ADVANTAGE OF APPLYING THE R LANGUAGE IN DATA MINING TECHNIQUES

Sonja Pravilović[1,2], Annalisa Appice[2]

[1] ***Montenegro Business School, "Mediterranean" University, Podgorica, Montenegro***

[2] ***Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Bari, Italy***

Abstract: *The development of technology, computers and the Internet has significantly contributed to easier organization of data, which would also become more useful if turned into information and knowledge. Knowledge can refer to the users of products and/or services, or the market in which the company operates. From the perspective of the enterprise, the business intelligence system places the primary emphasis on the users of products and services. Data mining combines concepts, tools and algorithms of machine learning and statistics to analyze very large data sets, so as to gain insight, understanding and effective knowledge, and it is applied for this purpose in many organizations. In the recent years, the market of data mining tools has become more and more flooded, with more than fifty commercial tools. The tools like SPSS PASW Modeler (Clementine), Excel, Rapid Miner, SAS, SAS Enterprise Miner and many others have been much more used until recently. However, the R language is increasingly taking over this role today. R is statistical software, and an object-oriented high-level programming language used for data analysis, which includes a large number of statistical procedures such as t-test, chi-square test, standard linear models, instrumental variables estimation, local regression polynomials, etc. The R language has built-in functions for the nearest neighbor method allowing the automatic classification, the asociation rules showing the connection probability between two or more events, decision tree models, numerous methods of single and multiple regression, and many others, which makes it a very high-quality tool in data mining techniques.*

1 INTRODUCTION

The development of technology, computers and the Internet has significantly contributed to easier organization of data, which would also become more useful if turned into information and knowledge.

Knowledge can refer to the users of products and/or services, or the market in which the company operates. From the perspective of the enterprise, the business intelligence system places the primary emphasis on the users of products and services. The knowledge that can ensure the company its survival on the market relates to understanding customers and their needs and the fact is that the most desirable type is a loyal customer, from the perspective of long-term client cooperation and expected future financial results of the company. This is an obvious connection between the knowledge economy and the success in the market.

The methods of data mining have enabled the process of improving decision-making processes at the strategic business level and providing hidden data through Business Intelligence (BI) methodology.

The methods of data mining discover relations, logic, correctness, and generality of any structure in data.

The term data mining is often identified with two distinct processes: detection and prediction of knowledge. The process of knowledge discovery implies understanding the explicit information that is vital to have in a readable format, while prediction refers to future events.

Business intelligence is currently very topical in the world and represents a continuation of the information decision support system with which it is closely related. The beginning of development is closely linked to automation of business process in the company. Different transaction systems have proved to be very high quality generators of large amounts of data that made the "explosion" of data. So, the creation of a huge database needed to be simple or easier to access. A growing awareness of the usefulness of information has contributed to the development of the new discipline called Business Intelligence (BI).

Data mining is a set of techniques and methods relating to the extraction of knowledge from large amounts of data (through automatic or semi-automatic methods) and further scientific, industrial or operational use of that knowledge. Data mining is closely related to the statistics as an applied mathematical discipline with an analysis of data that could be defined as *the extraction of useful information from data.*

The common methods of data mining are: predictable (classification, regression analysis, time series analysis, forecasting, ...), descriptive (clustering, sumarization, association rules, strings discovery, etc.) and new methods derived from data mining (decision tree, neural networks, genetic algorithms, text mining, and many others).

The most commonly used methods derived from data mining are: 1) the ***k*** nearest neighbor method that allows automatic classification; 2) asociation rules in the „if-then“ form, indicating the likelihood of an event to bind to another, 3) decision tree, which operates on the basis of decision-making based on familiar situations and decisions, 4) neural network designed to work similar to the human brain and used in risk analysis and forecasting; 5) genetic algorithms are based on the imitation rules of biological development using the optimization and machine problem solving.

Classification analyzes data sets, reveals hidden connections and defines elements (functions) to group them into one of several classes. Data association defines characteristics that occur together with more samples, i.e. connections between arbitrary attributes. Grouping (Clustering) is the process of determining which data groups are similar, but different from other groups of data and identifying variables for exercising the best grouping.

The complexity of business problems require procedures that include steps ranging from the preparation of data to the interpretation of obtained results and they could include:

- adding a new and updated data base (various analysis and data collection),
- cleaning and preparing the data for analysis (elimination of extreme values (outlayers),
- transformation and aggregation of data for analysis by periods,
- carrying values, indicators, etc.

