ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2007

Εργασία: *Πρακτική Άσκηση*

Επιβλέπων Καθηγητής: *Ι. Ντζούφρας*

Θέμα:

Κατασκευή λογισμικού ανάλυσης δεδομένων ποδοσφαίρου

(Football Analyzer)

Φοιτητής

Ονοματεπώνυμο: *Νίκος Μουστακίδης* **Α.Μ.:** *6030047*

Ημερομηνία Υποβολής: *01-06-2007*

Football Analyzer

User Manual

Table of Contents

Chapter 1: Introduction

§1.1 What is Football Analyzer?

Football Analyzer is a script created under R programming environment presented with a GUI built using R's "tcltk" package in order to provide to the end user an easy way to use menu band operation. The user only needs to provide the program with a properly formatted dataset containing the football data. The program uses the data to formulate a statistical model based on the Poisson distribution describing the performance of the teams.

§1.2 What can Football Analyzer do?

Football Analyzer is a program addressed to the average user. Thus, it lacks any "advanced" features that might be of interest to a statistician, mathematician or any other seeking to delve further into the subject of football analysis. The features it hosts include:

- Future Game Prediction: The probabilities of each outcome (win, loss, draw) of any given game are estimated and presented as well as the expected number of goals for each team. Multiple games are supported.

- Predict Final Score: The user can create a virtual tournament and the final score of a single team will be calculated based on that. A detailed view can also be extracted including the probability (or odds) of each game's outcome and the expected number of goals scored by each team.

- Performance Plot: This feature enables the user to observe the performance of a selected team based on the goals scored and received in each game.

- Model Parameters: Estimated parameters of the underlying fitted Poisson model are given. This might be useful to an average level user since these parameters can be interpreted in terms of attacking and defensive abilities. Moreover, estimated home effect parameter is provided giving a quantitative feeling of the advantage each team has when playing at each home ground.

- Reproduction of Final Rank: By using this menu the user can reproduce results of a virtual tournament. The final ranking of each team is provided based on their expected scores.

Chapter 2: Setting up Football Analyzer

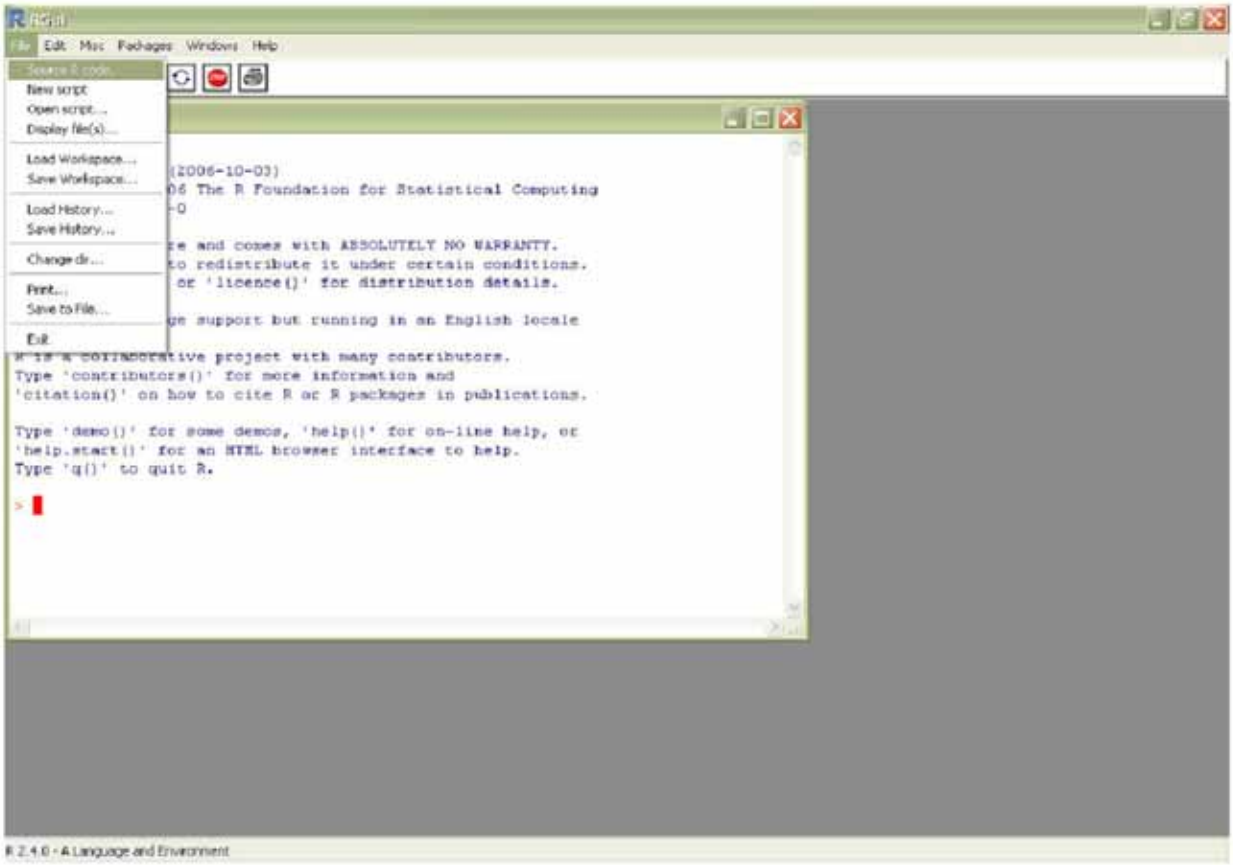
§2.1 Requirements & Installation

In order to run Football Analyzer you should have preinstalled in your computer:

- ✓ R version 2.4.0 or above and
- ✓ Package *tcltk*

To install the latest version of R visit R Project (http://www.r-project.org/). You should further download from the same site the *tcltk* package mentioned above in .zip format in case you wish to install it manually from the R menu "*Packages*"→"*Install package(s) from local zip files…*"(you can also install any package available through the R menu "*Packages*"→"*Install Package(s)…*" in case you have an internet connection available).

Once you complete these required steps, open R and go to the Menu "File" and select "Source R code" according to the following figure:



From the dialog window that appears, locate and open the football analyzer R script file. Alternatively, you can use the "`source`" command or drag and drop the script file into the command editor in R. When the execution of the script is carried out a function is created by the name football.analyzer.

After the function is installed and ready to use, we need to specify the working dataset. Data may be provided in either matrix form or data frame format. In both cases the sequence of the variables must be as given in the following table:

Name of Home Team	Name of Visitor Team	Home Team goals scored	Visitor Team goals scored
A	B	2	1
⋮	⋮	⋮	⋮

Therefore a 4-column matrix or data frame is needed. The first two columns should contain the names of the home and visitor teams and the last two should contain their respective scores (the last two variables are expected to be integer numbers. In a different case the dataset will not be accepted as valid by the Football Analyzer).

The command "`read.table`" can be used to easily import a dataset previously saved in text or text-like format (e.g. .txt or .dat) as a matrix or data frame in the R environment (for further help type in the R command editor "`help(read.table)`" without the quotes).

Example:

```
Scoreboard<-read.table("C:\\football.dat",header=T)
```

Chapter 2: Setting up Football Analyzer

§2.2 Starting the program

After following the above we must call the corresponding function by typing:

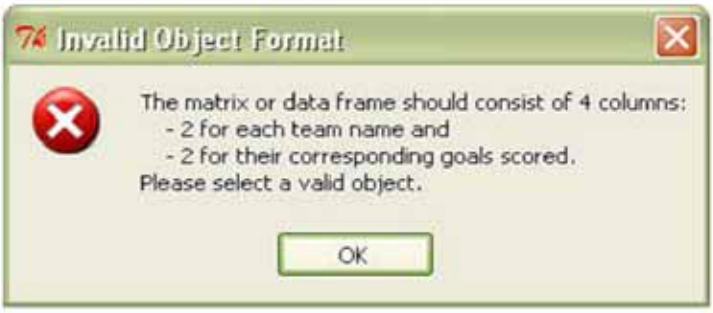
```
football.analyzer()
```

The program immediately opens an interactive window where you must select the dataset to use:

When selecting an object to use it as a dataset for Football Analyzer there are a few things that must be taken into account:

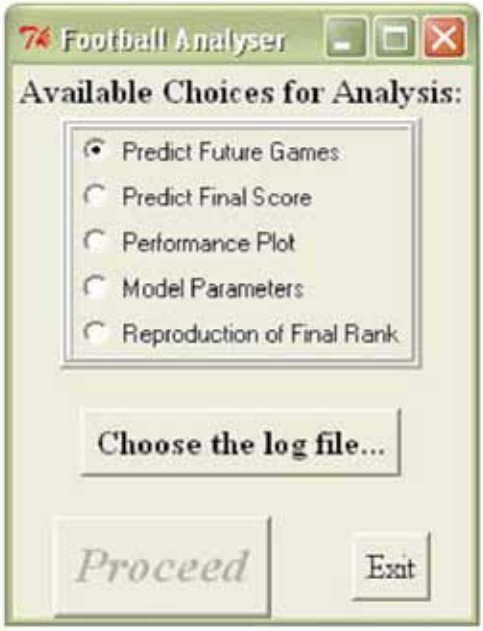
- The program will reject objects of a type other than matrix or data frame:

