Universidad
Carlos III de Madrid

Department of Signal Theory

BACHELOR'S THESIS

ANÁLISIS DE RESULTADOS PARA APUESTAS DEPORTIVAS: TENIS

Author: Lourdes Gómez González

Supervisor: Víctor Elvira Arregui

Leganés, 24th September 2014

Acknowledgements

Completed this stage of my life, I would like to mention and recognize those who have made this possible.

I would firstly want to thank you my family for all the support they have given me along these four years. Mom, dad, Cris, Almu, thank you for supporting me even when I was stressed because of the exams and insufferable. For always encourage me to never give up and continue the work. For having the confidence when I did not. Thank you just for being there every time I needed.

I would also want to thank you Daniel, for always being there and being my support.

To my supervisor Victor, without him I could not have done this project. Thanks for your support , commitment and dedication .

To Aran, a friend, a library companion, a teacher whenever I needed, with you all this has been easier. Thank you pupetina.

To María, always there, always supporting and encouraging. Thank you so much.

To Kris, for being by my side every year. Thank you.

To Lauris, Bowen and Alice for helping me and animate me this years. Thank you.

Of course thank you to my telematics, Manu, Pablo, Nitin, Nacho and Willy, for making these four years more enjoyable.

Abstract

Nowadays sports is one of the most popular social activities. The combination of sports and betting causes certain excitement in people, but this not always comes along with profits.

Numerical models can help predict the outcome of sporting events. The features within these models rely on data associated with the competitors. In this work we propose a system based on results of sporting events, obtaining a number of significant features, that will provide us the information to predict the outcome of a future sporting event, in this case for Tennis, so once we have the outcome probabilities we can bet.

As well as using the data such as the players involve in the events, novel features such as the surface are incorporated. This model will be used to simply predict the probability with which a tennis player will win a match, and also in the work we will see the Kelly Criterion, that will help us decide how much money to bet in the future events.

We will have a database of all the matches occurred in the latest years of the ATP World Tour, that will give us all the information about the events and of some of the gambling companies, or bookmakers, to compare the different rates given by the bookmakers and decide whether to bet or not.

By adding more informative features to the model, we will predict the outcome of future events with highly success.

Keywords: Probabilistic model, sporting event, Kelly Criterion. .

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Goals

1.1 Motivation of the project

Nowadays, the sports betting has become one of the biggest attractions for the customers sports fan. With the advantage of the internet, the number of bets has increased considerably. To this fact we have to add the massive amount of information that the users can obtain freely on internet.

The technological advantages of recent years have enable a great development in the market, offering the clients new ways to participate and get a great offer.

Both in the game and another aspects of our lives, we use the estimated probabilities to predict the outcome of an uncertain event such as the result of a match.

The motivation of this project is related to the high demand on online betting game and because of it, the opportunity of winning some money from predicting sports results. With this premise, a world of possibilities may develop different probabilistic models, in which many factors are involved, so that the field of sports betting offers a wide range of possibilities when gambling.

1.2 Objectives of the project

The main objectives of this project are to calculate the probability with which a player will win an specific tennis match, based on some parameters related to the players, previously calculated, and decide based on the Kelly Criterion how much money we should bet.

1.3 Structure of the Document

In order to help the reader to understand this document, here is a summary of its content divided by each of its chapters.

- **Chapter 2 - History of betting and tennis:** In this chapter we will explain the different types of bookmakers and odds. We will talk about some history of sports betting and how to bet in tennis as well as the mainly differences of betting in other sports. Finally, we will also explain the Kelly Criterion, which is the algorithm used to optimally determine how much money we should bet in a match.
- **Chapter 3 - State of Art:** A long this chapter we will see some other models use to predict the outcome of sport events, so we can be able to compare the existing systems with the one we create in this project.
- **Chapter 4 - Problem statement:** Here we will talk about the database used in the project in order to obtain our probabilistic model. As well we will explain in detail the algorithm used in order to achieve the goals proposed and also the Criterion used to determine the bet, which is one of our main objectives. And finally we will see some extensions done in the project.
- **Chapter 5 - Results:** In this chapter we will describe the results obtained in the implementation of our model proposed in order to achieve the goals.
- **Chapter 6 - Conclusions and future work:** Finally, we will give some conclusions and some ideas that could be done in the future to improve the project.

- **References.**
- **Appendix - Plan and Budget:** We will analyze in detail the budget that this project has been assumed.

Chapter 2

History of betting and tennis

2.1 History of sports betting.

The first betting took place in Greece, whose citizens gathered in the stadium to cheer on the best athletes of the time. Later, this same custom also appeared in Roman Empire, where they bet in gladiator combats and races. In the Middle Age, the bets where focused on the archery competitions, the gambling tournaments and launching knives [3].

However, the bets got its real power in the XVIII and XIX Centuries, thanks to the press evolution, who started to create certain sections devoted exclusively to sports betting. In the second half of the XIX Century, the stakes reached America [2].

In the XX Century betting websites began to multiply. The first to open was one in Liverpool and began to expand to Europe and North America. From this time, the stakes were worldwide popular. With the advent of the Internet, a lot of companies specialized in sports betting were created to bet online. Firstly in Canada and after in the United States.

Currently there is a high competition between gambling companies, bookmakers, whose main goal is to attract more customers and get hold existing in a market growing and expanding. [1]

2.2 Types of bookmakers.

The most popular are the traditional ones, where we can appreciate the contribution that the team predictions of the Bookmaker has set, and decide whether we want to bet or not [5].

Nowadays there exists another type of bookmaker which works in a totally different way than the others, these are the Betting Exchange Houses or P2P (People to people). Here they meet clients who want to bet in different events, that is why the bettor does not cross his bet with the bookmaker, instead of that he crosses his bet with another bettor, being the Betting House an intermediate between both bettors.

The Betting sport Exchange House takes a passive roll in the ante elections and the amount of money in the bet, because the bettors are the ones to decide this. These Houses make their benefits by charging a commission, which is calculated as an fixed quote by transaction or by a percentage of the net earning of the client. [4]

For the good bettors, these Bookmakers have the advantage that there are no limits in the bet or in the earning.

