from the Cassava Program, 8 from the Rice Program, 2 from Tropical Forages, 3 from Savanna, 3 from Hillsides and 1 from the Bean Program. It is hoped to renew the training activities for National Institution researchers from Latinamerican and African *institutions*, CIAT collaborators. The new Microcomputer Training Laboratory is expected to be used for this purpose. During the five years of existence of the old Laboratory, the Biometry Unit has offered a total of 35 one to two-week training courses, with a total number of 344 National Institution researchers trained from Latinamerica (260), Asia (24) and Africa (50). An approximate number of 105 participation from CIAT research associates/assistants have benefit from this effort during the last 6 years 1987-1992.