- It will also reject matrices or data frames that appear to have less or more than 4 columns:



- Finally, acceptable objects (4-column matrices/data frames) will be rejected if non-integer values are detected in their last two columns:

Once a valid object has been selected the current window will disappear giving its place to the main menu of the program:



Chapter 3: Exploring the main menu

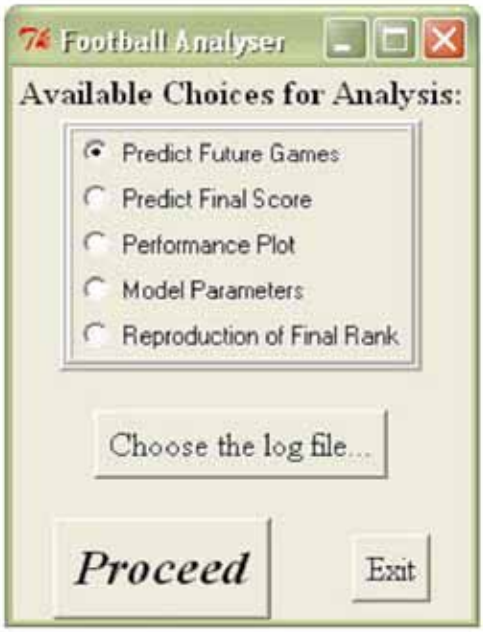
§3.1 Selecting a log file

Once a valid dataset has been chosen the main menu of the program will appear on screen:

You will notice that the main button "Proceed" is disabled so none of the program's features can be used. Instead the most prominent button is the "Choose the log file…" button. A log file must be defined to enable you to save your work at each step should you wish to do so. To choose a log file either click on the aforementioned button or press the "Return" ("Enter") key on your keyboard. In the dialog window that appears the default name of the log is composed of the dataset's name and the current date making it easier to later identify each work session saved by the program. You can also choose the file format of the log. The two available options are:

- ".falog" a custom text-like format that can be viewed by any simple text processor (e.g. Notepad). This custom format helps you better discern files created by Football Analyzer by giving them a unique extension
- ".txt" the standard simple text format

Once the log file has been defined, the layout changes enabling the use of the "Proceed" button:



Now you can select any of the options available and then click on the "Proceed" button (or you can simply press the "Return" / "Enter" key on your keyboard as above).

If you wish to change the log file at any given time click the "Choose the log file..." button again and redefine it. However, in case you cancel this definition of the new log file, the program will reset the old name and thus disable the "Proceed" button.

Besides the main menu, all other menus make use of the "Return" / "Enter" button to activate the main button (denoted in bold) of each menu as well as the "Escape" ("Esc") button to exit the current menu.

Another convenient feature is the fact that new result forms or menus deriving from the main menu do not shut down previous ones allowing for multiple features to be used simultaneously. For example one can have both the "Predict Final Score" and "Performance Plot" menus active and use both of them at the same time. You only have to select one of them, click the "Proceed" button and then without exiting the newly created menu go back to the main menu select the other menu and click the "Proceed" button once again.

Chapter 3: Exploring the main menu

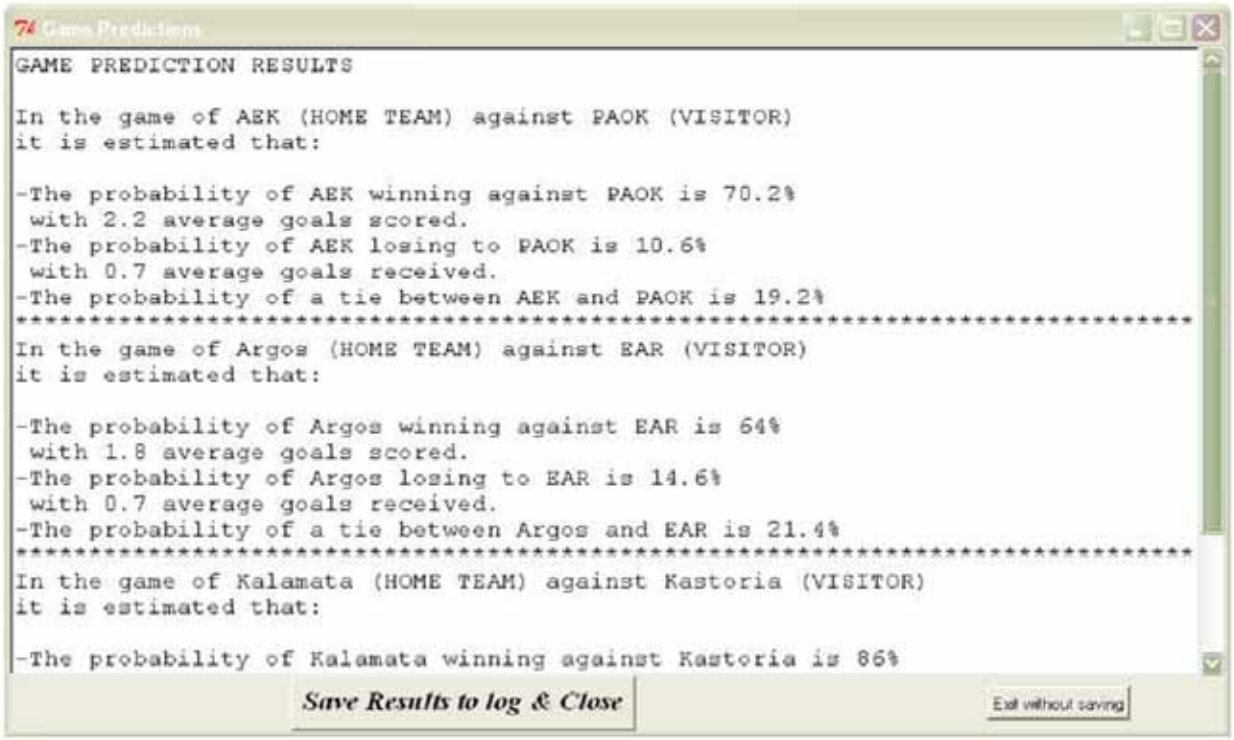
§3.2 The "Predict Future Games" menu

To activate this menu, after having specified the log file, choose "Predict Future Games" from the main menu and finally click on the "Proceed" button. A new menu will emerge:

Working within this menu, the prediction of the outcome of a game between any two teams appearing in the dataset can be estimated. Initially the main button "Make the Prediction" is disabled since no competing teams are selected. To enable it click on one of the "Available Teams" which you wish to play the role of the home team and then click on "Add/Remove to/from Home Teams list". The next step is to define the visitor team. This is done by highlighting another team from the "Available Teams" list and clicking on "Add/Remove to/from Visitor Team list". This procedure can be repeated as many times as you wish until you specify all pairs of competing teams you wish to evaluate. In case you have entered a different team than the one intended just use the appropriate "Add/Remove to/from .....Teams list" button to correct your entry. To clear the lists and start from the beginning use the "Reset Team Lists". Once all desirable games have been entered click "Make the Prediction".

Note: The program will detect games where a team is selected to compete against itself and produce an error message forbidding such an action.

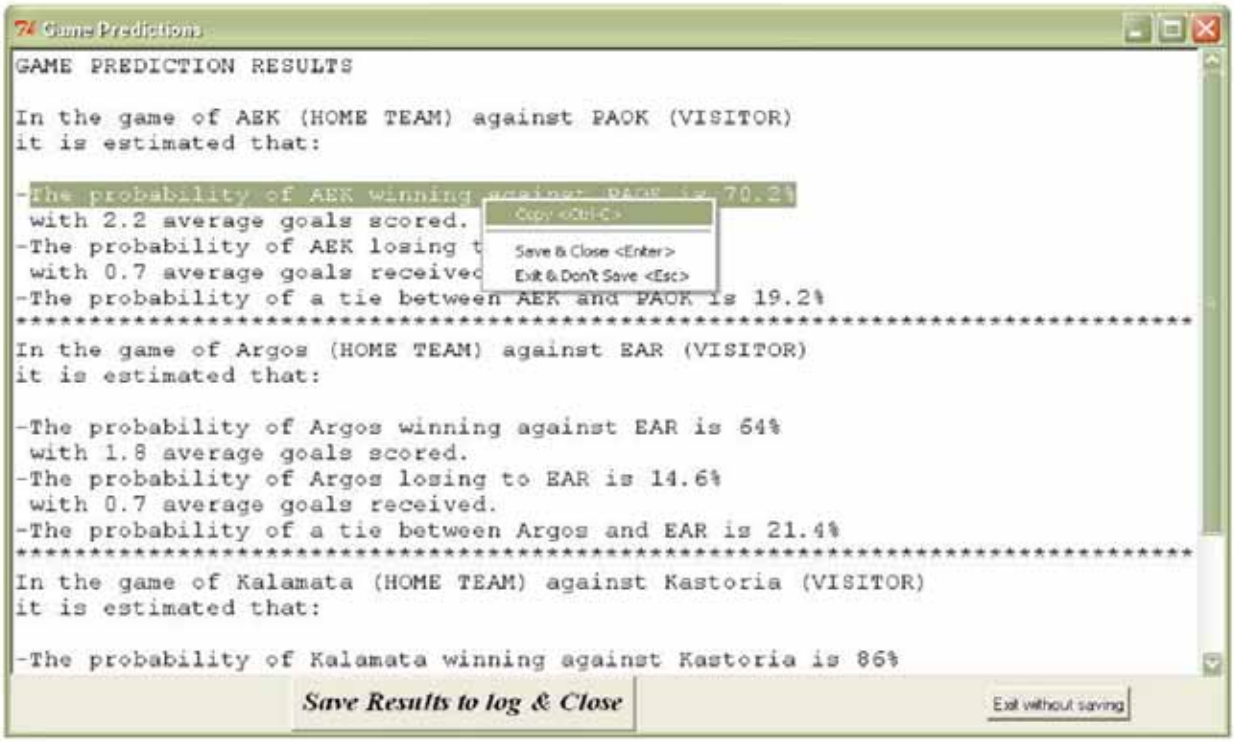
However, once a valid team matching has been provided the program will produce an output form with the estimated probabilities for each outcome:



