TITTLE: GALILEO, an intelligent system to support the experimental design planning in agriculture.

NAMES OF THE AUTHORS: Hércules Antonio do Prado, MSc
Alfredo José Barreto Luiz, MSc.
Marcos Mota do Carmo Costa, PhD
Orfeo Apolo Droguet Affin, BSc

INSTITUTIONS: EMBRAPA - EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA AND FICB - FACULDADES INTEGRADAS DA CATÓLICA DE BRASÍLIA

MAILING ADDRESS:
EMBRAPA/CPAC - Centro de Pesquisa Agropecuária dos Cerrados
KM 18 da BR-020 (Rod. Brasília-Fortaleza)
Caixa Postal 08223
70.301-970 - PLANALTINA - DF

COUNTRY: BRASIL

TELS.: (061) 389-1171 (061) 389-3364

FAX: (061) 389-2953

ELECTRONIC MAIL: alfa@brunb

EVENT FOR PRESENTING:
Simpósio Iberoamericano sobre la Informática en la Agricultura y la Ganadería

KEY-WORDS: expert systems, statistics, methodology, agricultural research, design of experiments.

TITTLE: GALILEO, an intelligent system to support the experimental design planning in agriculture.

Abstract:

In this paper the expert system approach is proposed to assist the agricultural researcher in selecting an appropriate design of experiments. A prototype of this system, named GALILEO, is introduced. A generic session of the system has two steps, the first is the classification of the researcher demand in five possible types: exploratory research, hypothesis test, parameter estimation, comparison of alternatives, and estimation of populational parameters; and the second is the definition of the bettter experimental design, according to the objectives of the researcher, the informations about the environment, and the existence of constraints. The experimental designs considered were completely randomized, randomized blocks, split-plot, and factorial, with theirs combinations; in sampling were considered simple randomized and stratified. An example is given that illustrates an application of the system and future extensions are proposed. It is also presented a little glossary of statistics terms used in the text.

## 1. INTRODUCTION

Expert systems (ES) has had crescent application to agriculture, mainly in crop diseases diagnostics, decision support, and interpretation of data to advice the adequate use of resources. Although the use of ES in modelling events has a great potential of application, it is not enough explored.

Traditionally, agricultural research has been done empirically, stressing the use of experimentation. As a result of this practice a great amount of data is produced. To obtain feasible information from this data it is necessary the use of statistical methods. By its turn, Statistics obligates that the experimentation be done according to such well defined criterions.

An experimental design, for example, is defined by the nature of the involved variables, environmental and resource constraints, the research objectives, and the already existent knowledge about the observed phenomenon. In addition, as the activities of analysis and inference, the design of experiments hardly depends on the interaction between the researcher and the statisticians.

In this aspect, we face crucial problems for research:

(a) The agricultural researcher, almost always, does not have sufficient knowledge about Statistics.

(b) The statistician, almost always, does not know deeply the agricultural research area.

(c) As a professional that have a highly specialized knowledge, the statistician are not available, particularly by attending many demands.

As consequence of the distance between the agricultural researcher and the statistician, (a and b), we may observe the same problems that Coleman pointed, from which we highlight the followings:

(1) Unwarranted assumptions of process stability during experimentation.

(2) Undesirable combinations of control-variable levels in the design.

(3) Violation or lack of exploitation of known physical laws.

(4) Unreasonably large or small designs.

(5) Inadequate measurement precision of responses or factors.

(6) Undesirable run order.

(7) Inappropriate control-variables settings.

(8) Misunderstanding of the nature of interaction effects, resulting in unwisely confounded designs.

(9) Inadequate identification of factors to be "held constant" or treated as nuisance factors, causing distorted results.

(10) Misinterpretation of past experiments results, affecting selection of response variables or control variables and their ranges.

Concerned with the availability of the statistician (c), its very common not having this kind of professional in the agricultural

research centers for immediate consults, mainly in developing coutries. So, to obtain some support the researcher often waits an excessive period of time.

