Can Non-Expert' Users Analyse Data?
A Survey and a Methodological Approach

Spinakis Antonis[1]
Pantouvakis Angelos [2]
Thalassinos Eleftherios[3]

Abstract

Nowadays, statistical methods are applied in a vast extent of different disciplines by a lot of people that, very often, have a rather little knowledge of statistics. This fact can lead to misuse of statistical methods. The statistical community has recognized this fact and now statistical packages try to offer some guidance to their users. This paper presents the design of a statistical package, oriented to users with medium or little knowledge of statistics. The results of a survey about the user requirements of such software are presented. The general architecture and the functionality of the system are discussed.

Keywords: semi-expert system; attributes; user requirements; statistical practice;
JEL Classification: C42. C44

Introduction

In recent years we have seen an explosion of data availability together with an increasing appeal of statistics in the everyday life, especially in socio-economic matters. Large databases can be created containing billions of data and even simple statistical procedures are needed in order to summarise the tremendous amounts of data. On the other hand, this impact of statistics has been recognised and almost every university department has at least one course about elementary statistics (see [1]). Combining these two events, one can see that, in recent years, statistical methods are used in a wide range of disciplines ([2],[3]). Unfortunately this expansion does not imply that statistics is used in the correct manner.

It is fact, that the majority of socio-economic analysts and some of the official statisticians are not necessarily able to use statistical data analysis methods in the most appropriate way.

The need for some kind of guidance has been recognised by commercial packages that started offering it via "statistical wizards", like those offered by SPSS or Statgraphics among others. However, these statistical software packages are designed to help the user with a moderate statistical background rather than the user with a small level of knowledge.

[1]	QUANTOS SARL,191, rue Saint Honort,75001 Paris, FRANCE
[2]	University of Ioannina, School of Natural Resources and Enterprise Managemen,Agrinio, Greece
[3]	Professor, Dept. Chairman, University of Piraeus, Karaoli and Dimitriou 40, 13582, Piraeus, Greece, tel. ++30210-4142525, e-mail: thalassi@unipi.gr, thalassinos@aias.gr

The design of "Statistical Expert Systems" has been discussed many years ago (see [4]). It has been recognised very early that it is quite difficult to enter the statistical knowledge in an expert system. For this reason, the "expertise" has been focused on specific aspects of generic software, or it has been limited to particular application domains (see [5],[6],[7]). The basic idea is not to replace the statisticians but to protect the user from misusing them ([8],[9]).

In this paper we discuss the architecture and the design of an 'expert-system' for 'non-expert users'. As 'non-expert' users we define all those that use statistical methods but they are not statisticians and they have a small knowledge about statistics, perhaps by some courses offered during their study, or they do not have any knowledge about statistics at all. This software, named "Extended Statistical Information System" (hereafter X-STATIS) aims at helping the 'non-experts' to use statistics in a correct manner and being able to understand the results of an analysis.

The main concept of X-STATIS is to enable these analysts to make use of modern IT-based systems and tools, in order to improve the quality of work. Especially, X-STATIS will provide a simple, interactive, user-friendly environment, which will assist non-experienced users in taking the appropriate decisions following the right paths. It will also provide sufficient information so that the non-expert user can understand and evaluate the obtained results. The challenge of this project is to identify, define and then bring together a working structure element that are essential for giving the proposed system the ability to guide and assist the user during a data analysis session.

The remaining of the paper proceeds as follows. Section 2 describes the results from a survey in three European countries about things that a 'non-expert' user wants from such statistical software. Section 3 describes the functionalities of such a system while section 4 the general architecture. In section 5 we focus on multiple regression and we describe how 'statistical expertise' on multiple regression could be enhanced into statistical software. Concluding remarks can be found in section 6.