Therefore, the market competition is getting stronger and developed distribution channels and supply of goods and services have brought about the application of business intelligence tools through which companies seek to effectively connect people with their businesses, with customers, suppliers and partners, enabling business users to access vast amounts of complex data.

The basic tools of business intelligence include:

- tools for queries (OLAP),
- data mining tools, and
- visualization tools (Dashboard / Scorecard tools).

Procedures for data mining can be conducted with the help of three technologies:

- multiprocessing computer technology,
- technology for massive data collection,
- algorithmic techniques.

Data mining combines concepts, tools and algorithms of machine learning and statistics to analyze very large data sets, so as to gain insight, understanding and effective knowledge, and it is applied for this purpose in many organizations.

In essence, data mining is a mathematical analysis carried out on large databases. The term data mining became especially popular in the 90's, and today it has a double meaning:

- using advanced analytical techniques to extract implicitly hidden information knowledge already structured in data in order to make it available and directly useful,
- research and analysis of large amounts of data, performed automatically or semi-automatically, with the aim of discovering significant patterns.

In both cases, the concepts of information and its meaning are closely related depending on the domain of data mining and the application of these data in a particular area.

The only difference between the two disciplines is that data mining is a new discipline that is related to significant or large data sets.

Today, data mining is a crucial activity in many areas of scientific research, but also in other areas (for example, in market research, economics, finance, medicine, agriculture, meteorology, etc.). In the professional world, it is used to solve various problems ranging from customer relationship management (CRM), fraud identification, consumer behavior, web pages optimization, etc.

The main factors that contributed to the development of data mining are:

- large amounts of data in the electronic form,
- cheap data storage,
- new methods and techniques of analysis (machine learning, pattern recognition).

Data mining techniques are based on specific algorithms. Patterns can be identified, the starting point for new hypotheses can be set, and then the causal relationship between the events can be tested, which can be further used in a statistical sense for the production and prediction of new data.

Among the most commonly used techniques are the following:

- grouping;
- neural networks,
- decision trees,
- association analysis (identification of simultaneously purchased products), etc.

2. ADVANTAGE OF R LANGUAGE AS A TOOL OF DATA MINING TECHNIQUE

There are numerous packages for statistical data processing: SAS, SPSS, Stata, R. Certainly, it is a very good idea to have a number of tools available, but in the end they users themselves should decide which of them to apply. The technical reports of the strengths and weaknesses of these packages rarely take into account the accuracy in choosing the statistical software. In some models, such as nonlinear regression, it is the accuracy that will have a bigger problem of practical importance than others.

In recent years, the market of data mining tools has become more and more flooded, with more than fifty commercial tools. The tools like SPSS PASW Modeler (Clementine), Excel, Rapid Miner, SAS, SAS Enterprise Miner and many others have been much more used until

recently. However, the R language is increasingly taking over this role today.

With data mining procedures it is possible to identify the following types of information in the existing data: classes, clusters, or categories, associations, sequences and make forecasts. Today there is a large number of statistical software packages, such as: SAS (Statistics, SPSS, etc.); mathematical software packages (Matlab, Mathematica); tools included in the data warehouse (OLAP) or database management system (Microsoft SQL Server Business Intelligence - including Enterprise Miner), and specialized tools for general use or for business use (DataMiner, IntelliMiner, etc.).

In today's modern business environment, when data cease to be poor resources, with the help of information technology, the emphasis needs to be placed on methods, methodology and algorithmic procedures extracting knowledge that is hidden in this abundance of data.

R is statistical software, and an object-oriented high-level programming language used for data analysis, which includes a large number of statistical procedures such as t-test, chi-square test, standard linear models, instrumental variables estimation, local regression polynomials, etc.

With procedures and functions incorporated in the R language it is possible to identify the following types of information in the existing data: classes, clusters, or categories, associations, sequences, and make forecasts.

Data mining techniques allow easier analysis, and a high level of knowledge of analytical skills is highly attractive and sought after in many successful companies.

Today, the R language represents an ideal solution for many challenging tasks associated with a data mining and business intelligence. R language provides breadth and depth in computational statistics, and much more than what offer other commercial closed source products. R is primarily a programming language for highly qualified statisticians.

R is a statistical software, and an object-oriented high-level programming language used for data analysis, which includes a large number of statistical procedures such as t-test, chi-square test, standard linear models, instrumental variables estimation, local regression polynomials, etc. Besides, R provides high-level graphics capabilities.