Another big secret for the Bookmakers success is what is call the Trading, the art of buying and selling bets as they were shares of stock.

The major advantages of the Bookmakers vs the Betting Exchange Houses are:

- **Share out**: The Bookmakers make a departure fee in advance, defined by its experts. However, in the Betting Exchange Houses the ones who define this quote are the bettors.
- **Liquidity**: Even though the Bookmaker does not limit the bet, it depends on the clients to increase the liquidity or not.

On the other hand, the main disadvantages that can be found in a Bookmaker are:

- **Better quotes**: As in this type of houses each one bets against the

others, there are no margins as in the traditional Bookmakers. We could be winning over a 15 or 20 percent. If we add the difference with the antes of some of them, we obtain the best ante in the market.

- **No limitations**: In the traditional Bookmakers there is a limit in the bet and in the maximum earning that a bettor can have in a specific time.

In our project we will work with the traditional Bookmakers, as the Betting Exchange Houses are forbidden in Spain. [7]

2.3 Types of odds.

The odds offered by the Bookmakers are used to calculate the gain that an user can get if he wins the bet.

As we said before, in the Bookmakers the rates are set by a technician group, with use several algorithms and forecast. And in the Betting Exchange Houses are set by the bettors itself. [9]

The quotes set for each player are inversely at their probability of winning, that is, if we bet for a player with a low quote, it will have a big probability of winning, whereas that if we bet for the one with a big quote, the probability that this player has of winning the match is small. [8]

In the Bookmakers we can find different types of odds, or quotes. Generally, there are three ways to present the rates, depending on the region or part of the world were they are used.

- The most frequent one is the one in **decimal format**, usually used in Europe except in the United Kingdom. They are characterize for having 2 or 3 decimals and one easy way to calculate this quote is to divide 1 over the probability of winning that the player has. [10]

  For example, if the Bookmaker gives a player the probability of winning of 0.50, the rate that will appear will be of $1/0.5$ which is 2, provided that we consider an ideal situation in where the Bookmaker will not save a percent of gain.

When the percentage offered by the Bookmaker is lower than the one we calculated, is the time to bet.

Lets imagine we bet 10 for a player whose quote is 3. This will let us know how much we can win if the tennis player wins the match.

In case our player wins the match, to calculate the gain:

$$g = m_b \cdot q$$

Where g is the gain, m_b is the money we are betting, and q is the quote of the bookmaker.

If we want to know the benefits b:

$$b = m_b \cdot (q - 1)$$

So if our player wins the match, our gain will be 30, while the net benefit will be 20. As we know, the Benefit = Gain - Money bet (20 = 30 - 10). [8]

- On the other hand, they use another type of ante, the **fractional format** which is the one used in the United Kingdom. In this type the numerator indicates the gain and the denominator, the money bet. [11]

  To calculate the benefit with this fractional format, we use:

  $$b = m_b \cdot f_q$$

  Being f_q the fractional quote.

  We can also change from fractional format f_q to decimal d_q:

  $$d_q = f_q + 1$$

- Finally, there is a third type of format which is the **american format**, used in the United States. Is normally used in the traditional American sports and it can be either positive or negative. The negative quotes are to indicate the amount of money it has to be bet in order to obtain a benefit of 100 units, while the positive quotes indicate the benefit obtained if we bet 100 units, euros in this case.

  If we compare this American format with the decimal one, the positive quotes will be bigger than 2 in decimal format, and the negative ones, smaller than 2.

To calculate the benefit b obtained in this American bookmakers:

$$b = \frac{m_b \cdot q}{100}$$

For the negative quotes, we do not have to include the sign in the equation.

As we had for the fractional format, we have a way to put the American format a_q in decimal d_q:

$$d_q = 1 + \frac{a_q}{100}$$

With this formula, the American Quote will not take negative values. [8]

2.4 Betting on tennis.

Tennis is a popular sport in most cities of the world and most of the bettors try to specialize in it because there are many factors on which depends a victory of a player.

The main reasons why tennis is one of the main sports of gambling are:

- Is not played against a clock but against an objective.
- Its punctuation style makes many changes occurring advantages in the match, perfect for live variations, for example, is ideal for trading for excessive fluctuation.
- There is no tie.
- The mini-breaks are great to operate.
- Is an individual sport, where the psychological factors can change the development of the match. [13]

To star betting in tennis, there are some factors to consider:

- It is an **individual sport**, in where having a bad day, bad physical form and things like that can affect the game.
- The **surface** where the match is taking place make the results variate. Many tennis players are better in some surface, like hard, than in others.
- The **type of tournament** affects in many factors as the surface, the type of ball, the humidity and temperature, that directly affect at the tennis players.
- **Mood and motivation.** The more motivation, the harder for a tennis player to loose if it is considered the favorite.
- **Styles of playing.** We can classify two bug groups of tennis players, the ones who play to attack, and the ones who defend. It is also important to know if the players has a good service or instead has a great return.
- **Number of sets and duration of the match.** In some tournaments as the Grand Slam or Copa Davis the match is five sets, whereas the other tournaments are three sets long. This is important because is harder to surprise a leader in a long match of five sets. Also, is more important the physical form of the players in a longer match.
- **Accumulated fatigue.** The players performances drops when they are in no good shape, above all the ones that do not have a good physical training.
- The **injuries.** In case the is a player injured, he automatically looses the match. The players who decide to play even though they have any injure, have more difficulties to win the match. [12]

2.5 Kelly Criterion.

The Kelly Criterion was developed by John Kelly in 1956 [14]. It was originally created to bet in the horse races, but it can be used in any type of bet.

Is a criterion created to maximize our long term budget available, calculating the correct amount of money to bet and nowadays is an useful tool to manage our budget.

With this method we can calculate the optimal bet taking into account the fee of the bet and its forecast. To bet more is an unnecessary risk and on the other hand, to bet less carries a lower yield.

