ANALYSIS OF STATISTICAL SOFTWARES USING ANALYTICAL HIERARCHY PROCESS

José Fabiano da Serra Costa*
Rio de Janeiro State University (UERJ) - Brazil
E-mail: fabiano@ime.uerj.br

Marcelo Goulart Correia
Rio de Janeiro State University (UERJ) - Brazil
E-mail: mgoulart@gmail.com

Letícia Tissi Tracierra de Souza
Rio de Janeiro State University (UERJ) - Brazil
E-mail: lettissi@hotmail.com

ABSTRACT

This article presents the usage of the Analytical Hierarchy Process (AHP) method to support the selection of the most suitable statistical software for a public university. AHP is a renowned multicriteria methodology for decision making, based on individuals' capability of building and understanding hierarchies. In this case, the method was applied to respond the need of Institute of Mathematics and Statistics (IME) of University of State of Rio de Janeiro (UERJ) to acquire an statistical software to be used by students and faculties. In this context, AHP and its main characteristics are presented. It also reinforces the importance of statistics software as a tool to help students learn and become familiar to computational statistics methods, highlighting techniques broadly used in the market place.

Keywords: AHP, Statistical Software, Multicriteria.

1. Introduction

Computational Statistics is truly a multidisciplinary area and the diversity of problems of these fields has created an atmosphere of researches and developments. This perception has taken place since the beginning of its usage and statistics labs efforts and their applications had led to a fast evolution. A great number of statistical software products, both specific and general, are available on the market. Many of these products offer a wide range of statistical procedures and a great flexibility to exploit and analyze data, however access and large scale use have not followed such development.

Through statistical teaching perspective – for specific statistics graduation courses as well as for other courses which offer statistics in their curricula – the use of such software is fundamental to prepare the students to develop stochastic models and to make education practices more dynamic. In addition to that, most companies demand knowledge in determined software solutions when hiring. Therefore, statistical software usage is mandatory in information technology laboratories of higher education institutions.

* Corresponding author

In this article, the acquisition of a statistical software product to be used by students and faculties in Institute of Mathematics and Statistics (IME) of Rio de Janeiro State University (UERJ) is analyzed by using a multicriteria decision making model (Analytical Hierarchy Process – AHP). The analysis was based on specialists' evaluation of five products regarding a set of criteria.

2. Analytical Hierarchy Process - AHP

AHP can be considered a decision making approximation that involves structuring of multiple choosing criteria in a hierarchy (DSS, 2008). The method evaluates the relative importance of those criteria, compares alternatives to each criterion and determines a ranking for the alternatives. So, AHP aims the selection and choice of alternatives, in a process that regards different evaluation criteria. It is important to notice an advantage of AHP method since it allows comparing both quantitative and qualitative criteria. It is considered one of the most known and disseminated decision making tools, having the greatest number of applications related in the literature (VAIDYA & KUMAR, 2006).

Many alternatives to conjugate information given by different evaluators (specialists) have been proposed and many of those came very close to consistency (FORMAN & PENIWATI, 2008; COSTA & BELDERRAIN, 2009; EHRLICH, 2004; SAATY & PENIWATI, 2007; FREITAS *et al*, 2008). However, what really matters is that reciprocity and transitivity basic properties of the matrix are respected.

3. Computational Statistics

This rapid and sustained growth in processing power of the computers observed from the second half of twentieth century had a strong impact in statistics practice. Older statistic models were mostly always linear, but modern computers and proper numerical algorithms caused an increased interest in non-linear models (especially neural networks and decision trees) and propitiated the creation of new types, such as the generalized linear model and multilevel model.

The development of Computational Statistics field was completely fragmented with new advances in many areas – some of these areas were developed by people with particular interests and computational experiences, and others by people whose interests were to investigate specific applications, and therefore needed to resolve computational problems. According to Hair *et al* (2005), it is almost impossible to discuss an application of multivariate techniques without analyzing the impact of Computational Statistics. The broad application of computers to process huge and complex data bases has significantly promoted the usage of multivariate statistic methods.

For the purposes of this work, five representative software products were selected:

3.1 R Project

R is a language and an integrated development environment to statistical and graphical calculations. It was created by Ross Ihaka and Robert Gentleman in Auckland University, New Zeland, and then developed on a collaborative effort by people in several locations in the world. It is highly expansible with the use of packages, which are libraries oriented to specific functions or study areas (THE COMPREHENSIVE R ARCHIVE NETWORK, 2010).

3.2 SAS

SAS (Statistical Analysis System) is probably the biggest computational package offered on the market. The number of analysis tools of SAS Company has gradually increased and the software main purposed today is no longer statistical analysis. Statistics package is only one of the many tool options placed by the company on the market. From 25 modules (one main module and 24 assembled) offered, at least six provide some sort of data analysis support (SAS BRASIL, 2010).