Capturing User Requirements

A user requirements study was carried out at the beginning of this project in order to understand the needs of target group (non-experts) in the market. This study consisted of three major parts. Primarily, a research on several statistical programs had been developed (market survey). This survey extends the results of other authors (see [10]) as it includes general statistical packages (SAS, SPSS, Statgraphics, Statistica, SPLUS, SYSTAT, MINITAB, SPAD) as well as packages for specific methods or packages that offer some kind of statistical methods (STAMP, ECOTRIM, TRAMO/SEATS, X-12 ARIMA, TSP, Microfit, Eviews, RATS, GAUSS, SigmaStat, Mathematica, Autobox, Forecast Pro, Microsoft EXCEL among others). The comparison can be found in [11].

In a second step, a qualitative survey (focus groups survey) was conducted, where the main aim was to reveal the underlying factors building opinions, behaviours or attributes of target group regarding the "ideal" statistical software. Finally, a

quantitative survey was judged as necessary in order to identify with improved accuracy the requirements of 'non-expert' users. The survey was conducted in three European countries (Italy, France, and England) and the target population was employees in a variety of disciplines from medical to economic sectors that have a moderate or small statistical background. The method used was telephone survey using random telephone numbers from a catalogue containing companies in each country. In total, 525 questionnaires were filled.

The main findings of the survey will be briefly given. The majority of them avoid the use of statistical packages. Microsoft EXCEL was quite popular. 87% of the responders thought that they have sufficient knowledge of Excel. Among statistical packages there were some local preferences (e.g. in France they use SPAD, which is almost ignored in other countries). A large portion (54%) answered that they used their own companies' software implying a general-purpose software for the jobs inside the company. The majority of them feel uncomfortable with statistical packages. The statistical methods that they use were descriptive statistics, including group comparisons, time series and regression methods and rarely more advanced statistical procedures. They reported that guidance for the selection of the appropriate method and the explanation of the results would be quite helpful. They also provided information about IT functionalities that were considered as helpful.

The above results provide a guidance of what the 'non-expert' wants from statistical software and they have been embodied in the functionalities of the system described in the next section. An extended presentation of the user-requirements phase is available in [11].

The Main Functionalities of X-STATIS System

The main goal of the system is to provide users with sufficient information to understand and evaluate the obtained results of statistical analysis. It also aims at creating a user-friendly information system since it will offer to the public facilities that fully will utilize modern IT-based systems and tools and will improve the quality of non-expert work. In the following, we describe some of the main functions of X-STATIS.

Friendly Interface

This module makes available to the user in a user-friendly way functionalities implemented by the system, targeting to fulfill the objectives of the project. This module is responsible for interacting with the user, process and forward to the appropriate modules user requests, and present in an appropriate way any result obtained.

Data Base Management

This is managerial activity that will apply information system technology and management tools to the task of managing an organization's data resources to meet the information needs of non-expert. The database management will control the creation, maintenance, and use of the databases of an organization and its end users.

Visual Query Builder

This will allow the import of several data file format into the X-STATIS system by using the ODBC drivers. The user can connect easily with a database containing the desired data, and select the data to be examined by constructing a visual query.

DDE Data Manager

This module will allow the communication with other DDE applications (e.g. MS Excel) for automatic exchange of data.

Automatic Descriptive Statistics Analysis

This module will perform Summary Statistics and Graphical Analytic Techniques for one, two and more than two variables. It will also help the user with interpretation of the generated results, by providing sufficient information and guidance in order to evaluate and understand the outcome of the analysis.

Automatic Multiple Regression Analysis

The system will automatically find the multiple regression model, and describe the relationship between the response variable and the predictors variables. The rules parameters can be designed and modified by experts in the field of statistics. In the sequel we will treat this module in more detail. It must be mentioned that transferring statistical knowledge from experts to an intelligent system is quite challenging though difficult due to the multiple criteria that an expert takes into account into the implementation and the interpretation even of a simple regression model.