The formula used in this criterion for events with two possible endings is:

$$r = \frac{p(b+1) - 1}{b}$$

Where b is the rate given by the bookmaker minus 1, and p is the probability of winning for Player1. [14].

This algorithm has the following characteristics [15]:

- The calculation of the probability is subjective, so the efficiency of this method resides in the ability that we have to estimate an appropriate probability event.
- The key of this algorithm is to calculate a probability that conforms more the reality that the ones given by the bookmakers, which determine their own fees.
- Depending on the bookmaker, these fees will be imposed by their own algorithm or through the betting users. For example, in Spain legally the fees are imposed by the bookmakers.

Chapter 3

State of the Art

3.1 Other prediction models

Numerical models can help predict the outcome of sporting events. The features within these models rely on data associated with the competitors.

There are many ways to predict sport results based on numerical models.

We are going to briefly explain some examples of different models used to reach the same goal, sport results prediction, not only tennis as we have performed, but any other sport event.

- **Logistic regression model** was used to predict the outcome of American Football matches.

  The first relied on the theory that the home team within a sporting event has a certain advantage over the away team. Therefore, this basic predictor chose the home team to win within every match (home).

  The other baseline algorithm used previous results between the two competing teams to achieve a prediction. After analyzing the accuracy of differing amounts of data, it was found that the forecasts became less accurate when older results were used. Thus, the most accurate predictions came from just using the previous years results between the teams (prev_{res}).

When tested on all of the matches, the unsupervised *home* baseline achieved an accuracy of 57.8% and the supervised $prev_r es$ obtained 58% accuracy. A logistic regression model was used to improve on the 58% accuracy found by the baselines. The output of a logistic model is a binary result (1/0) so as tied games were ignored, this made the approach suitable to forecast either a home or away win.

The logistic model encompasses features that represent data relevant to the result that is trying to be predicted. Then during the training of the model, the relationship between each feature and the result is assessed to see if it has a strong or weak correlation with the end result. This relationship is then used when trying to predict the outcomes within the testing phase. [28]

- A model is proposed for predicting the result of a football match from the previous results of both teams. This model underlies the method of identifying nonlinear dependencies by **fuzzy knowledge bases**.

  Acceptable simulation results can be obtained by tuning fuzzy rules using tournament data. The tuning procedure implies choosing the parameters of fuzzy-term membership functions and rule weights by a combination of genetic and neural optimization techniques. [29]

- **Bayesian networks** provide a means for representing, displaying, and making available in a usable form the knowledge of experts in a given field.

  The objective is to determine retrospectively the comparative accuracy of the expert Bayesian Network compared to some alternative **Machine Learning models** that were built using data from a two-year period in football.

  The additional Machine Learning techniques considered were: MC4, a decision tree learner; Naive Bayesian learner; Data Driven Bayesian; and a K-nearest neighbor learner. The results show that the expert Bayesian Network is generally superior to the other techniques for this domain in predictive accuracy.

  The results are even more impressive for Bayesian Networks given that, in a number of key respects, the study assumptions place them at a disadvantage. For example, we have assumed that these networks prediction is incorrect if a Bayesian Network predicts more than one outcome as equally most likely.

  Although the expert Bayesian Network has now long been irrelevant the results here tend to confirm the excellent potential of Bayesian Networks when they are built by a reliable domain expert. The ability to provide accurate predictions without requiring much learning data are an obvious bonus in any domain where data are scarce. [30]

- The **Poisson model** further. Parameters representing the teams inherent attacking and defensive strengths are incorporated and the most appropriate model is found from a hierarchy of models. Observed and expected frequencies of scores are compared and goodness-of-fit tests show that although there are some small systematic differences, an independent Poisson model gives a reasonably accurate description of football scores. Improvements can be achieved by the use of a bivariate Poisson model. [20]

Chapter 4

Problem statement

4.1 Database for players and matches.

In order to be able to create an algorithm for the probability calculus and application of the Kelly Criterion for betting, we had to use some database for the players and the matches.

We took this database of the last three years, from 2011 to 2013, from the ATP Men's Tour [16] so that based on the results of the matches given, we could create our own base to develop our algorithm.

This database gives us the information about all the matches from the Masters Series. From here we can obtain the following data:

- The **tournament**.
- The **date**, which is important for timing the succession of different events.
- The **surface**, so that we can differentiate between them and compare the results of the players depending on the surface of the event.
- The **court** can be outdoor or indoor. Most of the tournaments take place in outdoor courts, except for the ones in winter which will be indoor and always in hard surface.

- The **winner** and **looser** of the match. This way we can see how the players are doing in the lasts events.
- The **ranking** which indicates the players position in the ATP Tour. With the database given we will make our own ranking for the players which we will use in our algorithm.
- The **state of the match** whereas is completed, walkover or retired. We will only take into account the events completed.
- The **odd** given for each event by several bookmakers.

4.2 Probabilistic model.

As we have one event with only two players, a first approach could be that the probability for each one of them, of winning or loosing the match, is 50%, the same qualities as flipping a coin. This method is possible but since we know that not all the players are as good, or as bad, than their opponent and we have many data that confirms it, we will use a probabilistic model based on Bernoulli Distribution. [17] [18] [26]

First of all, given the database of the players and the matches, we need to calculate certain parameters of the proposed model for each player that will give us a numerical representation of how good or bad are these players.

Once we already have the ranking and all the players organized, we will calculate their parameters. In order to do it, we will maximize the likelihood, which is the joint probability function,that is the multiplication of the Bernoulli probability functions of each match.

The Bernoulli Distribution is a probabilistic distribution with a parameter p. Were the probability mass function is:

$$f(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases} \tag{4.1}$$

being p the probability of $k = \{0, 1\}$, and has a value of $0 \leq p \leq 1$.

Our probabilistic model supposes that the probability p of the Bernoulli Distribution is approximately as follows:

$$p_{ij} = \frac{1}{1 - e^{\theta_j - \theta_i}} \tag{4.2}$$

where p_{ij} is the probability of player$_i$ winning over player$_j$. And Θ is the vector with the parameters of all the players from the data available.