R is an object-oriented programming language. This means that everything what is done with R can be saved as an object. Every object has a class. It describes what the object contains and what each function does. For example, plot (x) does not give the same output if the regression result is x or vector.

It should be noted that companies have cheaper data mining software available today. Some of the most popular are IBM intelligent Miner, Oracle Darwin, SAS Institute's Enterprise Miner and SPSS Clementine. But, the price of these programs can range from tens of thousands of dollars (for more complex ones) up to several million dollars. A new generation of data mining software requires neither the engagement of experts, nor detailed knowledge of statistics by the managers. Its use and application is still quite simple.

The objective of the R style as a programming guide is easier reading, using and verification of the code.

From the mentioned method of data mining, the R language has built-in functions for the *k* nearest neighbor method allowing the automatic classification and clustering (Figure 3), association rules showing the probability of connection between two or more events, models of decision trees (Figure 4), numerous methods of single and multiple regression, and many others.

For example, k-means method aims to partition the points into *k* groups so that the sum of squares from the points to the assigned cluster centres is minimized.

The clustering technique allows grouping of similar data. Grouping actually means sorting the elements into the set, which has the greatest similarity data (customer segmentation - by age, occupation, income, consumption ..). The division must meet two criteria: 1) each group is homogeneous (similar data), and 2) each set must be distinguished from other sets (a significant difference).

The decision trees are very popular method for classification and decision making. There are based on the relationship between the strategies and conditions used to solve problems in finance, banking, marketing, insurance. It predicts the outcome by using a series of questions and rules for categorizing data. Decision tree branching occurs as a consequence of the fulfillment of conditions of classification issues dividing the data into subsets that are more homogeneous than the senior set. If the question has two answers, then the response to the question produces two subsets (binary tree). Decision tree learning, used in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data not decisions. The R language has built-in function that form and display result model tree in a very simple way.

3 CREATION OF R LANGUAGE

The elegant and widely accepted S language, as a permanent software system with an outstanding comprehensive conceptual solution is the result of the effort and hard work of John Chambers. In 1998, the Association for Computing Machinery (ACM) presented it with its Software System Award as "the S system" and forever changed the way people would analyze, visualize and manipulate data. R is inspired by the S environment developed by John Chambers, and significant contributions of Douglas Bates, Rick Becker, Bill Cleveland, Trevor Hastie, Daryl Pregibon and Allan.

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics, University of Auckland, New Zealand. After that, a large group of

individuals contributed to R by sending codes and so-called bug reports. The current R is a result of a collaborative effort with contributions from around the world. In mid-1997 a core of the group known as the R Core Team was formed, whose members can change the archive of the R source code.

R is a dialect of the S language. The S language was developed at Bell Labs as a computing environment for data analysis and graphical display. The graphical display and interactivity, which can be found in R, are powerful tools for data research since they provide an understanding of data in the easiest way.

4 R LANGUAGE

In the last ten years, the language R statistics has exploded in popularity and functionality and has become an election tool of scientists around the world. Today, R is used by more than 2,000,000 analysts. Since it has completely unveiled its elegance and power to the academic community so far, the members of the academic community have embraced the R language for solving their most challenging problems in the fields ranging from computational biology, quantitative finance to training people in these fields.

The result is the explosion of the R analytics and applications, which has led to enthusiasm and acceptance of R as a primary analytical method in companies such as Google, Facebook and LinkedIn. R is now available to everyone and is becoming a powerful and revolutionary support in any workplace, improving productivity.

Each technique of data analysis is now at hand instead of limiting the analysis and features with paid added modules. R encompasses virtually all data manipulations, statistical models, dates, and charts, that is, everything that a modern scientist needs. Even when the predictive model doesn't give excellent results, it is simple and easy to find help by assuming and using the most modern community of reviews of the methods in statistics and predictive modeling of leading scientists, absolutely free of charge.

To make a beautiful and unique visualization of the date by presenting the complex data and charts is just one of the essential elements of the process of data analysis. R surpasses the traditional bar charts and line plots, and facilitates the extraction of meaning from multidimensional data with a multi-panel scale, 3-D surfaces and much more.