The problems related above present typical characteristics for an application of expert system. The approach of prototyping used to develop this kind of system may be applied to this case with no restrictions. Such approach consists in developing prototypes for knowledge portions, using a fashion of incremental development.

## 2. OBJECTIVES OF GALILEO

GALILEO is an expert system that intends to help the agricultural researchers in research planning and defining the better experimental design, through a systematic procedure, using the available data and knowledge. We do not intend to reach a complete representation of planning and designing process of experimentation. Rather, we hope the researchers surpasses the knowledge expressed in GALILEO, and by improving his own knowledge, give feedback to the system.

## 3. METHODOLOGY

The explicitation of the statistician knowledge occurs by the following manners:

(1) Interview with the statistician about their experience.

(2) Exhaustive case analysis, searching for situations not explicited in (1).

(3) Use of meta-level knowledge ( e.g., about scientific method) in structuring the knowledge base.

(4) Growth of the knowledge base with other experiences, as well with the feedback of its use.

Items (1), (2) and (3) are considered in this version of GALILEO. (1) and (2) are related to the representation of the statistician knowledge and to analysis of a sufficient amount of problems of relative complexity; (3) addresses to the use of knowledge on research process, specificaly about the scientific method.

As the knowledge in this area will continue to expand, it will be required the improving of the knowledge base, that demands in (4) a continuous effort without a defined end.

## 4. CONTEXT AND DELIMITATION OF THE PROBLEM

Due the natural complexity of the knowledge acquisiton process, it is necessary to consider the context in which GALILEO takes place, in order to define precisely its limits.

The diagram below describes the process that start on the researchers demands with respect to Experimental Statistics.

RESEARCHERS DEMANDS

|

ANALYSIS BY STATISTICIANS:
DEFINITION OF THE PROBLEM CLASS

|

PLANNING AND DESIGN OF THE
EXPERIMENT

|

EXECUTION OF THE EXPERIMENT

|

ANALYSIS OF THE DATA

Considering that the best way to develop an expert system is by prototyping, starting with a knowledge nucleous and then growing in complexity, GALILEO was initially implemented to solve the demands related with simple cases of sampling and experimentation (e.g., simple random sampling and completely randomized design).

## 5.CLASSES OF RESEARCHERS DEMANDS

The demand of agricultural researchers on Experimental Statistics may be grouped, a priori, in five main classes:

(a) Exploratory research: occurs when the researcher believes that there is a cause-effect relation between independent and dependent variables, but he is not sure about the nature of this relation. In this case it is necessary an approach that permits to accept or reject that belief and gives some idea about the relation.

Example: we are concerned to the savanna area and believe that some environmental conditions are related to the presence of a kind of tree specimen.

(b) Hypothesis test: in this case we have a hypothesis about the relation between independent and dependent variables which is expressed by a mathematical function. We want to test this hypothesis. Example: we want to test the hypothesis that the growth speed of the *Sitotroga cerealella* population increases linearly with the moisture of the rice grains.

(c) Parameter estimation: occurs when we already have a generic mathematical function to explain the phenomenon and want to estimate the parameters of this function for a specific situation of the researcher interest. Example: we know what is the generic mathematical function of the corn yield response related to the soil nitrogen level. We want to estimate the function parameters, to a corn hybrid, under a given climate condition, in a specific soil.

(d) Comparison of alternatives: this approach is suitable when we just want to know the values that dependents variables assume when the independents variables change its values. Example: we want to know the grain production of each one of many soybean cultivars in a specific region.

(e) Estimation of populational parameters: in this case one is just interested in obtain estimated values to determined parameters. Example: what is the estimated amount of orange trees affected by *tristeza* virus in a region?

We note that the design will depend on the class the experiment belongs, as well on specific information about the environment where the experiment will be done.