The output presents estimations concerning the probabilities of victory and loss for the home team as well as the probability of a draw between the two teams. It also provides information on the expected number of scored goals for each team.

In the output window there are several options available:

→ You can select a part of the output text and copy it to the clipboard using the right mouse button:



→ To save to the log file use any of the following approache:

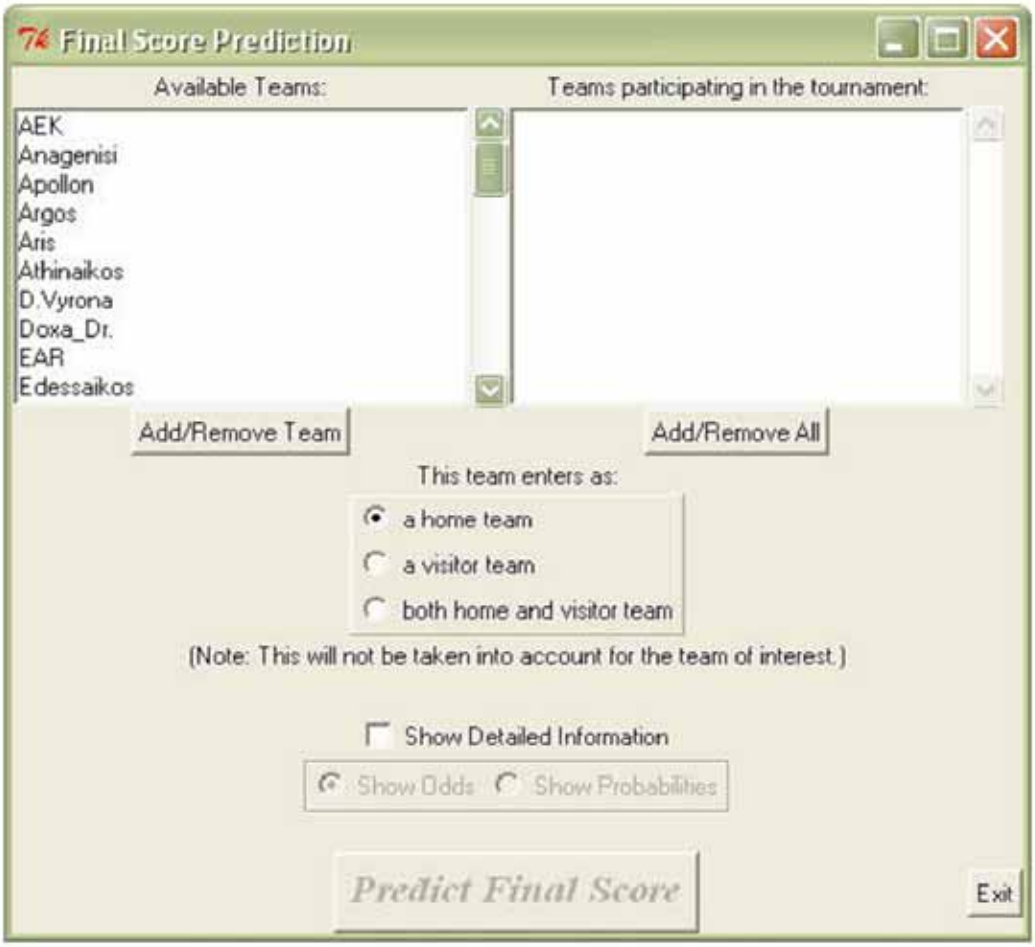
- Click the "Save Results to log & Close" button
- Right-click anywhere on the output form and select "Save & Close"
- Press the "Return" key on your keyboard

→ You can click the "Exit without saving" button OR right click anywhere on the output form and select "Exit & Don't Save". This will result in the immediate closing of the output screen without first storing the resulting text to the log file.

→ If you try to "bypass" these buttons by closing the form through the "X" window button or by pressing Alt+F4 OR by pressing the "Escape" key on your keyboard you will be prompted by a dialogue window to decide to exit the screen (or not) without saving the output to the log file.