3.3 SPSS

This is an evolution of the command oriented version projected originally to mainframes, and then to DOS, OS/2 and finally Windows. The software is arranged in modules, including: basic, professional, advanced and others, which have differentiated prices and are commercialized separately (PASW STATISTICS, 2010). The SPSS (Statistical Package for the Social Sciences) was originated in Social Sciences and research analysis fields, but nowadays it is used in many different areas, comprehending a sophisticated gathering of procedures to survival analysis used in medical studies.

3.4 Statistica

Statistica is a package produced by StatSoft, founded in 1984 by a group of university professors and researchers. It is a broad range system involving integrated data analysis, graphics, data base management and customized developments focusing on users from Engineering, Sciences and Business (STATSOFT, 2010).

3.5 S-Plus

To Ripley (1997), S-Plus (TIBCO Software) is one of the most complete statistical packages commercially available, presenting a development environment to several researches in statistical methodology and incorporating recent researches on data visualization. The package is constituted of two parts: one is the interface based on the graphic package Axum, and the other is S-Plus mechanism, which is implemented as a DLL. It also incorporates the language S, with extensions.

In Table 1, some characteristics of the selected products are shown:

Table 1. Characteristics of software products.

Software	Estimated cost	Maintenance cost	Open code	Operational System
R Project	Free	Contract specific, company doesn’t offer service	Yes	Windows, Mac OS, Unix/Linux
SAS	Not disclosed by the company, around US$ 30.000 / processor and type of module	Contract specific, company offers service	No	Windows, Unix/Linux
SPSS	not disclosed by the company, around US$ 1.600 / user and type of module	Contract specific, company offers service	No	Windows, Mac OS, Unix/Linux
Statistica	not disclosed by the company, around US$ 700 / user and type of module	contract specific, company offers service	No	Windows
S-Plus	not disclosed by the company, around US$ 1.000 / user or type of module	contract specific, company offers service	No	Windows, Unix/Linux

4. Five alternative model

From the main focus selection, it was possible to determine which criteria and alternatives would be evaluated by specialists. For that, it was taken into account the opinion of the specialists of the Informatics Lab of Institute of Mathematics and Statistics (IME). After some considerations, the model was established consensually.

The selected criteria were:
- Product Price ($P) – Software product price, disregarding updating and maintenance costs.
- Maintenance Cost ($M) – Including technical support and corrective maintenance.
- Training Time (TT) – Verification of the learning curve of each software product.
- Required Physical Structure (RPS) – Evaluation on the necessary hardware to execute the software.
- Amount of Statistical Analysis (ASA) – Number and variety of statistical analysis provided.
- Tool Complexity (TC) – Complexity level and skills requirements to use the software.

The alternatives were: R Project (A_1); SAS (A_2); SPSS (A_3); Statistica (A_4); S-Plus (A_5).

After defining criteria and alternatives, it was necessary to choose specialists to evaluate those aspects. Since this decision is an important matter for development, the chosen experts were ten professors of Institute of Mathematics and Statistics (IME) of University of State of Rio de Janeiro (UERJ) who had good knowledge of statistical software.

Each specialist received seven surveys: one to evaluate the priority of the criteria in relation to the main focus; and six others to evaluate the alternatives in relation to the criteria. The deadline was estimated in fifteen days. All specialists answered the surveys in time, providing their opinion. After receiving the surveys and proceeding the calculations, the following Local Weight (LW) were obtained:
LW $P = (0,536; 0,106; 0,108; 0,116; 0,134) – Product Price.
LW $M = (0,541; 0,095; 0,129; 0,119; 0,115) – Maintenance Cost.
LW TT = (0,169; 0,206; 0,250; 0,183; 0,191) – Training Time.
LW RPS = (0,301; 0,137; 0,165; 0,167; 0,230) – Required Physical Structure.
LW ASA = (0,209; 0,290; 0,237; 0,069; 0,194) – Amount of Statistical Analysis
LW TC = (0,216; 0,268; 0,261; 0,090; 0,165) – Tool Complexity.
LW MF = (0,227; 0,162; 0,071; 0,116; 0,295; 0,129) – Main Focus.

Table 2 presents the elements of Global Weight (GW) which store the performances (priorities) of the alternatives in the light of the Main Focus.

Table 2. Global Weights in five alternative model.

Alternatives	GW
R Project	0,346
SAS	0,190
SPSS	0,186
Statistica	0,110
S-Plus	0,168

The use of Consistency Reason (CR) allows one to evaluate the inconsistency regarding the judgment matrix order. As far as it concerns the analyzed matrices, CR value was under 0,1, therefore, it is correct to say the judgments were consistent.