Given the probability p_{ij}, the Likelihood is:

$$p(Y|\Theta) = \prod_{n=1}^{N} p(Y_n|\Theta) = \prod_{n=1}^{N} p_{i(n),j(n)}^{y}(1 - p_{i(n),j(n)})^{1-y} \tag{4.3}$$

where Y_n are the binary results of all the matches, indicating $Y_n = 0$ if player$_j$ wins the match, and $Y_n = 1$ if player$_i$ does.

In order to obtain the parameters desired to be able to implement our probabilistic model we have to maximize this likelihood.

To obtain the inference algorithm we are going to introduce a prior of the parameters with the main goal of avoiding whats is called "overfitting". [33]

In statistics and machine learning, overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

Even if the model is not very complex, this happens when we have only a few data of a player, for example, we only have one match where he has played and even though he is not a really good player,he wins. Then the model will give him a good parameter as the only data we have from him is good, but this will not be realistic. So we want to infer the parameters of the players that will maximize the posterior.

By Bayes, the posterior is:

$$p(\Theta|Y) = \frac{p(Y|\Theta)p(\Theta)}{p(Y)} \tag{4.4}$$

We do not know $p(Y)$, but do not need it to calculate the posterior, as we are actually calculating the "Unnormalized-posterior".

$$p(\Theta|Y) = p(Y|\Theta)p(\Theta) \tag{4.5}$$

that is the Likelihood $*$ Prior.

As we already calculated the likelihood, we now need to find the parameters that minimize this likelihood, for that we use the prior and take its logarithm.

$$\begin{aligned} -\log p(\Theta) &= -\log\left(\prod_{n=1}^{M} p(\Theta)\right) \\ &= -\sum_{n=1}^{M} \log \frac{\Theta(n)^{k-1} e^{-\frac{\Theta(n)}{s}}}{s^k \gamma(k)} \\ &= (k-1)\sum_{n=1}^{M} \ln\left(\Theta(n)\right) - \sum_{n=1}^{M} \frac{\Theta(n)}{s} - M \cdot k(s) - M \cdot ln((k-1)!) \end{aligned}$$

Where s is the scale, which is a fixed value of 2, k is also another fixed number of value 2. γ is a function that calculates the factorial of k and M corresponds to the number of players.

4.3 Inference algorithm.

In order to obtain the parameters of the players we have to maximize our model, or whats it is the same, minimize the logarithm of it. So we calculate

the logarithm of our likelihood and the unnormalized-prior [27]:

$$L = -logp(\Theta|Y) = -logp(Y|\Theta)p(\Theta) \tag{4.6}$$

Given this function we use a Matlab function called 'fmincon', that will give us the global minimum of the function L since the function is convex, which are the parameters that we are calculating for our players.

To evaluate the matches and find out the parameters, we gather together all the events, of all the tournaments, taking place in the same day. The 'fmincon' function looks for the minimum each day, taking all the matches of that day, starting always from the first day. This way our model learns from the first day until the last.

As we have three tennis seasons, there is a lot of data to evaluate so 'fmincon' needs too much time to be able to determine the minimum. To solve this problem we decided to evaluate each day taking as initial point the day before instead of the first day, this way the minimum we are looking for will we near the last one and the function will not take that long to achieve it.

Once we have these parameters, we can calculate the probability of winning of the players, with the formula $p_{i,j}$ already seen.

We implemented the model for:

- Season 2013.
- Seasons 2012-2013.
- Seasons 2011-2013.

4.4 Model extensions.

Once we have obtained all the results with the method explained before, we decided to do some extensions in order to improve the algorithm.

First we decided to take a forgetting factor, α. This is because we know that is more important for an event happening today between two players, i.e. the performance of a couple of players in recent events, and not so important the matches occurred a year ago. This is why we design an forgetting factor based in a decreasing exponential where the latest events have more weight that the ones occurred months ago.

The expression of this factor α is the following:

$$\alpha = a_o - (a_o - b_o)e^{-\beta(t-1)} \tag{4.7}$$

where a_o is the upper limit, 1, and $(a_o - b_o)$ is lower limit, 0.1.

And beta is a parameter calculated to obtain a minimum value at the end of the year, that is that $e^{-365\beta} = 0.3$, this gives as $\beta = 0.03$.

Once we calculated this factor we multiply it by the Likelihood, so when calculating the parameters for the players, we take into account this exponential and this characteristic parameters are modify to be more important for the latest matches.

Another extension that was going to be done was taking into account the surfaces. We could not finish this extensions because of the limit time in the project, and all the time needed for the execution of it.

We divided the matches into four groups depending on the surface, clay, grass, hard outdoor and hard indoor.

To obtain the results of the matches only on one of the surfaces, i.e. clay, we take all the events occurred in this surface and the rest of the events that took place in the other types of surface will also be counted but with a Surface Factor.

This surface factor, S

$$S = \begin{pmatrix} 1 & 0.6 & 0.4 & 0.2 \\ 0.6 & 1 & 0.2 & 0.2 \\ 0.4 & 0.2 & 1 & 0.6 \\ 0.2 & 0.4 & 0.6 & 1 \end{pmatrix}, \tag{4.8}$$

represents the factor given to a match of a certain surface, i.e., S_{ij} corresponds to the factor when evaluating the surface i and the match takes place in surface j. If $i = j$, we are evaluating the same surface where the match is taking place, that is why S_{ij} will be 1.

Continuing with the example of clay surface, we will have the matches in clay surface multiply by factor 1, the ones in grass by 0.6, the ones in hard outdoor by 0.4 and finally in hard indoor by 0.2. This is because the type of game in clay and grass is similar that the one indoor, so we have a bigger factor for this. The same will be for the rest of the surfaces.

Adding these extensions finished to the implementations already done with our model, we ended up with five different variations of our model:

- Season 2013.
- Seasons 2012-2013.
- Seasons 2011-2013.
- Seasons 2011-2013 with Forgetting Factor.

We also made some extensions when betting. In order to bet the money according to Kelly's Criterion we made several options for the budget and comparing it with some of the bookmakers, like Bet365, or with the best rates between all the bookmakers given. So we could bet with:

- A budget for all the season.
- A budget per tournament.
- A budget per match.
- With an specific bookmaker.
- With the best rates among all bookmakers.