Chapter 3: Exploring the main menu

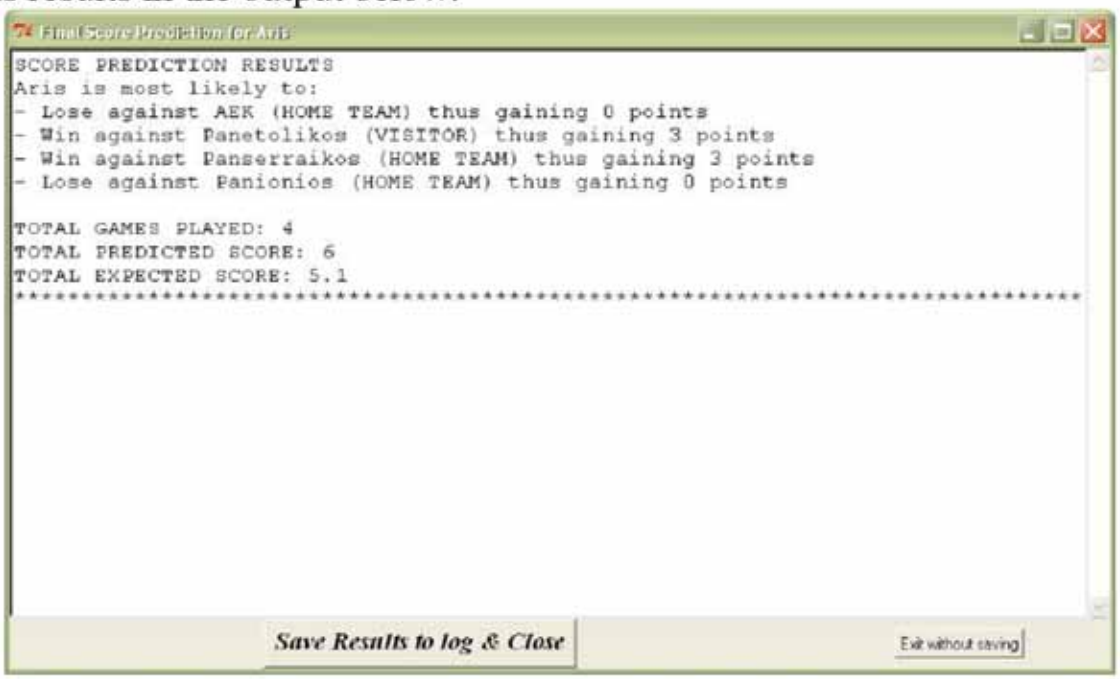
§3.3 The "Predict Final Score" menu



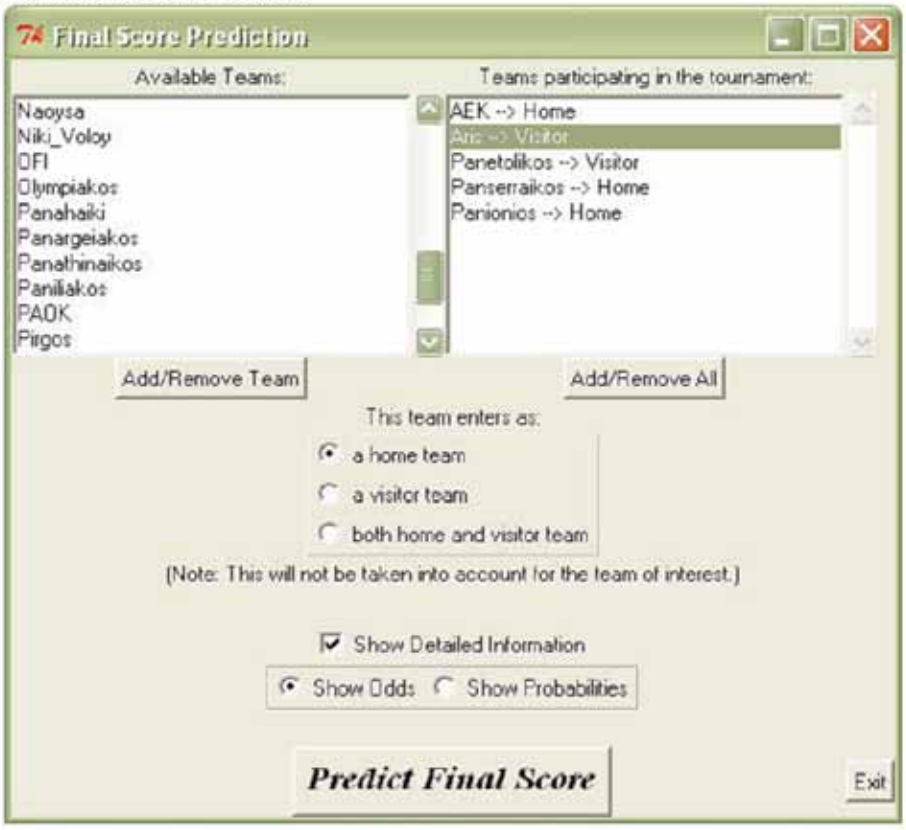
This menu featured in Football Analyzer helps you create virtual tournaments and predict the final score of a single team out of the selected ones against all the rest.

The first step is to select one by one the teams that you want to include into the virtual tournament. After selecting a team in the "Available Teams" list, toggle the radio button to properly define it to enter as a home, visitor or both as a home and visitor team. Then simply click the "Add/Remove Team" to submit the selected team. If you wish to include all available teams in your tournament, then this can be done directly using the "Add/Remove All" button. Once all teams have been selected, you need to highlight the team whose score you wish to include in your virtual tournament from the "Teams participating in the tournament" list and then click the "Predict Final Score" button or press the "Return" key on your keyboard.

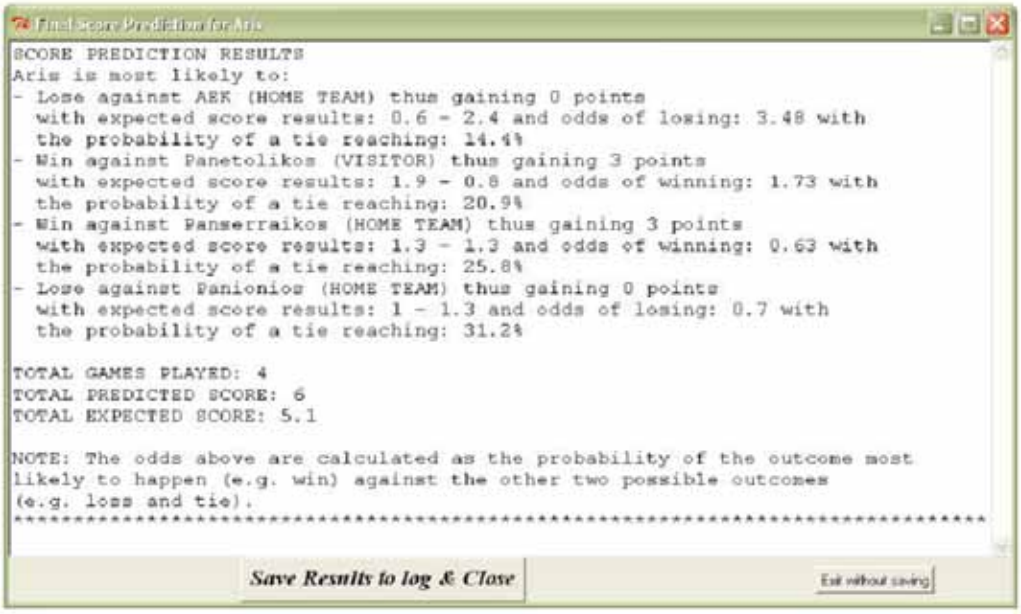
This results in the output below:



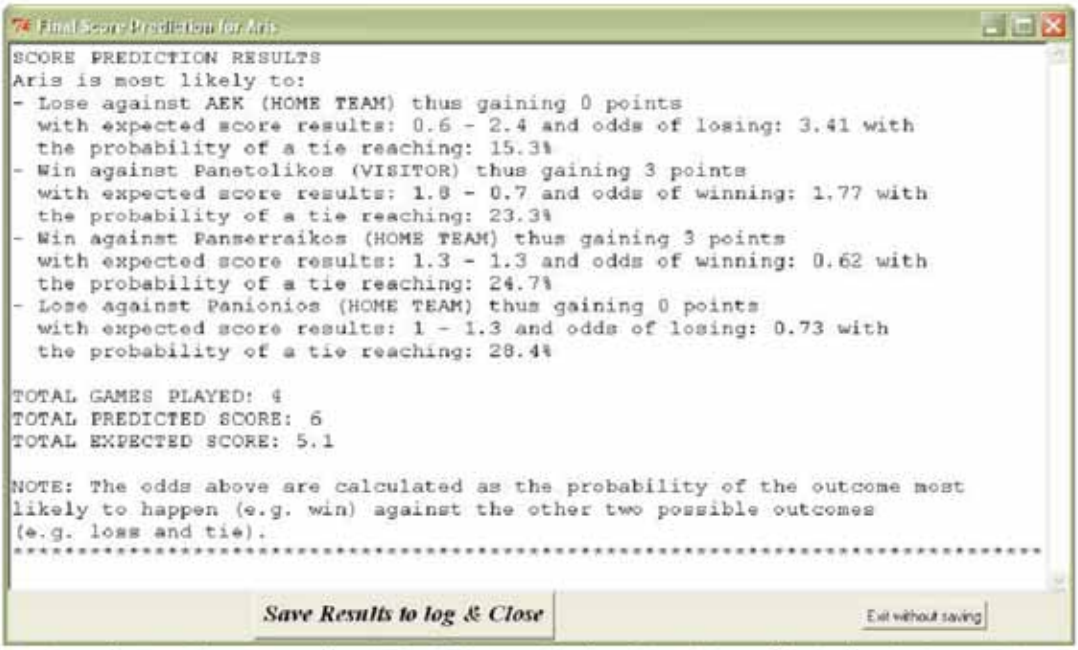
In case you wish to obtain more details activate the "Show Detailed Information" check box and then select whether you wish odds or probabilities to be calculated:



Selecting the "Show Odds" option will result to an output including information concerning odds and average goals scored:



Here we consider 3 possible outcomes (victory, loss and tie). So the odds calculated above describe the ratio of the most possible outcome against the other two and not for example victory versus loss.
Selecting the "Show Probabilities" option will result to an output including information concerning probabilities and average goals scored:



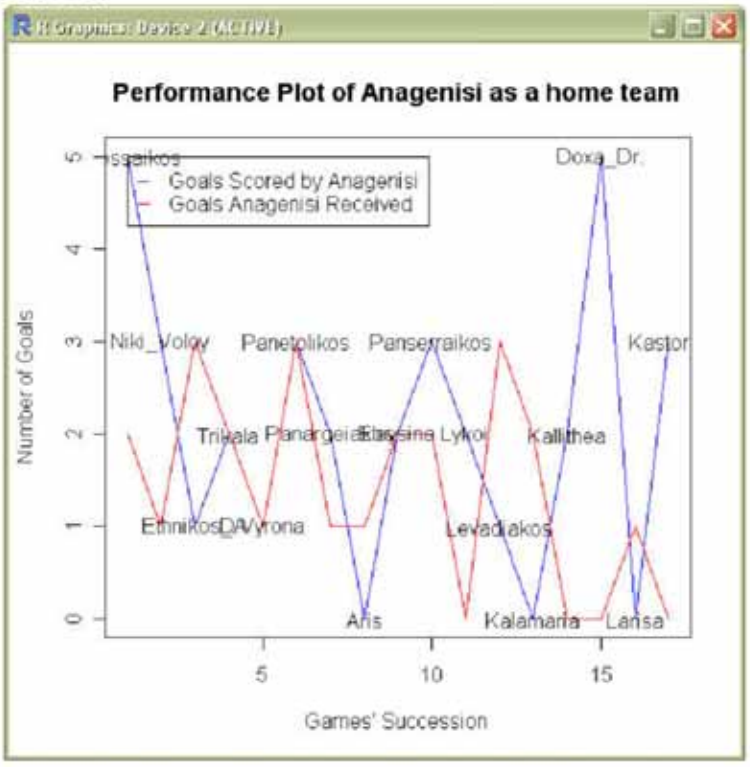
Note: Entering all teams in "Both" mode in the "Predict Final Score" menu gives you a detailed view on the Final Rank Reproduction for that team.