Automatic Analysis of Time Series

In this case, expert knowledge of time series is embodied in the X-StatIS software. Predefined rules guide the user towards the appropriate analysis technique. The user is then guided through the process of applying methods and interpreting results.

Automatic elimination of outliers

Using suitable diagnostics, possible outliers can be detected and examining the influence of them, some of them can be eliminated automatically.

Dynamic Graphical Methods

Additional to the traditional graphical techniques some dynamic graphs will be developed. The essential characteristic of a dynamic graphical method is the direct manipulation of elements of a graph on a computer screen; the manipulation is carried out using an input device such as a mouse, and in high-performance implementations, the elements change virtually instantaneously on the screen.

Manual Use of Methods

The system will allow the manual use of methods if any of the users want to work on this direction.

Scripting Modules

This component will offer functionality, which allows data analysis to be repeated again in the future on a new data set. It will allow the creation of macros through automatic recording of user interactions, ask the user to specify the data set(s) that the recorded macro is to be applied, and save all this information into a repository which could be executed later on without any user interaction.

Architecture of X-STATIS

The main goal in designing the architecture of the X-STATIS application is to produce a modular, open and generic structure, which will provide the appropriate flexibility to customize the application according to the user needs. This will be achieved by providing an architecture where plug-in modules will be able to attach it. The architectural presentation of the X-STATIS system in terms of operation is depicted in the figure 1 where the main software modules are illustrated. These modules are:

Figure 1
X-STATIS architecture
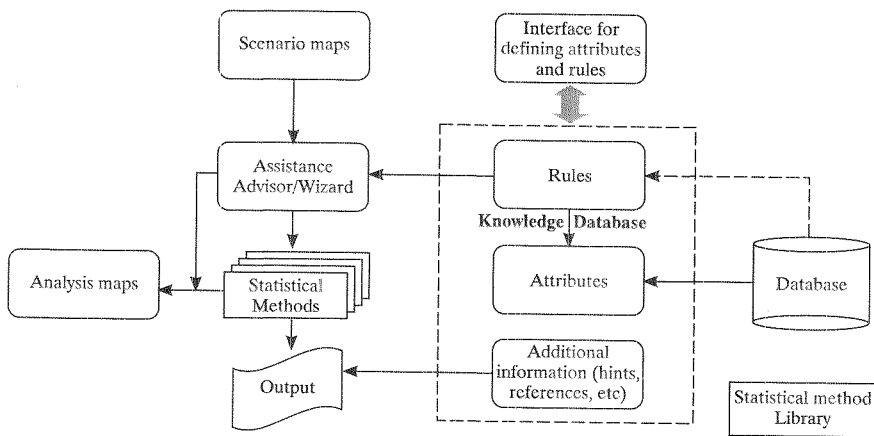
Assistance Advisor Wizard

This is an open and parameterised component, which, based on an intelligent dialogue interface, will interact with the user trying to figure out what the user wants to do with the data and directing him/her appropriately in the selection of appropriate data analysis methods, manipulation and interpretation of results. It will give the proposed system the feel of a semi-expert system since it will interact with the user and the other system components in such way that it will hide the complexity and the statistical expert knowledge required for data analysis. The user will be guided "smoothly" and intelligently to the appropriate method without the need to be a statistical expert in order to understand and apply the methods offered by the system.

Knowledge Database

This is a backbone database containing the statistical attributes and general statistical rules used in the assistance of the appropriate statistical method for data analysis supported by the software. It will also contain links to external references that can help in data analysis, warnings like what actions are not appropriate to perform on the data, advises on how to proceed with data analysis, references to other statistical information, etc. The database architecture will be open in order to be expandable with new statistical attributes and rules that the Metadata Advisor wizard supplies when the X-STATIS system is applied to a new database collection.

Statistical methods library

These modules will implement statistical methods – algorithms. They will have a plug-in form, which will allow the user to customise the X-STATIS system by integrating into it only some data analysis methods. Implementing and integrating these plug-in data analysis modules in the system helps staying in line with the expandability and the customisation that the proposed system will offer.