5. Four alternative model

Subsequently, the option R Project was removed from the alternatives regarding the strong influences of the criteria Product Price and Maintenance Cost over the final results. Comparing a free product to others which are paid for was not compatible for this analysis. So, to avoid the order inversion problem, the questionnaires were developed considering four alternatives (SAS, SPSS, Statistica and S-Plus). Again the survey was sent to the same specialists to gather their opinion. Based on the calculation of Geometric Means, it is possible to normalize the results to obtain Local Weight:
LW $P = (0,110; 0,224; 0,390; 0,277) – Product Price.
LW $M = (0,118; 0,274; 0,286; 0,322) – Maintenance Cost.
LW TT = (0,138; 0,395; 0,244; 0,222) – Training Time.
LW RPS = (0,152; 0,279; 0,251; 0,318) – Required Physical Structure.
LW ASA = (0,480; 0,273; 0,081; 0,166) – Amount of Statistical Analysis.
LW TC = (0,330; 0,307; 0,156; 0,207) – Tool Complexity.
LW MF = (0,197; 0,138; 0,101; 0,110; 0,324; 0,130) – Main Focus.

The elements of Global Weight store performances (priorities) of alternatives in the light of the Main Focus, as shown on Table 3.

Table 3. Global Weights in four alternative model.

Alternatives	PG
SAS	0,267
SPSS	0,281
Statistica	0,215
S-Plus	0,237

As the calculated Consistency Reason was equal to zero to all criteria as well as to the Main Focus, the consistency of the matrix can be considered very good.

6. Final Considerations

The results achieved in this work indicated that R Project obtained a higher priority due to the criteria Product Price and Maintenance Cost, which can be easily understood by the fact that is a free software product. So, the next option was to create a new analysis line, excluding R Project of the roll of alternatives. The same specialists were involved in both phases. The work rate to recollect the data was smaller in the second round, it is believed that this was a consequence of the previous experience of the specialists in dealing with the used survey form. Therefore, to each model, the results obtained were:

- Analysis 1:
  - *Analyzed software products:* R Project, SAS, SPSS, Statistica and S-Plus.
  - *Prioritized software:* R Project.
  - *Outcome explanations:* R Project is a free product, which determined the best results in the two criteria related to cost and consequently increased significantly its priority.

- Analysis 2:
  - *Analyzed software products:* SAS, SPSS, Statistica and S-Plus.

- *Prioritized software:* SPSS.
- *Outcome explanation:* SPSS has the best results in Training Time and the second best in Required Physical Structure, Amount of Statistical Analysis and Tool Complexity.

Regarding the consistency of the results, the analysis pointed out that judgments and results were quite satisfactory, fulfilling all the model requirements.

REFERENCES

Costa, T. C., Belderrain, M. C. N. (2009) Decisão em Grupo em Métodos Multicritério de Apoio à Decisão, *Anais* do 15º Encontro de Iniciação Científica e Pós-Graduação do ITA, SP.

Decision Support Systems Glossary. Available at: <http://dssresources.com/>. Access in July 2009.

Ehrlich, P. J. (2004). *Procedimentos para Apoio às Decisões: Gestão dos Riscos e de Objetivos Conflitantes*. Escola de Administração de Empresas da Fundação Getúlio Vargas, SP, 2004.

Forman, E., Peniwati, K. (1998). Aggregating Individual Judgments and Priorities with the Analytic Hierarchy Process. *European Journal of Operacional Research*, n.108, pp. 165-169.

Freitas, A. L. P., Trevizano, W. A., Costa, H. G. (2008). Uma Abordagem Multicritério para Problemas Decisórios com Múltiplos Grupos de Avaliadores. *Investigação Operacional*, 28, pp. 133-149, Lisboa, Portugal.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C. (2005). *Análise Multivariada de Dados* – 5 edição. Porto Alegre – Bookman, 2005.

Lurie, P. M. (1995) A Review of Five Statistical Packages for Windows. *The American Statistician*. EUA, v. 49, n. 1, pp. 99-107.

Pasw Statistics. Available at: <http://www.spss.com/statistics/>. Access in: July 2010.

Ripley, B. D. (1997). *S-PLUS for Windows version 4.0*. EUA- Maths&Stats Newsletter.

Saaty, T. L. (1980). *The Analytic Hierarchy Process*. EUA, Nova Iorque- McGraw-Hill.

Saaty, T. L., Peniwati, K. (2007). *Group Decision Making: Drawing out and Reconciling Differences*. Pittsburgh: RWS Publications.

Saaty, T. L. (2008). Relative Measurement and Its Generalization in Decision Making, Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors, The Analytic Hierarchy/Network Process, *RACSAM - Review of the Royal Spanish Academy of Sciences, Series A, Mathematics*, Vol. 102 (2), pp. 251-318.

SAS Brasil. Available at: <http://www.sas.com/offices>. Access in: July 2010.

Statsoft. Available at: <http://www.statsoft.com/>. Access in: July 2010.

The Comprehensive R Archive Network. Available at: <http://cran.r-project.org/> - Access in: July 2010.

Vaidya, O. S., Kumar, S. (2006). Analytic Hierarchy Process: An overview of applications. *European Journal of Operational Research*, v. 169 (1), pp. 1-29.