Chapter 3: Exploring the main menu

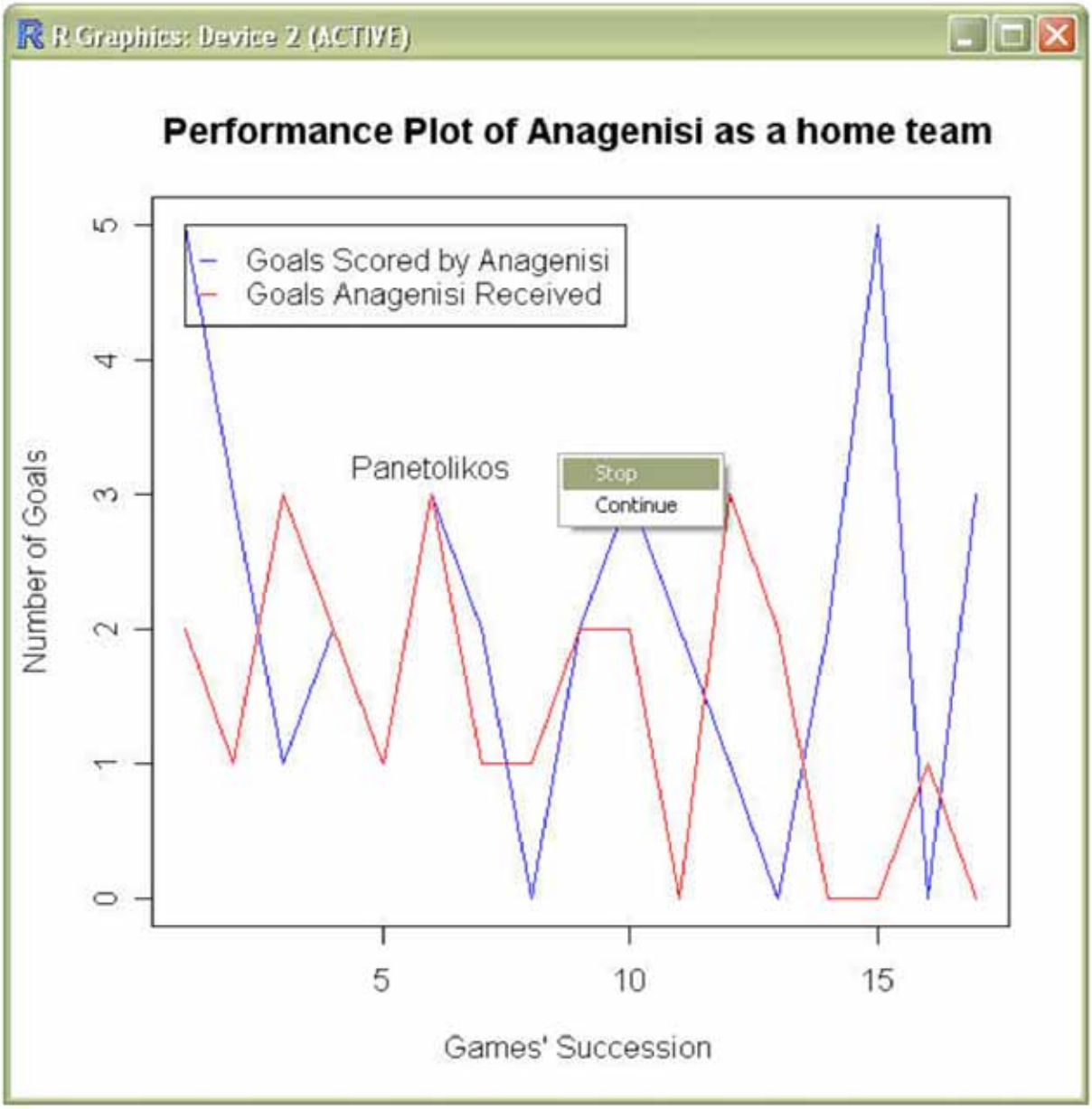
§3.4 The "Performance Plot" menu

This menu provides the user with graphic representation of the observed data for any single team. The available options here include:

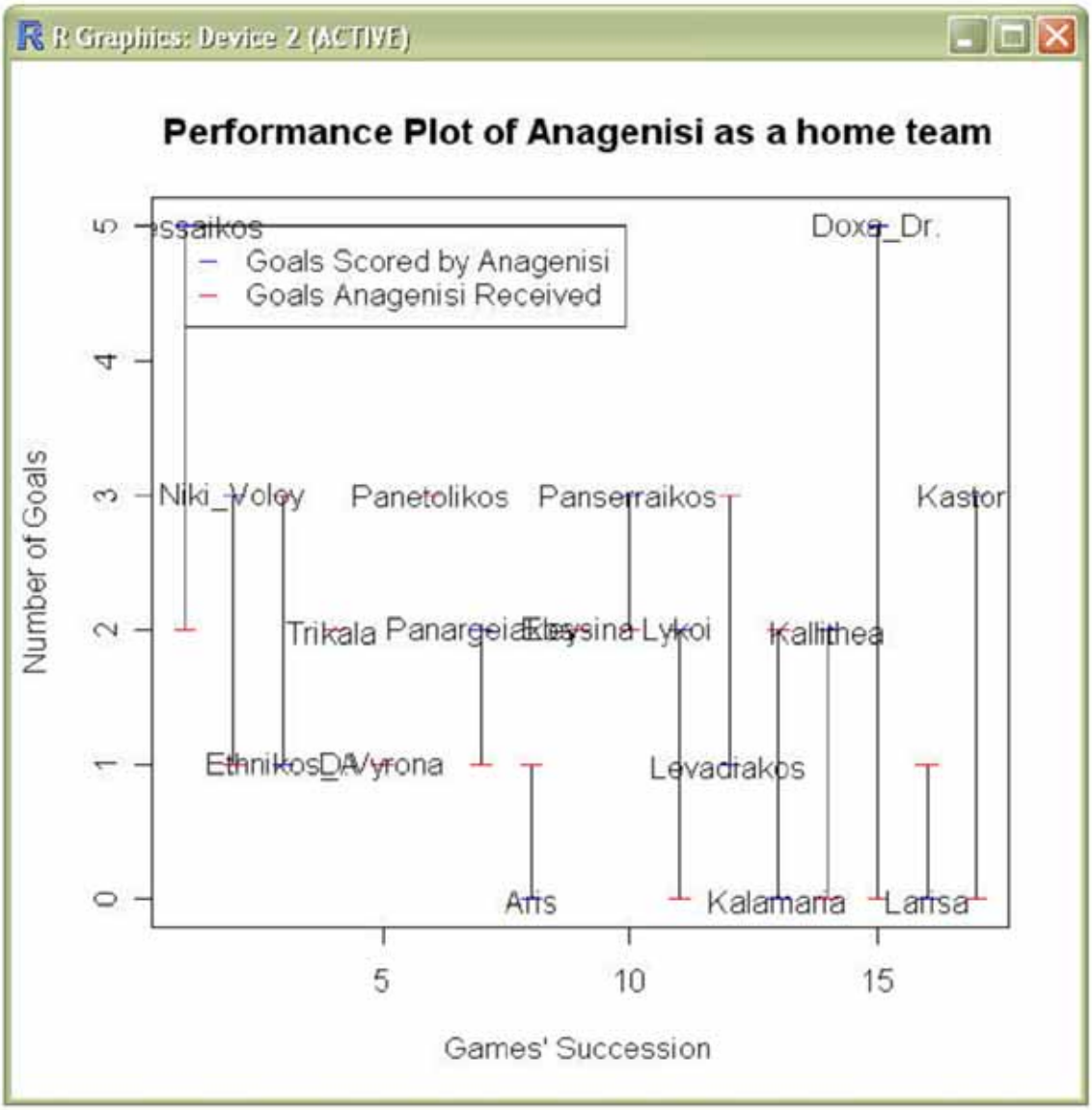
→ Home or visitor team selection. You can select to view the performance of the team as a home or as a visitor team in the games it participated in:



→ You may include in the graph the names of all opposing teams or the names for specific games you need. You can turn on and off each team's ID. This option is helpful when the plot is very cramped. The cursor turns into a cross with which you can click anywhere on the blue line spikes to name the current opposing team. This way, you can name the teams in specific games of special interest or name all of them one by one and avoid name tag overflow. To terminate the identification mode right click anywhere on the plot and select "Stop":