## 6. KNOWLEDGE REPRESENTATION AND STRUCTURE

GALILEO is a rule-based expert system. This form of knowledge representation is widely used in expert systems and it shows to be suitable to the development of prototypes, that is validated by the specialists, in this case, by the statistician. The adequation of the production rules become evident since the problems of experimental design may be subdivided into independents subproblems that, when solved, lead to the solution of the problem. Moreover, the flexibility of the production rules shows to be useful in modelling the knowledge of a very dynamic area, where the problem solving process evolves constantly, being enriched by each solved problem.

The knowledge structure was defined according to the process of attendance of the researchers demands, i.e., firstly specifying the research problem and then classifying it according to the demands reported in **5. CLASSES OF RESEARCHERS DEMANDS**. Secondly, the problem is analysed and the experimental design specified according to its objectives and informations given by the researcher about the investigated area and related specifically to the problem.

A generic session of GALILEO, in its first version, traverse two steps: classification of the researcher demand and design definition.

The dialog with a consultant follow the script below:

1) Report the independent variables by descending order of importance:

2) Report the dependent variables by descending order of importance:

3) The variable _____ is:

( ) quantitative ---> *4*

( ) qualitative ---> *7*

( ) ordinal ---> *7*

( ) binary ---> *7*

4) The quantitative variable is _____:

( ) continuous ---> *5*

( ) discrete ---> *6*

5) The minimum and maximum values for the quantitative continuous variable is:

| MINIMUM | MAXIMUM |
|---|---|
| ( ) minus infinite | ( ) real negative |
| ( ) real negative | ( ) zero |
| ( ) zero | ( ) real |
| ( ) real | ( ) plus infinite |
| ( ) _____,__ | ( ) _____,__ |

*for any combination of minimum and maximum ----------> 7*

6) The minimum and maximum values for the quantitative discrete variable is:

| MINIMUM | MAXIMUM |
|---|---|
| ( ) integer negative | ( ) integer negative |
| ( ) zero | ( ) zero |
| ( ) integer | ( ) integer |
| ( ) _____ | ( ) _____ |

*for any combination of minimum and maximum ----------> 7*

7) The levels of the variable _____ will be pre-defineds?

( ) Yes. ---> *8*

( ) No. ---> *9*

8) Report the levels of the variable _____:

*for any case ---> 9*

9) At the environment where the data will be collected are there non-controlled factors, presented in different levels, that may influence the values assumed by the variables?

( ) Yes ( ) No.

*for any case ---> 10*

10) Is there some constraint to the randomization of the levels of the independent variable _____?

( ) Yes. ( ) No.

*for any case ---> 11*

11) Is there interest in investigating the existence of interaction between the independent and dependent variables _____ and _____?

( ) Yes. ( ) No.

*for any case: if in 2) there is no variable ---> 15*

*else ---> 12*

12) Your actual knowledge about the studied phenomenon allow you to state that there is a cause-effect relation between the independent and dependent variables?

( ) Sure. ---> *13*

( ) Probably. ---> *16*

13) The relation between the independent and dependent variables are known and accepted?

( ) Yes. *if in 1) the first variable was quantitative ---> 18*

*else ---> 17*

( ) No. ---> *14*

14) Do you have a hypothesis to the relation between the independent and dependent variables and want to test it?

( ) Yes. ---> *19* ( ) No. ---> *17*

15) Do you confirm that your interest is only to estimate the values assumed by the independent variables at the studied event?

( ) Yes. ---> *OK (class: estimation of populational parameters)*

( ) No. ---> *End: there is nothing to conclude.*

16) Do you confirm that your interest is to investigate the existence of possible cause-effect relations between the independent and dependent variables?

( ) Yes. ---> *OK (class: exploratory research)*

( ) No. ---> *End: there is nothing to conclude.*

17) Do you confirm that your interest is to estimate the values assumed by the independent variables at the studied event, comparing them?

( ) Yes. ---> *OK (comparison of alternatives)*

( ) No. ---> *End: there is nothing to conclude.*

18) Do you confirm that your interest is to estimate quantitatively the dependent variables responses to changes at the independent variables levels?