Scenario maps

This component will offer functionality to help users with little or no knowledge about data analysis, to benefit from an environment, which visually guides them through their data analysis. To achieve this the component will offer appropriate interfaces where expert analyst could use and create visually a sequence of steps that should be taken. Novice users could use later these scenarios by following the predefined steps and perform a complete statistical data analysis. At each step, available help will be offered to the user in order to understand the analysis purpose.

Analysis Maps

The analysis map component offers functionality that allows recording of the steps taken by the user during the analysis process. It will do this visually by creating diagrams of the steps taken during a data analysis session. Therefore, the structure of an analysis map will change by adding new item in the diagram as the user proceeds through the analysis. The component can be thought as a monitor that records the several actions taken by the system or as a "history" concept. In addition, the functionality, which offers, provides an overview of the data analysis session. Each analysis step will be associated with the appropriate report including the data analysis results that correspond to the applied method. Another major service offered is the ability to move back to a specific data analysis step and continue from that point on by following a different data analysis approach.

The Outlines of System through Multiple Regression Method

General approach

X-STATIS will run on MS Windows and will develop in each database separately taking into account the specialities of them. There will be a system of attribute infor-

mation, which will be attached to the database structure. The attribute information will consist of the variable types (continuous, nominal, etc.), the variable labels, value labels and other information related with data. This attribute information will provide a basis for the development of a structured way of handling metadata. It is important to point the role of experts in the creation of the knowledge database, which consist of attributes, rules and other information.

Opening the system the main screen will have a form like Excel. The user can load a data table, which consists of the variables that he desires to use for the analysis. When the table has been imported into the program, the main screen will include all the information of selected data. Selecting the variables, a module will be activated which will be suitably designed to protect from misusing statistical methods where they are not appropriate. This module consists of rules regarding each method and will decide which of them will be available in the next steps. More analytically, the module of rules will be consisted of the following four categories:

* Rules for selecting each method
* Rules for selecting the most appropriate method for the data
* Rules for controlling the flow of each method
* Rules for particular messages appearing at the output

Figure 2 illustrates the operational way of system. Hence, the system automatically will run an algorithm, which will examine the attributes associate with the data and will decide which method of those available will be applied to the particular choice of variables.