Instead of using the point-and-click menu or inflexible black box procedures, R is a programming language designed specifically for data analysis. Experienced programmers of R have created a data analyzer which is faster, more efficient and flexible than the inherited user-friendly statistical software in creating the mix and match models for the best results. The R code is automated and easily repeatable in research and implementation. [4]

As a successful open-source project, R is supported by a community of more than 2 million users and thousands of programmers around the world. Whether the use enables optimizing portfolios, analyzing genomic sequences or predicting a certain component in the failure, experts in every field have made resources, applications and the code available online free of charge.

R runs on Windows, Macintosh, Linux and Unix platforms and the installation is free of charge and easy through http://cran.r-project.org/ . However, although the R language has a lot of benefits it is difficult to learn for users who have had no experience in programming. For users who have worked with data in other programming languages, switching to R makes everything incredibly easier and faster.

Working in R can sometimes be problematic, since it requires that all objects are stored in memory, indicating the limited size of data sets. Optimization routines in R such as *optim* or *nlm* require passage through a function whose argument is a vector of parameters (eg. log-likelihood). However, the function object may depend on a variety of other things in addition to its parameters (data). In writing the optimization code, the existence of bonding parameters is desirable in order to allow the access to the user.

R has functions to access databases. A typical solution with large data sets is the storage of data in the database and the entry of those that are necessary to R, as needed. Another weakness is the lack of coherent documentation that covers the entire R.

The basis of the R language consists of: data entry, dataframes, graphical display of 2D and 3D data, tables, mathematics and mathematical functions, the classical tests of data mining, statistical models, regression, analysis of variance, covariance, general linear models, count data, count data in the tables, the proportions of data, binary variable, general models, mix-effect models, tree models, time series analysis, multivariate models, spatial statistics, simulation models, and many others. [1]

R is a high-level language with the environment designed for the analysis and graphical representation of data. Here, the term "environment" is used to characterize it as a fully planned and coherent system, and not a system that is gradually supplemented with specific and inflexible programming tools, which is very often the case with other data analysis programs.

R can be easily and considerably expanded by the installation through packages or libraries. There are more than 3000 packages in the Cran data warehouse (repository). The R language design was influenced by the schemes of two existing languages: Becker, Chambers and Wilks's and Sussman's.

The existing R language is very similar in appearance to S, but sublevels of implementation and the semantics are derived from this scheme.

R certainly has a very rich environment that can be enjoyed by beginners, intermediate-level users and experts in disciplines ranging from scientific activities,

economics, finance, sociology, political science, agriculture, medicine and engineering.

R includes the packages for preparation, processing, data display, graphics, mathematical functions, a wide range of statistical techniques, from the basic conventional tests, through regression and analysis of variance and general linear modeling, to more specialized topics such as spatial statistics, multivariate methods, tree models, mix-effective models and analysis of time series, as well as many others.

The idea is to offer R to users with very little knowledge of statistical theory, assuming that they do not know the basics of mathematics and / or statistics in order to assist them in their assumptions behind the tests, and encoure a critical approach to statistical modeling.

The question that may be asked is why users should begin to deal with R instead of implementing a perfectly appropriate statistical package to solve their problems? If the intention is to implement a very limited range of statistical tests, and not to do more (or otherwise) in the future, then it is absolutely fine not to switch to R.

However, the main reason for switching to R is to take advantage of the coverage and availability of new applications where R is the best in the areas such as effective generalized mixed models, as well as other general and additional models.

Another reason for learning R may be the desire of users to understand the literature, as more people in various fields of sciences publish their research results in the context of R. Thirdly, if some research is made in the field of data mining, then it is very easy to notice that today the best known researchers in this discipline have become highly specialized in using R. In addition, a large number of the world's leading statisticians use R, so of course, this also contributes to the importance of knowing R.

Another very important reason for learning R is a quality of back-up and support. There is a premium network dedicated to research and Web wizards are eager to answer the questions of users. If the user intends to invest enough effort in becoming a good computer statistician, the structure of R and the easiness of writing one own's functions are the main attractions. The last, but certainly not the least important reason is that the R product, as one of the best integrated software in the world, is available for free.

Resampling methods (bootstrapping, random permutation tests, etc.) are very useful and easy to operate (function within the for loop). In addition, it enables a direct access to packages through the network or via the mailing list. Eg: through install.packages (name) you get a package called name (assuming that the computer running R is connected to the Internet), which makes R good, in collaboration with other programs, and widely available.

5 ORGANIZATION OF R LANGUAGE

The "official" R consists of several packages that are created by the core R team. In addition, there are hundreds of packages that have been enriched by the users. Some of these packages represent the latest statistical research library. Most statistical research is first conducted in the R language.