( ) Yes. ---> *OK (class: parameter estimation)*

( ) No. ---> *End: there is nothing to conclude.*

19) Do you confirm that your interest is to test your hypothesis on the relations between the independent and dependent variables?

( ) Yes. ---> *OK (class: hypothesis test)*

( ) No. ---> *End: there is nothing to conclude.*

In the scheme above the subsequent step related to each question is placed immediately after the answer , as below

Yes. ---> *19*

Based on the answers to questions of the dialog, the system use its heuristics to indicate a design. As example we consider the following answers to the dialog:

1) Vegetable species.

2) (a) External bacterias in roots.

(b) Internal bacterial en roots.

3) For Vegetable species ---> qualitative.

For External bacterias in roots ---> quantitative.

For Internal bacterias in roots ---> quantitative.

4) For External bacterias in roots ---> discrete.

For Internal bacterias in roots ---> discrete.

6) For External bacterias in roots ---> minimum = zero.
maximum = integer.

For Internal bacterias in roots ---> minimum = zero.
maximum = integer.

7) For Vegetable species ---> Yes.

8) (a) Rice.

(b) Beet.

(c) Onion.

(d) Potato.

(e) Carrot.

(f) Coentro.

(g) Corn.

(h) Lettuce.

(i) Tomato.

9) No.

10) No.

12) Sure.

13) No.

14) No.

17) Yes.

At this moment, the system concludes that the researcher demand belongs to the "Comparison of alternatives" class and the best design is the "Completely randomized" with four replications.

## 7. DEVELOPING AND EXECUTION ENVIRONMENT

The initial prototype was developed in a PC AT/386 IBM-compatible, with 2 MB RAM, 60 MB HD and 1 drive de 3 1/2 ", with DOS version 5.0 . It was used the shell EXSYS (**E**xpert **S**ystems) in building the knowledge base, dialog specification and results exhibition.

## 8. FUTURE EXTENSIONS

After the knowledge about sampling and experimentation had been captured and sedimented on a level sufficient to help in the solution of a significant amount of cases, we intend to work in a module for analysis and interpretation of data that came from the implemented experiments, according to the indicated designs.

## 9. GLOSSARY

*Cause-effect relation:* relation such as $y=f(x)$ in which $y$ is regarded as "caused by" $x$.

*Completely randomized design:* a kind of design where the treatments are applied in a chance basis to the experimental units.

*Confounded designs:* a device whereby, in large factorial experiments, the size of blocks is limited by sacrificing some of the independent comparisons relating to the higher-order interactions.

*Constraint:* a constraint in a set of data is a limitation imposed by external conditions, *e.g.* that the total number of experimental units cannot be upper than 60.

*Continuous variable:* a variable that asumes values in the real numbers set.

*Control variables:* see independent variables.

*Dependent variable:* in a cause-effect relation is the *y*-type variable.

*Discrete variable:* a variable that assumes values in the integer numbers set.

*Estimation:* is concerned with inference about the numerical values of unknown populational values from incomplete data such as a sample.

*Experiment:* an action or operation undertaken in order to discover something unknown, to test hypothesis, or stablish or illustrate some known truth.

*Experimental design:* logical disposition of the experimental units, according to environmental characteristics, aspects that one wants to observe, and the existing constraints.

*Experimental unit:* minimum physical space on which each treatment is apllied.

*Factor:* quantity under examination in an experiment as a possible cause of variation, *e.g.* in a "factorial" experiment.

*Factorial experiment:* an experiment designed to examine the effect of one or more factors, each factor being applied at two levels at least so that differencial effects can be observed. The term is frequently used in a slightly narrower sense, as describing an experiment investigating all possible treatment combinations which may be formed from the factors under investigaton. The "level" of a factor denotes the intensity with which it is brought to bear. It may be measured quantitatively, as when fertilizer is applied to plots in a given weight per unit area, or qualitatively, as when plants are considered at two levels "inoculated" and "not inoculated".

*Gradient:* variation of physical characteristics of the environment inside the experimental area.