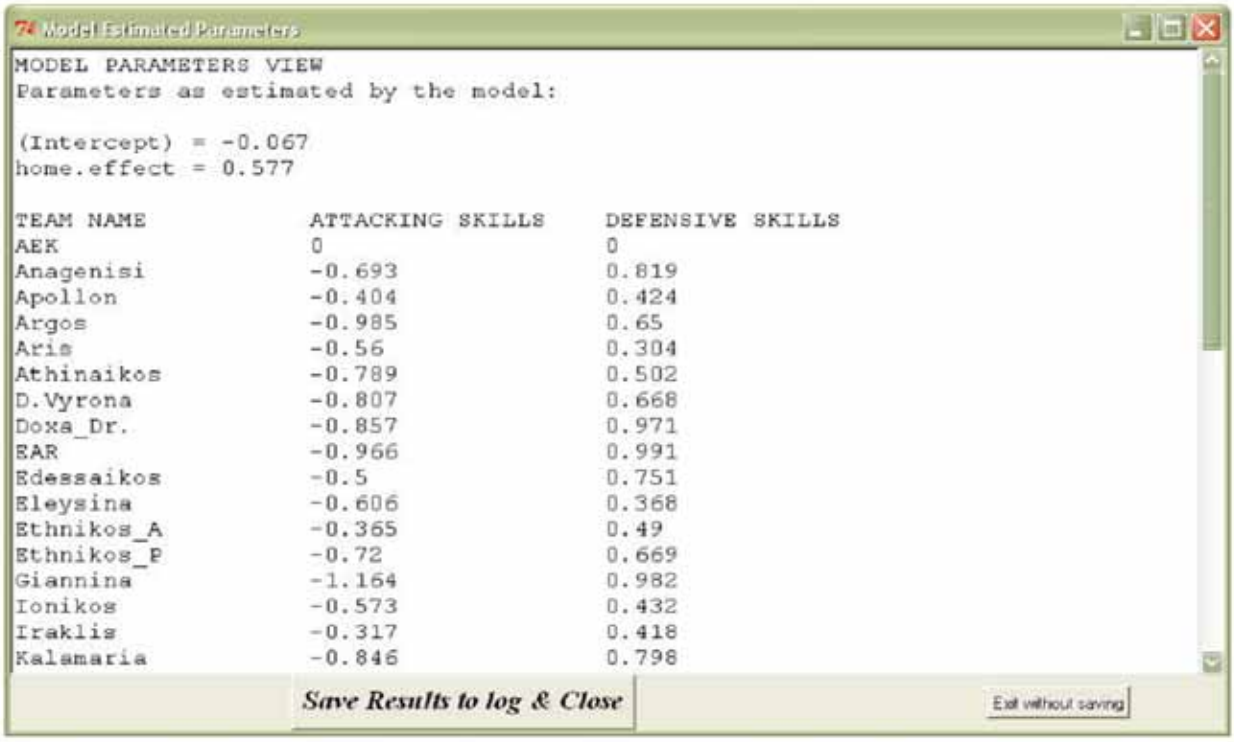
→ Draw a line or a bar plot. This option enables the user to choose whether to present the data in a two-line plot connecting the home and the visitor team goals with a blue and a red line respectively thus focusing on the change of goals scored and received throughout the time sequence of the games. Alternatively a bar plot can be produced which focuses on the difference in goals in each game.



All the options described for the plots are also operational for this type of plot.

Chapter 3: Exploring the main menu

§3.4 The "Model Parameters" form



Here the user can be informed about the technical details on the Poisson model used. At this form the output is user friendly for the average user enabling him to be presented with the attacking and defensive parameters of each team.

This section of the manual is addressed to slightly more experienced users. The fitted model is a simple general linear model which could be replaced by altering the code to enhance the predictive capabilities of the program.

$$goals_{i,j,k} \sim Poisson\left(\lambda_{i,j,k}\right),\ k = 1,2,\ i,j = 1...p$$

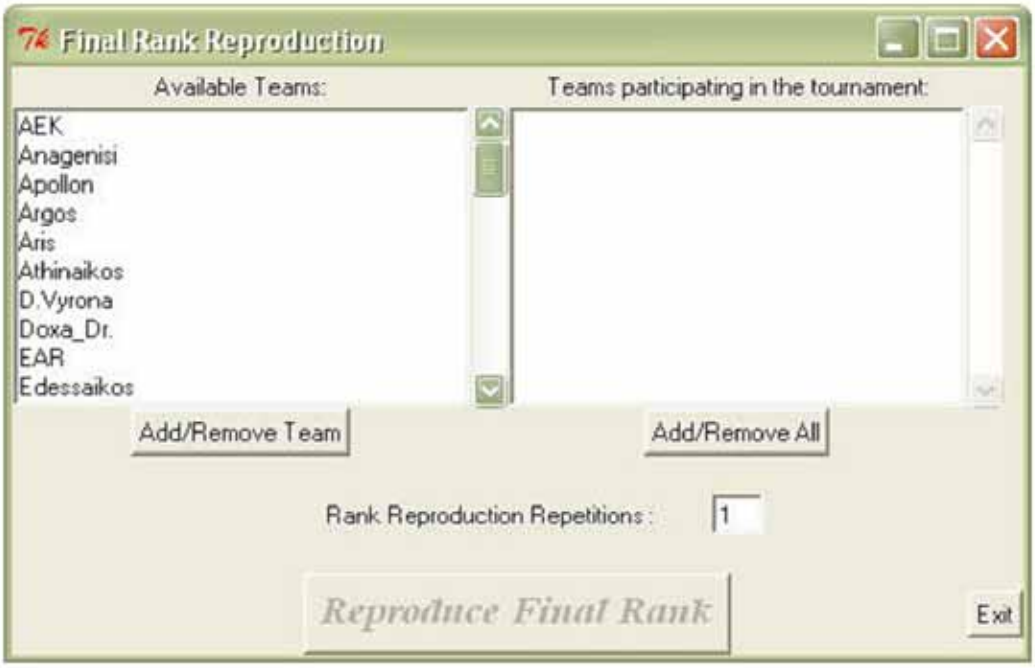
$$log\left(\lambda_{i,j,k}\right) = \mu + home.effect_k + defensive_i + attacking_j$$

Where μ is a constant (the intercept), *home.effect* is obviously the home effect parameter, *defensive* is the parameter describing the defensive capabilities of the team receiving the goal and *attacking* is the parameter related to the attacking skills of the team scoring the goal. Finally, p is the number of the teams. On a more technical note, we make use of corner-point parameterization in this model*.

* for further details see "On modeling soccer data" by Dimitris Karlis and Ioannis Ntzoufras, Student, 2000

Chapter 3: Exploring the main menu

§3.5 The "Reproduction of Final Rank" menu



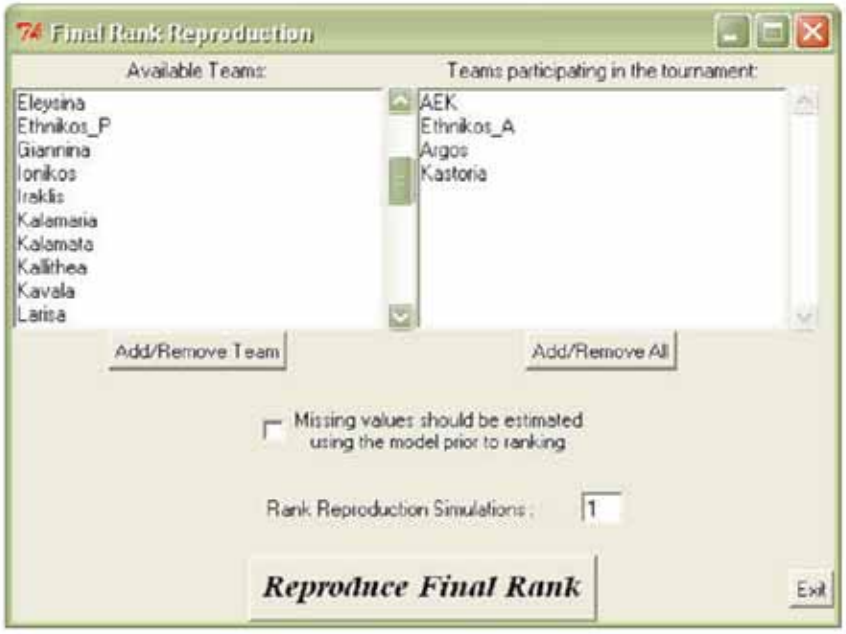
With this menu the user can create a virtual tournament where each team will play against all the other teams. The final ranking according to the score it will gather is calculated and presented at the end. To complete that, you need to select the teams you wish from the "Available Teams" list and pass them to the "Team participating in the tournament" by making use of the "Add/Remove Team" button. The same button can be also used to remove a previously selected team from the tournament.

In the case you wish to include or exclude all from the tournament you can use the "Add/Remove All" button.

The "Rank Reproduction Repetitions" field is used to input the desired number of "runs" for the specified tournament. If only one run is selected the output will contain the ranking based on the expected scores of each team. Otherwise the system will produce said runs, calculate the score in each runs and perform the final ranking based on the mean of these scores.

The main menu button "Reproduce Final Rank" will not be activated unless two or more teams are entered into the tournament list.