Chapter 5

Results

5.1 Numerical results

We can see several results of the developed model. On the one hand, we have the numerical results where we can see the money bet, how many times our predictions are correct and we win money, and the times we fail in our predictions as well. On the other hand, we have the data obtained with the model, such as the information about the quality of the players.

In order to qualify how good our bad our model predicts, we have used a evaluation factor called Log Loss or Predictive Binomial Deviance [31]. This factor is as follows:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)$$

Where y_i is the outcome of each game, $\hat{y}_i$ is the predicted probability of player$_i$ winning and n is the number of matches played. The smaller this factor is, the better our model is. [31]

We have used this factor in all our variations of the model and we also

evaluate with this factor the different bookmakers used in the work, to be able to make a comparative of the quality of the model. We have made it for all the probabilities of the matches and also the probabilities of the last three months of the seasons. These last three months will give us a better result of the Log Loss as at the end of the season we have learned a lot about our players and we are able to predict really good the outcome of the events.

Here we have the results of the Log Loss obtained of out model:

Season	Complete	Last 3 months
2013	0.6357	0.5961
2012-2013	0.6118	0.5853
2011-2013	0.6382	0.6153
2011-2013 with forgetting factor	0.6932	0.6632
Average Model	0.6140	0.5891

Table 5.1: Log Loss for our model.

Now the ones for the different bookmakers used:

Season	Complete Season	Last 3 months
2013	0.5546	0.5659
2012-2013	0.5486	0.5659
2011-2013	0.5476	0.5659

Table 5.2: Log Loss for Bet365.

Season	Complete Season	Last 3 months
2013	0.5562	0.5643
2012-2013	0.5509	0.5643
2011-2013	0.5501	0.5643

Table 5.3: Log Loss for Expekt.

Season	Complete Season	Last 3 months
2013	0.5535	0.5621
2012-2013	0.5474	0.5621
2011-2013	0.5455	0.5621

Table 5.4: Log Loss for Pinnacles Sports.

If we take a probability of 0.5 for all the events, like flipping a coin, we will get a LogLoss = 0.69. As we can see in the tables above, our model

Season	Complete Season	Last 3 months
2013	0.5568	0.5641
2012-2013	0.5508	0.5641
2011-2013	0.5500	0.5641

Table 5.5: Log Loss for Stan James.

Season	Complete Season	Last 3 months
2013	0.5558	0.5606
2012-2013	0.5501	0.5606
2011-2013	0.5493	0.5606

Table 5.6: Log Loss for Ladbrokes.

predicts the future events wide better than just giving a probability of 0.5 to each of the players. But this is not the worst case, indeed if we give a player a probability of 0 of winning a match, and this player wins the event, this will give us a LogLoss tending to infinity. For example, thinking that a player will win the 80% of the times and just do it half of them, makes that the 20% that we gave him as a looser also happens half of the time, and this gives us a big penalty in LogLoss = -(log(0.8)/2 + log(0.2)/2) = 0.92.

So we can conclude that the results obtained in our model with this evaluation factor are pretty good and near the results that the bookmakers get, specially in the ones evaluated for the last three months of the seasons. Here the smallest difference between our results and the bookmakers ones is 0.03 and the biggest 0.1, as we can see in the Tables. In conclusion, this is a good result taking also into account what do the bookmakers obtained and what the worst case will be.

As we have seen before we use Kelly's Criterion to know how much money we should bet in each of the events taking into account the probabilities calculated for them. We are going to see the evaluation done using this Criterion for each of the variations of our model.

So another interesting result to analyze is the amount of money won and lost along a tennis season.

Figure 5.1 is the amount of money won/lost along the season 2012-2013, having a budget per match of 1€:

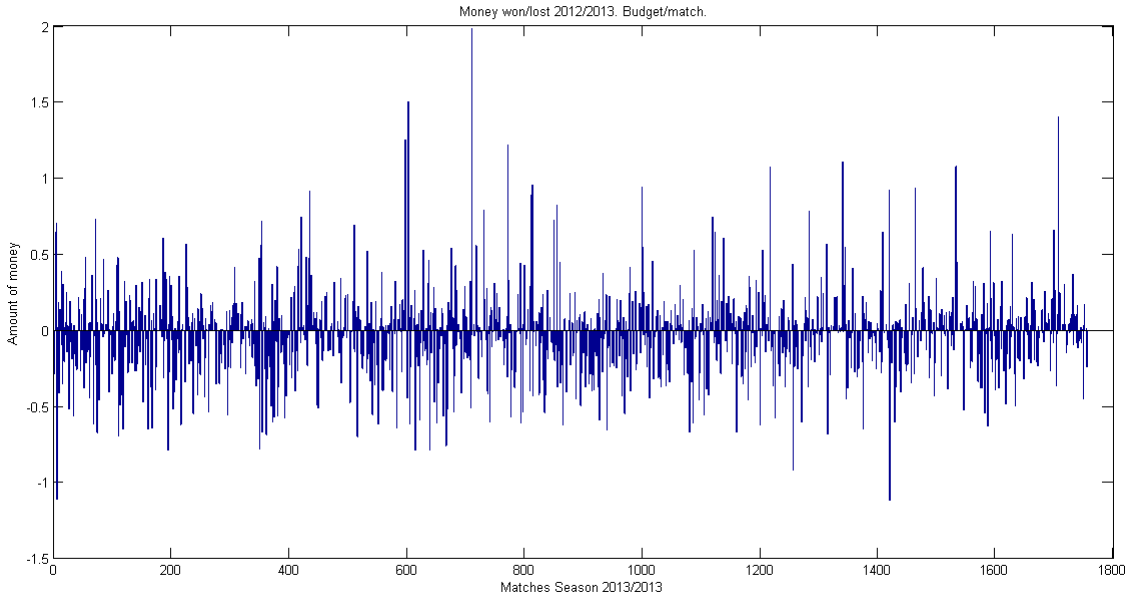

Figure 5.1: Money win/lost 2012-2013.Budget per match

As we can see in the Figure 5.1, we are analyzing match by match so we can appreciate that in some of the matches we win a lot of money as in the match 700, approximately, and in some of them we loose. In average we just loose 0.01 over a 1€ in each match, taking into account that in some of the we win more than 2€ over 1 of budget it is a really good result.