*Independent variable:* in a cause-effect relation is the *x*-type variable.

*Interaction effect:* is a measure of the extent to which the effect (upon the dependent variable) of changing the level of one factor depends on the level(s) of another or others.

*Latin square design:* one of the basic statistical design for experiments which aim at removing from the experimental error the variation from two sources, which may be identified with the rows and columns of the square.

*Level:* see factorial experiments.

*Ordinal variable:* the variable that marks position in an order or series, as first, second, and so on.

*Parameter:* an unknown quantitity which may vary over a certain set of values.

*Qualitative variable:* variables that assume non-numerical values (*e.g.* sex, nationality or comodity).

*Quantitative variable:* variables that assume numerical values *(e.g.* height, weight or price).

*Randomization:* allocation of the treatments to the experimental units in a such way that each unit has the same chance to receive a given treatment. Its function is to avoid bias of media estimate and the experimental error.

*Randomized blocks design:* an experimental design in which each block contains a complete replication of the treatments, which are allocated to the various units within the blocks in a random manner and hence allow unbiassed estimates of error to be constructed.

*Range:* the largest minus the smallest of a set of variate-values. It is an elementary measure of dispersion.

*Repetition:* a term denoting the execution of a statistical inquiry, at different points in space or time, usually as part of a coordinated programme, as distinct from replication.

*Replication:* the execution of an experiment or survey more than once so as to increase precision in to obtain a closer estimation of sampling error. Replication should be distinguished from

repetition by the fact that replication of an experiment denotes repetition carried out at one place and, as far as possible, one period of time. Current usage on this point is often rather loose.

*Response variables:* see dependent variables.

*Sampling:* the action of taking a part of a population, or a subset from a set of units, which is provided by some process or other, usually by deliberate selection with the object of investigating the properties of the parent population or set.

*Simple randomized sampling:* is a sampling design where the probabilities of selection of members are all equal and are constant throughout the drawing.

*Split-plot design:* an experimental method in which additional or subsidiary treatments are introduced by dividing each plot into two or more portions.

*Stratification:* the division of a population into parts, known as strata; specially for the purpose of drawn a sample, an assigned proportion of the sample then being selected from each stratum.

*Stratified sampling:* occurs when a sample is selected from a populational which has been stratified, part of the sample coming from each stratum.

*Treatment:* handling form of the material allocated to an experimental unit, which effect on each factor is observed.

## 10. CONSULTED BIBLIOGRAPHY

COCHRAN, W. G. **Tecnicas de Muestro.** Mexico: Editorial Continental,1976. 507 p.

COLEMAN, D. E.; MONTGOMERY D. C. A Systematic Approach to Planning for a Designed Industrial Experiment. **Technometrics.** v. 35, n. 1, p. 1-27, Feb. 1993.

DESLANDRES, V.; PIERREVAL, H. An Expert System Prototype Assisting the Statistical Validation of Simulation Models. **Simulation.** v. 56, n. 2, p. 79-89, Feb. 1991.

DUHEM, P. **Cadernos de História e Filosofia da Ciência.** Salvar os Fenômenos. Campinas, v. 3, 1984. Suplemento.

EHRENBERG, A. S. C. **A Primer in Data Reduction.** Chichester: Jonh Wiley & Sons, 1983. 305 p.

FORSYTHE, D. E. Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. **Social Studies of Science.** v. 23, n. 3, p. 445-477, Aug. 1993.

GOMES, F. P. **Curso de Estatística Experimental.** Piracicaba: Nobel, 1978. 430 p.

KENDALL, M.G.; BUCKLAND, W.R. **A Dictionary of Statistical Terms.** Edinburgh: Oliver and Boyd, 1960. 575 p.

KOSTOV, K. Logical-methodological Basis of the Script in Automation of Research. **Science of Science.** v. 4, n. 2, p. 149-157, 1984.

LITTLE, T. M. **Statistical Methods in Agricultural Research.** Davis: University of California, 1972. 242 p.