If there are missing values at the working dataset (Chapter §2.2) the "Reproduction of Final Rank" menu will detect them and ask, via a checkbox, to estimate them before the actual ranking is calculated:



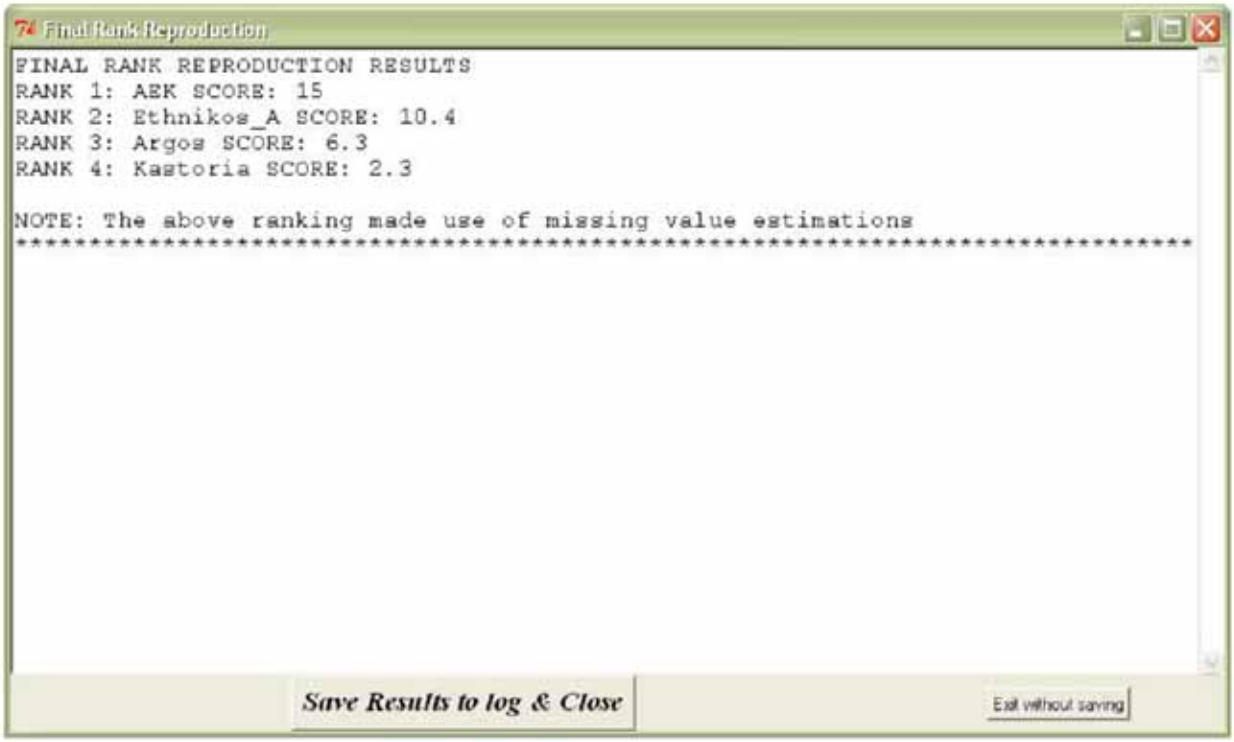
After selecting the participating teams and pressing the "Reproduce Final Rank" button the final ranking results appear on the screen:

The above output presents the ranking and the calculated score after a single tournament run according to the following formula:

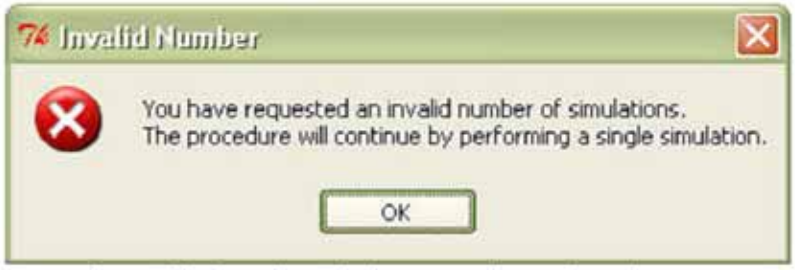
$$score = prob.win \cdot 3 + prob.draw \cdot 1$$

Where "prob.win" is the probability of a victory against the opposing team and "prob.draw" the probability of a tie between them.

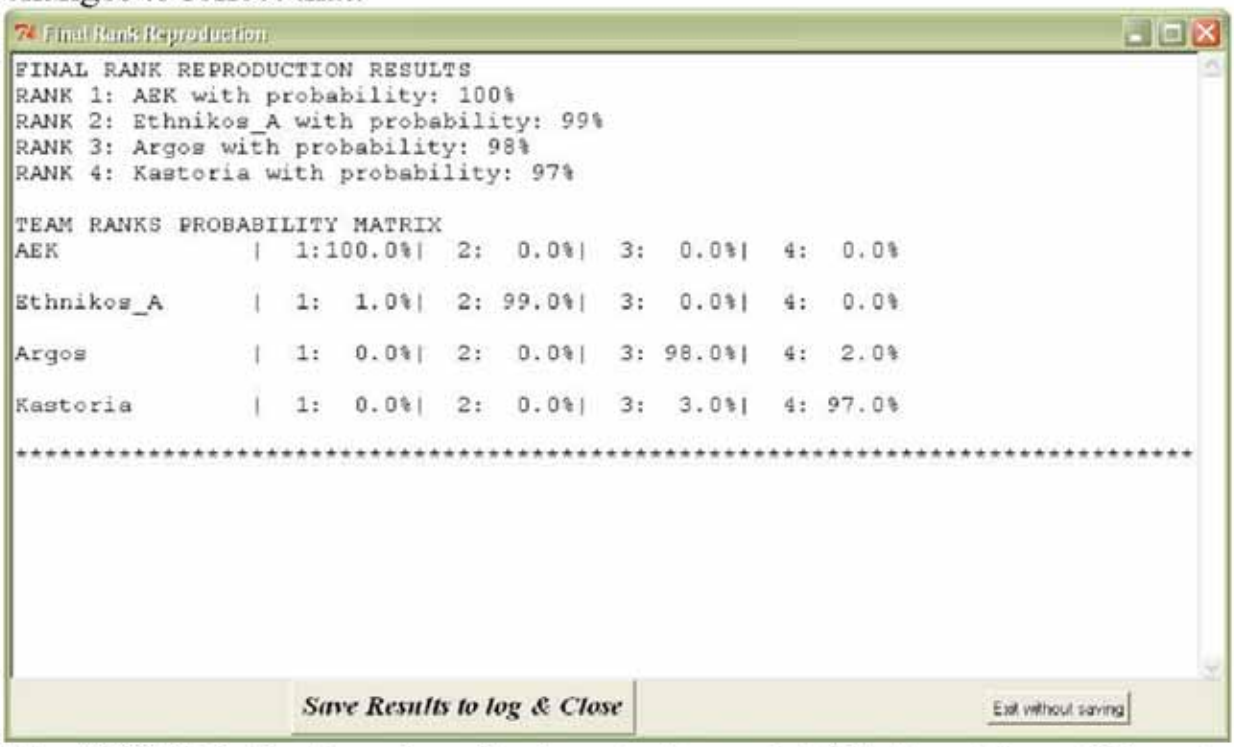
In case missing value estimation has been selected the output informs the user with the addition of a NOTE at the end of the output:



The "Rank Reproduction Simulations" stands for the number of simulations that will take place. If an invalid (non-positive) number is submitted the program will reject it and set the simulation number of repetitions to the default value ("1") and continue normally.



In case you demand multiple simulations to be taken into account the output changes to reflect that:



The SCORE indication gives its place to the probability to achieve different ranks. The output is dependant on the number if teams participating in the tournament. If more than 5 teams are selected the output presents the probabilities as a list for each team while suppressing ranks with zero probabilities since the matrix view gets too cramped. The probability illustrated above is calculated as the number of simulations where the current team achieved that rank against the total number of simulations so the accuracy of the above probability estimation relies heavily on the number of simulations taking place.

Chapter 4: Case Study Example

§4.1 English Football 2007

In this case study example we will examine the effectiveness of Football Analyzer by comparing the actual ranking of a real tournament against the outcome of the Rank Reproduction of the program.