R is not supported in the same way as the commercial software, but many users are finding better support through the R-help or through the network, rather than from commercial enterprises. Users will often use R because of its simple and clear graphical environment.

Many users use R as a statistics system, although it is much more than that. The R environment incorporates classical and modern statistical techniques and / or packages and libraries.

There is a significant difference in approach between S (toward R)and other statistical systems. For example, in S, a statistical analysis is usually conducted in a series of steps, storing the intermediate results of individual steps in objects. [3]

For example, SAS and SPSS give more extensive results from a regression or discriminant analysis. R, however, gives the minimal output and stores the results within certain objects, so they can be accessed by further functions for subsequent interrogation.

R is started within the graphical environment (Figures 1 and 2 and through a windowing system, and the programming code can be written directly within the window or copied from another editor. The results and code, input or output data, can be saved and loaded in various forms, such as. *.txt, *.dat. *.r data files.

Since R has a rich set of facilities and the programming language, it is, of course, difficult to master for the customers, but once learned, it is very easy to extend, amend, enhance, and apply in another field or find a simpler solution in R.

It is particularly difficult for users who are accustomed to working within the framework of statistical packages, since they have high expectations. There are hundreds of R packages, but only a few of them may be suitable for a given problem for a particular client.

Individual users have a much easier task. They need some basic knowledge of R, and then they need to learn some specific techniques that are related to the problems in their area. After solving the problem, the user can only retain an insight into how to solve the problem and try new packages within R that could help achieve even simpler solution. [2]