We have to know that even though we estimated really good our probabilities, near the ones given by the bookmakers, they always have an 5% extra being the sum of their probabilities not 1 but 1.05, so it is more difficult for us to win as much money as they do.

The Figure 5.2 represents the money won/lost along the same season, 2012-2013, but this time having a budget per tournament of 100€.

We can easily see that in the Figure 5.2 there is suddenly in one tournament where we win more money than in others, while in the rest of the tournaments we have less fluctuation of the amount of money won or lost. As we can appreciate in the graph that there are many times where we decide not to bet, and the value evaluated is 0 in the Figure. This happens when the probabilities calculated for that event are critical, or too big or too small, or maybe because when using Kelly Criterion the situation taking the odds

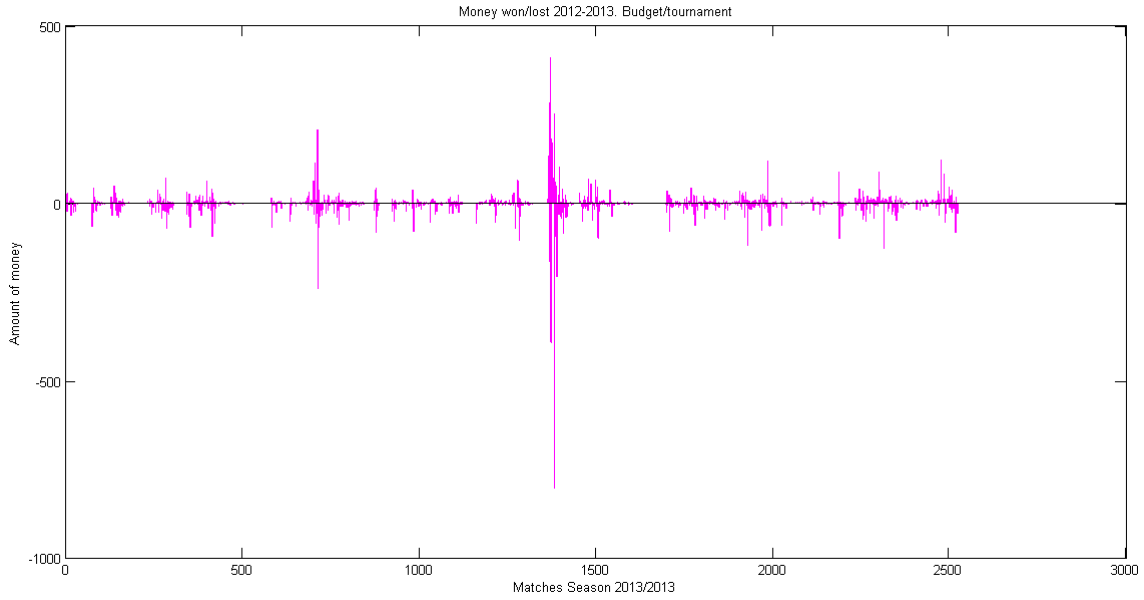

Figure 5.2: Money win/lost 2012-2013. Budget per tournament

of the bookmakers is not the best to bet.

We can also see how in the end of the season we win more money, this is because our predictions are more accurate as we have learned during the season and the information about the players is better. In Figure 5.3 are the last matches of the season, where we can appreciate this profits.

The Figure 5.3 represents the last matches of the season specifically the last two tournaments, the Masters Cup in London and the Paris Masters. In the first tournament, the BNP Paribas Master, we triple our original budget of 100€ with a net profit of 200€ . And in the Masters Cup we only loose 3.7€ over 100€ of budget.

In Figure 5.3 we can also notice that in the last match we loose money, exactly 30.5€ as the Kelly Criterion told us to bet 24% of our budget, 126.7€ in that moment. This match which is very illustrative is the Final Masters Cup, between Nadal(1) and Djokovic(2), where we knew that Nadal had a better parameter in that moment than Djokovic, but he loose. We lost money because our model does not consider that the number 2 in the ranking wins the number 1, that is why we gave Nadal a larger probability and we bet for

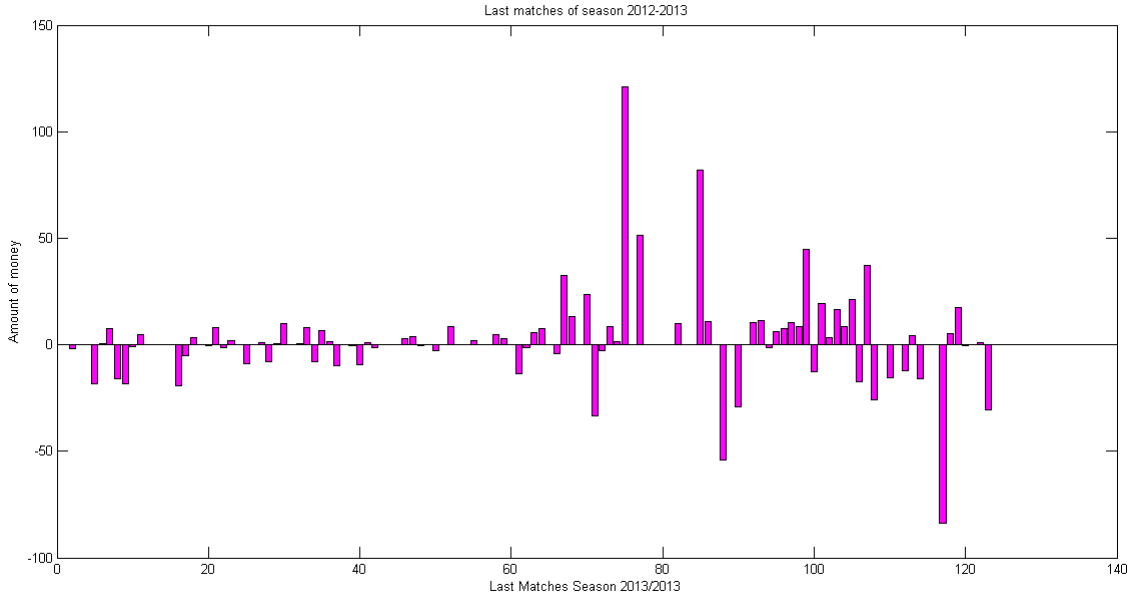

Figure 5.3: Money win/lost 2012-2013. Budget/tournament end of season

him but we loose[1].