Figure 2
Flowchart describing the operations for working with X-STATIS
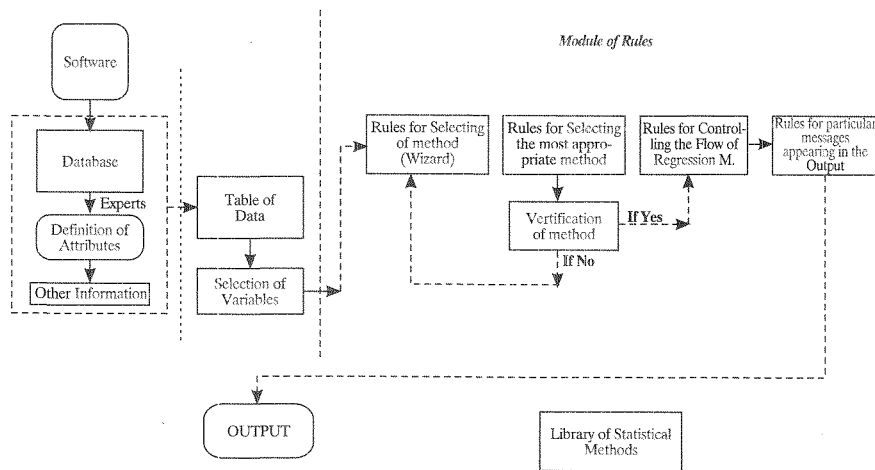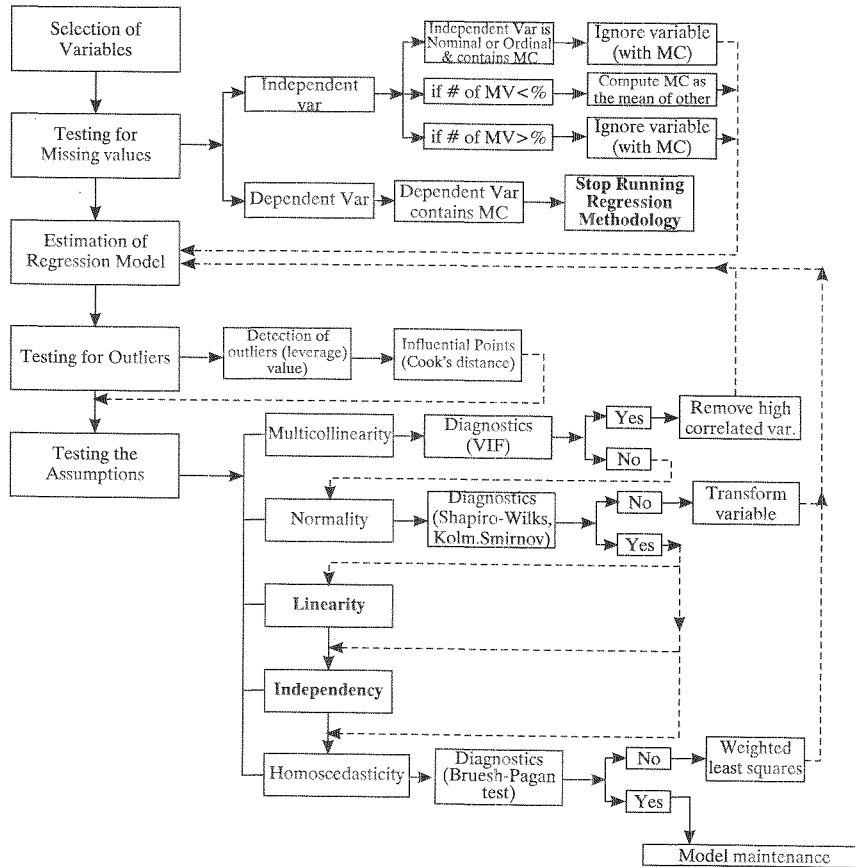
Application – Multiple Regression Method

In this subsection, we will try to describe the procedure (that it was considered appropriate) in order to perform multiple and simple regression ([12]). X-STATIS Regression will perform ordinary least squares multiple regression. Moreover, the X-STATIS Regression method will include regression diagnostics for examining the structure of the data and for checking model assumptions as well as appropriate messages when something went wrong. The challenge for a system that can run multiple regression in an 'expert' way is large. On the other hand it is clear that embodying all the statistical knowledge about regression is not at all possible but at the same time the 'non-expert' user is indifferent of such an advance approach. We propose a compromise by allowing diagnostics and checking the basic assumptions, but without going into sophisticated solutions. Note that we aim at expanding the module by offering more procedures at the near future.

Figure 3 presents the regression flowchart. The majority of these procedures are not transparent to the user. The only interaction between the user and the system will be realized during the selection of variables (setting the dependent and the independent variables). During the execution of method, only suitable warnings regarding the validity of assumptions is going to be presented to the users. The most usual statistical problems involved in regression analysis are described in the following:

- To obtain optimum estimates of the unknown regression parameters
- To test hypothesis about these parameters
- To determine the adequacy of the assumed model
- To verify the set of relevant assumptions.
- To predict the value of the response variable for some new observations

Figure 3
Regression Flowchart

Selection of Variables

Testing for Missing values

Estimation of Regression Model

Testing for Outliers

Testing the Assumptions

Independent var

Independent Var is Nominal or Ordinal & contains MC → Ignore variable (with MC)

if # of MV<% → Compute MC as the mean of other

if # of MV>% → Ignore variable (with MC)