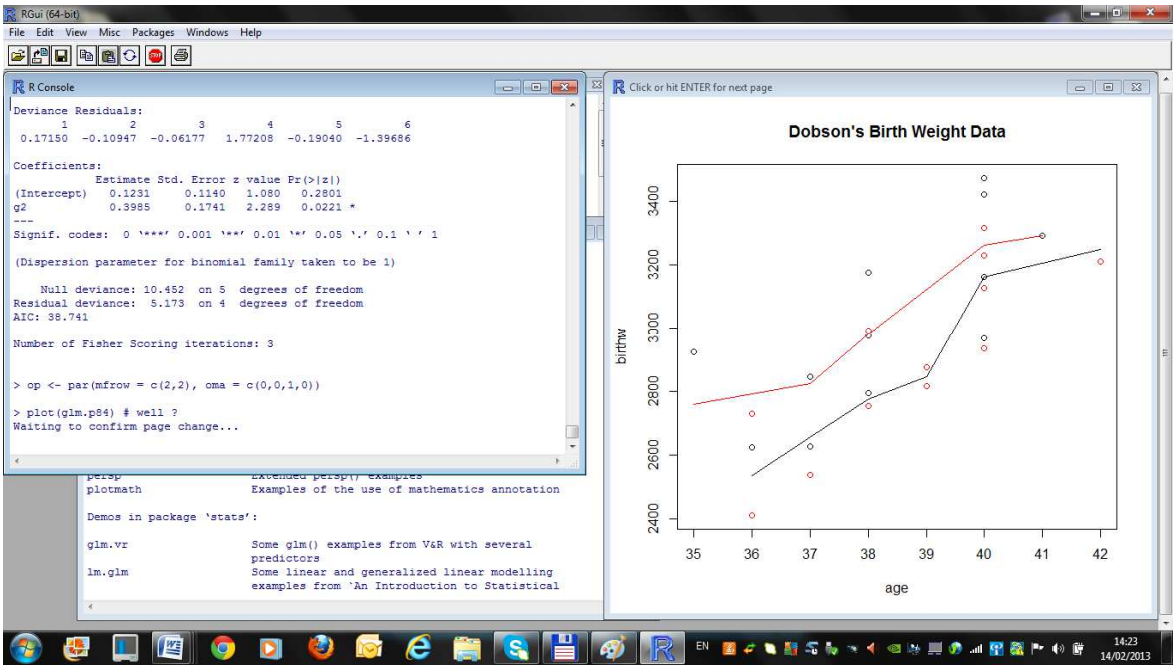

Figure 1. Display of the window in R

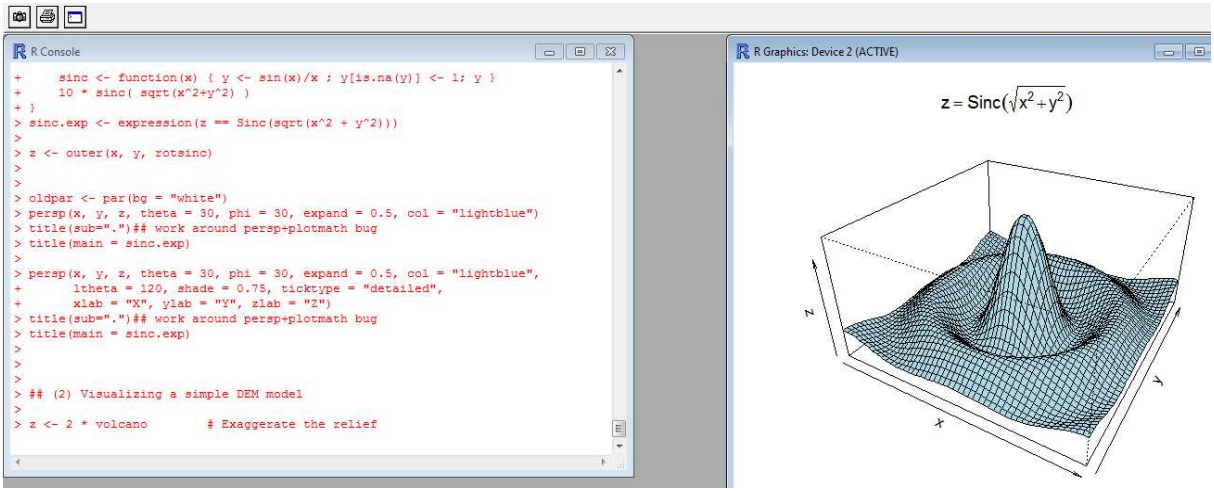

Figure 2. Display of the 3D function in R

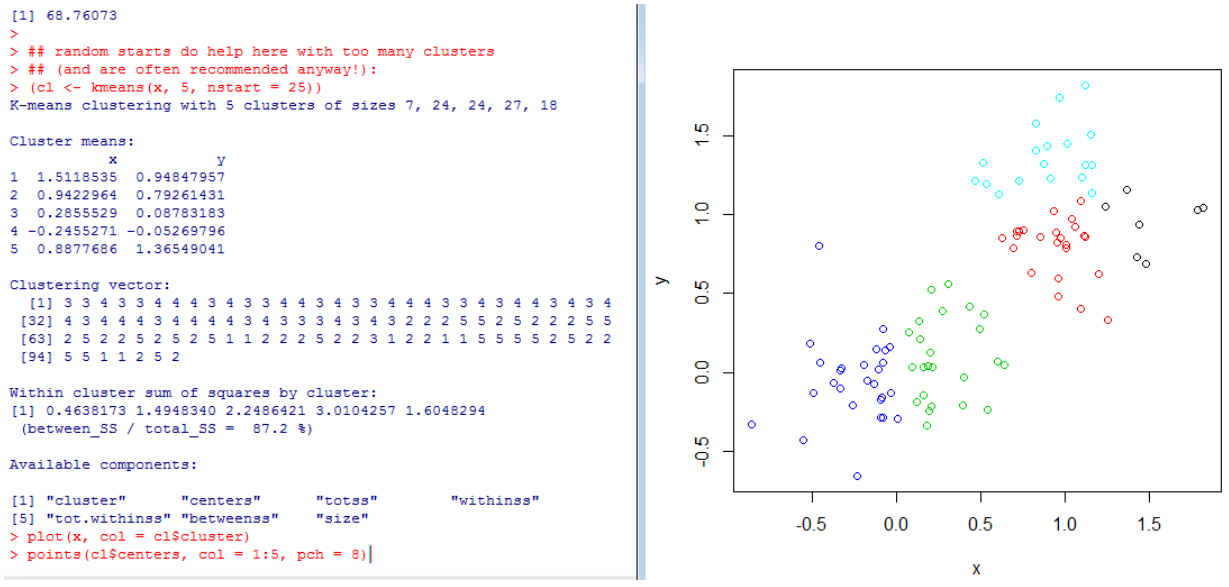

Figure 3. Display of k-means function in R

```
1) root 209 43.12000 1.753
  2) cach < 27 143 11.79000 1.525
    4) mmax < 6100 78  3.89400 1.375
      8) mmax < 1750 12  0.78430 1.089 *
      9) mmax > 1750 66  1.94900 1.427 *
    5) mmax > 6100 65  4.04500 1.704
     10) syct < 360 58  2.50100 1.756
       20) chmin < 5.5 46  1.22600 1.699 *
       21) chmin > 5.5 12  0.55070 1.974 *
     11) syct > 360 7  0.12910 1.280 *
  3) cach > 27 66  7.64300 2.249
    6) mmax < 28000 41  2.34100 2.062
     12) cach < 96.5 34  1.59200 2.008
       24) mmax < 11240 14  0.42460 1.827 *
       25) mmax > 11240 20  0.38340 2.135 *
     13) cach > 96.5 7  0.17170 2.324 *
    7) mmax > 28000 25  1.52300 2.555
     14) cach < 56 7  0.06929 2.268 *
     15) cach > 56 18  0.65350 2.667 *
> summary(cpus.ltr)

Regression tree:
tree(formula = log10(perf) ~ syct + mmin + mmax + cach
    chmax, data = cpus)
Variables actually used in tree construction:
[1] "cach"  "mmax"  "syct"  "chmin"
Number of terminal nodes:  10
Residual mean deviance:  0.03187 = 6.342 / 199
Distribution of residuals:
      Min.    1st Qu.     Median       Mean    3rd Qu
-0.4945000 -0.1191000  0.0003571  0.0000000  0.1141000
> plot(cpus.ltr);  text(cpus.ltr)
```

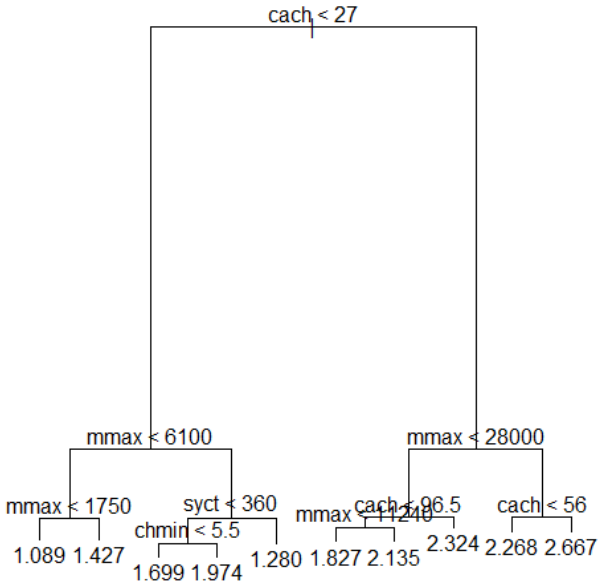

Figure 4. Display of the model tree function in R

Functions in the research are "first class objects", which means that they can be treated like any other R object. It is important that the results of the function can be input arguments in new functions. Functions can be nested, which means that they can be contained within one another. The output value of the function is the last expression in the body of the function. The object functions can be "created" so as to contain all the arguments necessary for evaluating the function.