Now we can see these same results but taking into account a threshold over Kelly's Criterion decision. This threshold is settle to ignore the decisions too extreme than the Criterion takes, these is when Kelly says to bet over a 50% of our budget, as we consider it too risky.

The Figure 5.4 and Figure 5.5 represent this threshold taking a budget per match of 1€, as before, and a budget per tournament of 100€.

We can see how, as happened with the graph without the threshold, at the end of the season in the Figure 5.5 we win much more money as we know much more about the players and when estimating we get probabilities more likely successful.

[1] Our model only takes into account that Nadal is a better player at that moment than Djokovic, but we do not consider that this Master takes place indoor, a surface in which is better Djokovic than Nadal. That is why in a complex model, as we say in the future work, we should take that surface parameter in consideration.

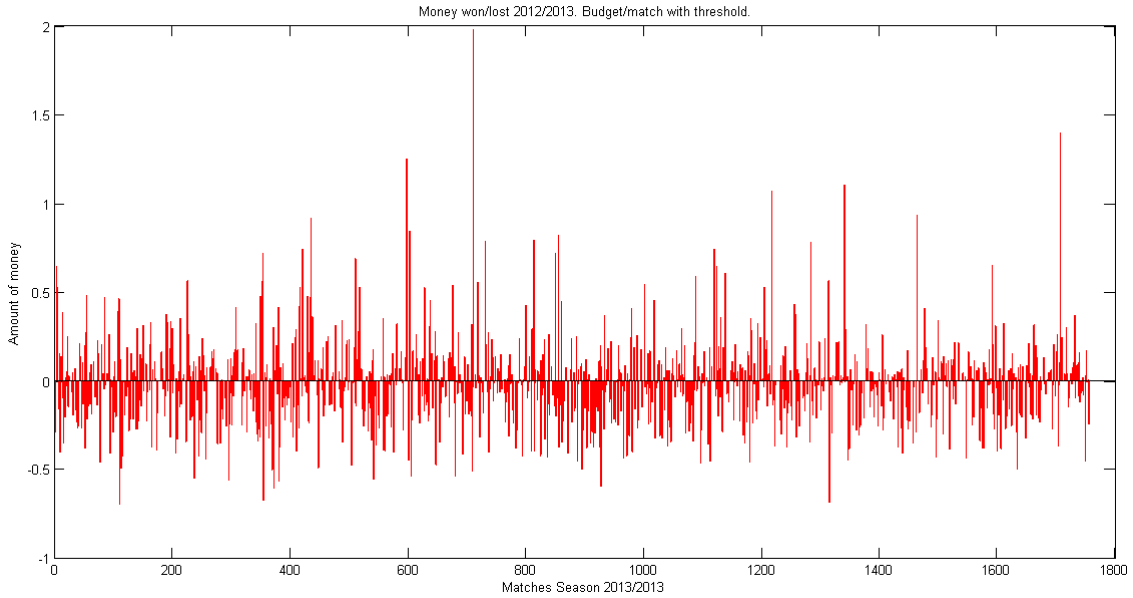

Figure 5.4: Money win/lost 2012-2013. Budget/match with threshold.

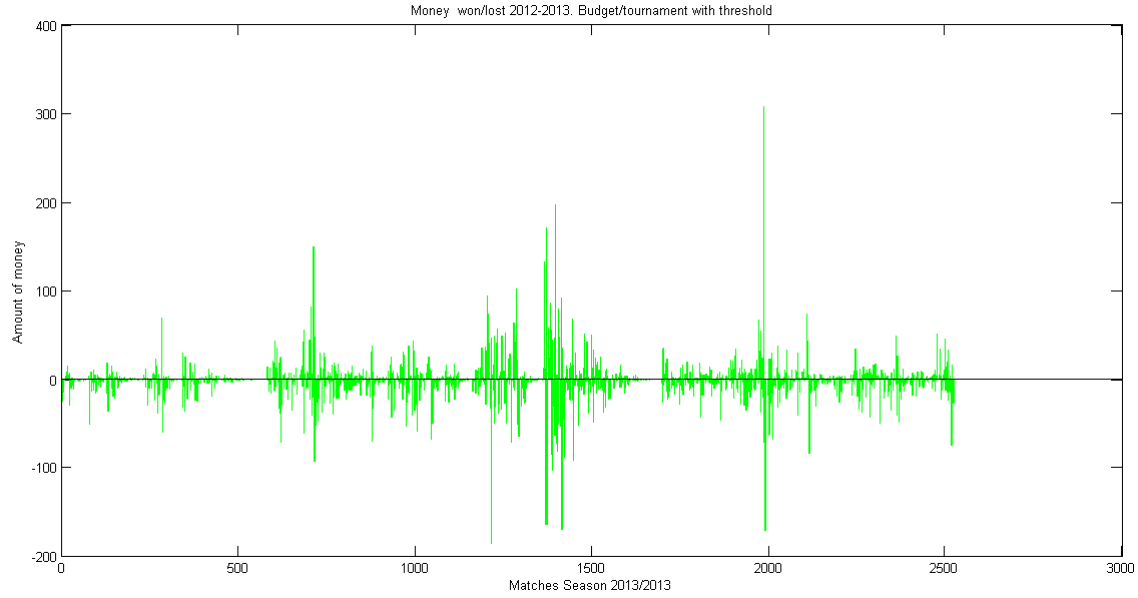

Figure 5.5: Money win/lost 2012-2013. Budget/tournament with threshold.