Dependent Var → Dependent Var contains MC → **Stop Running Regression Methodology**

Detection of outliers (leverage value) → Influential Points (Cook's distance)

Multicollinearity → Diagnostics (VIF) → Yes → Remove high correlated var.
No

Normality → Diagnostics (Shapiro-Wilks, Kolm.Smirnov) → No → Transform variable
Yes

Linearity

Independency

Homoscedasticity → Diagnostics (Bruesh-Pagan test) → No → Weighted least squares
Yes

Model maintenance

Conclusions

The X-STATIS system can be seen as an expert statistical system, which will address to non-experts data analysts. There is skepticism whether such software can be of an "expert system" type since statistical knowledge cannot easily be enhanced into a computer program. For example consider the case of detection of hetero-scedasticity in linear regression. Statisticians could detect it by a simple glance in the residual plot. This is not easy for the computer and hence we have to provide a measure for detection purposes. Which one is the best? No clear answer exists. Thus, we are limited in using some measure(s) that are known to the literature that 'work' but, clearly, controversial issues may arise. Such considerations are real challenges for statisticians. Consider another problem, that of using warning in several circumstances during the analysis. A 'non-expert' user is totally unable to

handle a warning about the inefficiency of the regression model due to violation of the normality assumption. A system that simply results in such a warning, in fact, does not prevent the user by misusing statistics. On the other hand, a system that did not provide any results due to such problems would not be any useful at all. To add sophisticated statistical methodologies in order to resolve the normality problem that occurred could lead to results without practical value for anon-expert user. This example reveals the limitations of such an intelligent system, as well as the need for compromises during the analysis phase. A computer-system in no way can substitute a well-educated statistician with a great experience in real data analysis. However, taking into consideration that we are in the beginning of this project, our basic aim is to enable non-expert data analysts to make use of modern IT-based systems and tools, and to improve the quality of their work.

Acknowledgements

References

[1] Loftsgaarden D.O and Watkins, A.E. Statistics teaching in colleges and universities: courses instructors and degrees in fall 1995. *American Statistician* 52, 308-314

[2] Kettenring, J.R (1997) Shaping Statistics for Success in the 21st Century. *Journal of the American Statistical Association,* 92, 1229-1234

[3] Moore D.S (2001) Undergraduate Programs and the Future of Academic Statistics. *American Statistician,* 55, 1-7

[4] Hand, D.J. (1984) Statistical Expert Systems. *Statistician,* 33, 351-369

[5] Woollard, R., Clark, C. and Jury. J.W. (1996) An intelligent statistical process control system for paper conversion. *TAPPI Journal,* 79, 137-141.

[6] Grabowski, B. L., and Harkness, W. L. (1995), "Expert Systems as an Instructional Strategy in Statistics: A Case Study" in Proceedings of the Section on Statistical Education, American Statistical Association, pp. 90-94.

[7] Prat, A, Sole, I, Catot, J.M. and Lores, J., (1998), FORCE4/R, A new software product for forecasting and seasonal adjustment. Proceedings of the International Seminar on New Techniques and Technologies for Statistics, Sorrento, 4- 6 November,, pp. 429-434.

[8] Hand, D.J. (1987) The Applications of Artificial Intelligence in Statistics. Paper presented in Doses Seminar, 1-3 December, 1987 Luxembourg

[9] Hand, D.J.(1986) Expert Systems in Statistics. Knowledge Engineering Review,1,2-10

[10] Morgan, W.T. (1998) A review of Eight statistics software packages for general use. *American Statistician*, 52, 70-83

[11] Project X-STATIS, D-2.1 Report: Definition and Assessment of User Requirements, including results of Market Survey & Evaluation of existing Statistical Software packages

[12] Draper and Smith (1981). *Applied Regression Analysis.* New York: John Wiley & Sons, Inc.