The dataset we have describes the outcomes of games played for the Premiership in english football during the period 2006-2007. The actual ranking was the following:

Rank	Team Name	Score
1	Manchester United	89
2	Chelsea	83
3	Liverpool	68
4	Arsenal	68
5	Tottenham	60
6	Everton	58
7	Bolton	56
8	Reading	55
9	Portsmouth	54
10	Blackburn	52
11	Aston Villa	50
12	Middlesbrough	46
13	Newcastle	43
14	Man City	42
15	West Ham	41
16	Fullham	39
17	Wigan	38
18	Sheff Utd	38
19	Chalton	34
20	Wattford	28

We format the dataset into a proper text file like so:

```
home.team,away.team,goals1,goals2
Sheff Utd,Liverpool,1,1
Arsenal,Aston Villa,1,1
Everton,Watford,2,1
Newcastle,Wigan,2,1
Portsmouth,Blackburn,3,0
Reading,Middlesbrough,3,2
West Ham,Charlton,3,1
Bolton,Tottenham,2,0
Man Utd,Fulham,5,1
Chelsea,Man City,3,0
⋮
```

and store it in drive C:. We then invoke it from within the R command editor using the "`read.table`" command:

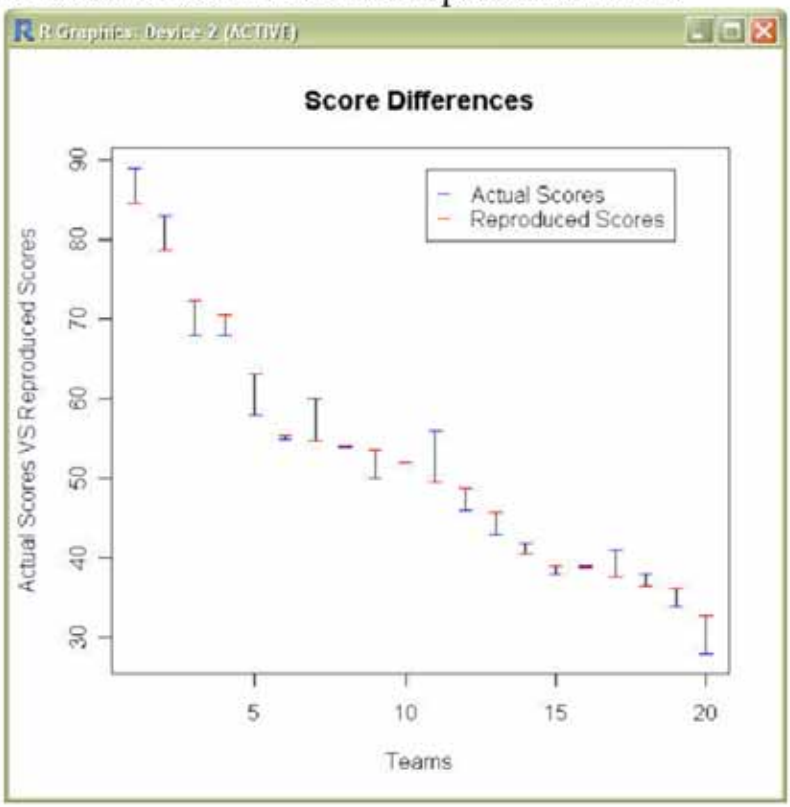
```
eng.foot<-read.table("C://eng_foot_2006-7.txt",
     header=T,sep=",")
```

We then call the Football Analyzer function and select eng.foot as our dataset. We find our way to the Final Rank Reproduction form and create a virtual tournament with all twenty teams included.

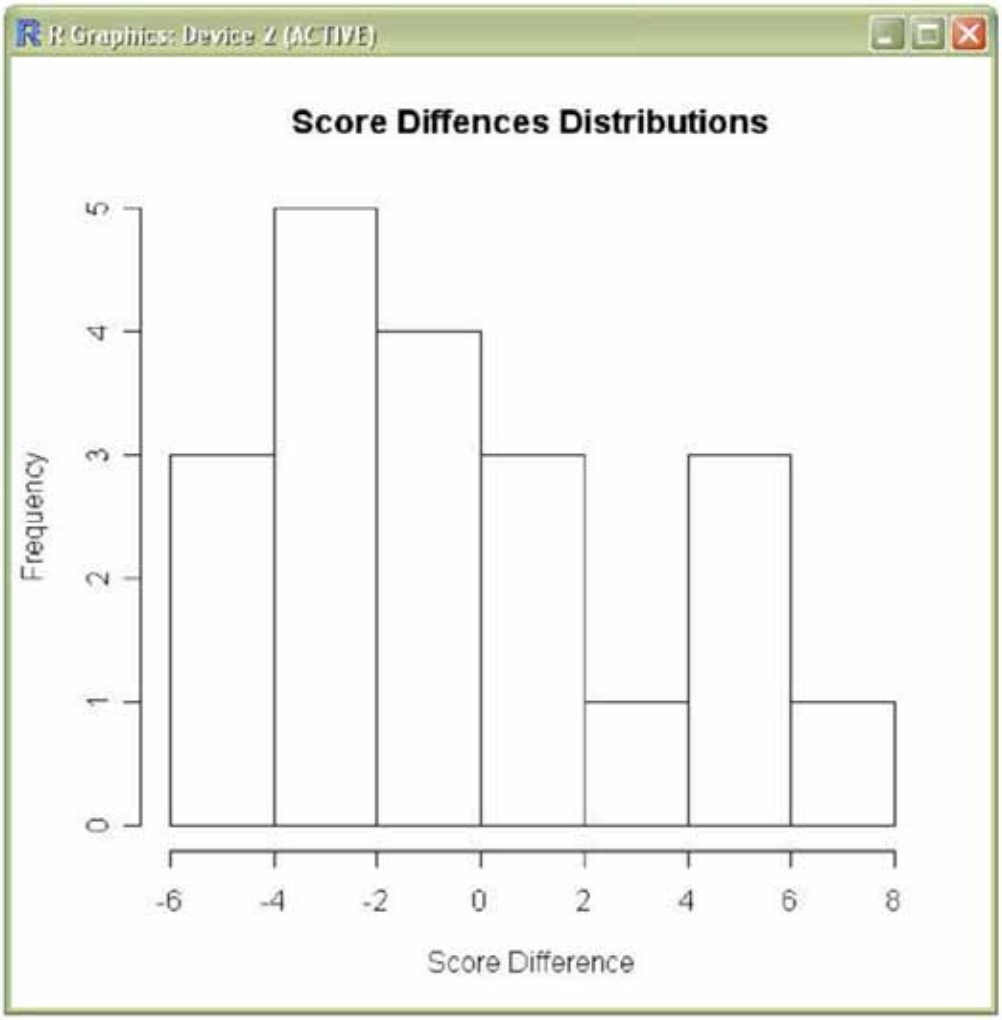
The Final Rank Reproduction output window looks like this:

```
FINAL RANK REPRODUCTION RESULTS
RANK 1: Man Utd SCORE: 84.8
RANK 2: Chelsea SCORE: 78.6
RANK 3: Liverpool SCORE: 72.3
RANK 4: Arsenal SCORE: 70.5
RANK 5: Everton SCORE: 63
RANK 6: Reading SCORE: 55.5
RANK 7: Tottenham SCORE: 54.7
RANK 8: Portsmouth SCORE: 54.1
RANK 9: Aston Villa SCORE: 53.6
RANK 10: Blackburn SCORE: 52
RANK 11: Bolton SCORE: 49.6
RANK 12: Middlesbrough SCORE: 48.9
RANK 13: Newcastle SCORE: 45.8
RANK 14: Man City SCORE: 40.5
RANK 15: Wigan SCORE: 39.1
RANK 16: Fulham SCORE: 38.8
RANK 17: West Ham SCORE: 37.7
RANK 18: Sheff Utd SCORE: 36.5
RANK 19: Charlton SCORE: 36.1
RANK 20: Watford SCORE: 32.7
**************************************************
```

With these results in hand we attempt a graphical depiction for the differences of the actual scores and the reproduced ones:



The largest observed deviation in absolute numbers of a reproduced score in relation to its actual counterpart was 6.4 while there were cases with zero deviance meaning that our reproduced scores were on par with their corresponding actual scores. To see how score differences are distributed we draw the following graph:



As we can see the distribution presents a slight positive asymmetry but due to the small number of data (merely 20 teams) we have inconclusive indications. However, we might skeptically claim that more often than not the statistical model underestimates the actual scores.

Now actual ranks come into focus. We have seen how reproduced scores are faring against actual ones so now the time has come to shift our attention to the ranking process.

```
RANK 1(1): Man Utd SCORE: 84.8
RANK 2(2): Chelsea SCORE: 78.6
RANK 3(3): Liverpool SCORE: 72.3
RANK 4(4): Arsenal SCORE: 70.5
RANK 5(6): Everton SCORE: 63
RANK 6(8): Reading SCORE: 55.5
RANK 7(5): Tottenham SCORE: 54.7
RANK 8(9): Portsmouth SCORE: 54.1
RANK 9(11): Aston Villa SCORE: 53.6
RANK 10(10): Blackburn SCORE: 52
RANK 11(7): Bolton SCORE: 49.6
RANK 12(12): Middlesbrough SCORE: 48.9
RANK 13(13): Newcastle SCORE: 45.8
RANK 14(14): Man City SCORE: 40.5
RANK 15(17): Wigan SCORE: 39.1
RANK 16(16): Fulham SCORE: 38.8
RANK 17(15): West Ham SCORE: 37.7
RANK 18(18): Sheff Utd SCORE: 36.5
RANK 19(19): Charlton SCORE: 36.1
RANK 20(20): Watford SCORE: 32.7
```

In the output above there have been added brackets indicating the actual ranking of each team.

We can see that 12 out of 20 team ranks were correctly reproduced achieving a good 60% accuracy ratio. That might appear to be low but one should consider the fact that errors come in pairs meaning that when one team is misplaced in ranking another team mistakenly takes its "rightful" place.

In more details, the most inaccurate case was that of Bolton whose reproduced rank was 4 ranks away from its actual one. Other than that, the rest of the cases keep a maximum distance of 2 ranks between the reproduced and the actual ranking a fact that shows that the reproduction is actually quite efficient.