There is no need for the transfer of long arguments of the list, which are useful for interactive and research work within the program. The code can be relieved and cleaned using comparametrisation. R has a vast amount of information available, is rich in the user databases and functionality. In addition, R is very suitable for the creation of reports on daily, weekly, yearly schedule, while the level of documentation is much higher than the average open source software, even than some commercial packages (eg, SPSS).

CONCLUSION

The aim of this study was to point to the application of modern programming languages and statistical packages without which modern science and research work in many areas of economics, finance, medicine, meteorology, engineering, and data mining cannot be imagined today.

Application of R as a programming language and statistical software is much more than a supplement to Stata, SAS, and SPSS.

Although it is more difficult to learn, the biggest advantage of R is its free-of-charge feature and the wealth of specialized application packages and libraries for a huge number of statistical, mathematical and other methods.

R is a simple, but very powerful data mining and statistical data processing tool and once "discovered", it provides users with an entirely new, rich and powerful tool applicable in almost every field of research.

REFERENCES

[1] Crawley, M.J. *The R book*, Imperial College London at Silwood Park, UK, John Wiley and Sons, Ltd 2007.

[2] Dalgaard, P. *Introductory Statistics with R*, New York, Springer-Verlag, 2002.

[3] Krause, A. and Olson, M. *The Basics of S and S-PLUS*, New York, Springer-Verlag, 2000.

[4] McCulloch, C.E. and Searle, S.R. *Generalized, Linear and Mixed Models*, New York, 2001.