5.2 Qualitative results

We present here some important data such as the players, the evolution of their parameters along the season and how we collect information in this process that makes are model more precise when predicting the outcome of the latest tennis events. We will also evaluate using an histogram the probabilities estimated it self, and we will compare the to the ones given by the bookmakers in the database used.

Firstly we are going to see some information about the quality of the players, we can notice in the Figures 5.6, Figure 5.7 and Figure 5.8 the evolution of some of the players along the seasons evaluated with the model and how the position of these players is almost the same as the one given by the ATP official ranking.

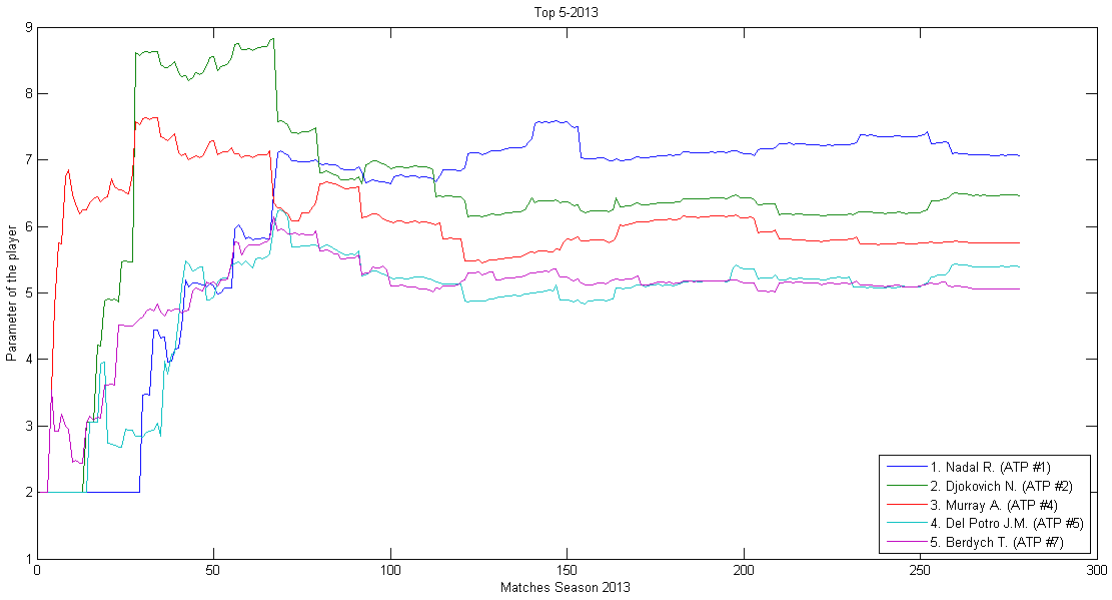

Figure 5.6: Top 5 players-2013.

As we said, here is the evolution of the Top 5 (Figure 5.6) and Top 10 (Figure 5.7) players in the 2013 Season. We can appreciate how some players, such as Nadal[2], start in a lower position of the ranking, and ends up in the first ones, in this case the number 1.

[2] Nadal started injured but end it up doing a great season. Our model also considers this fact when calculating the parameters.

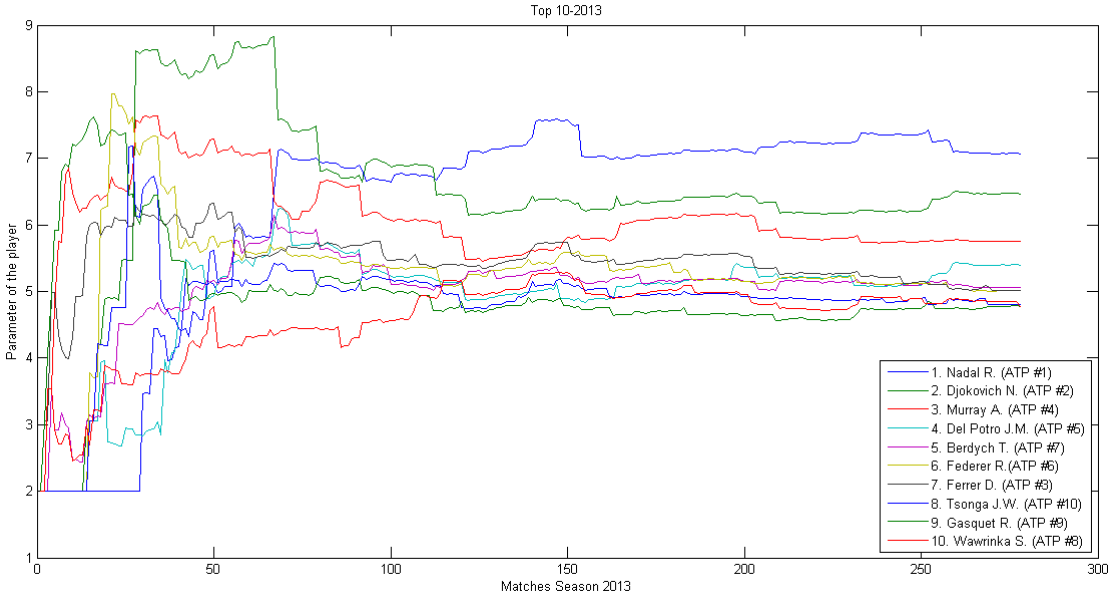

Figure 5.7: Top 10 players-2013.

Should be noted that our model considers the top 10 players the same ones as the ATP[3] does, even though our model takes into account the relationships between all the players, knows how good are the players facing, while the ATP gives points to the players without considering their ranking.

We can also see how in the official ranking of the ATP World Tour happens the same, at the end of 2012 and of 2013 the top 5 has changed. And we can notice the clear similarity between both rankings, the official ATP (Figure 5.8 and Figure 5.9) and the one we obtained by implementing our model. [32]

Now we will see some of the rest of the player in this season 2013, not only the top players (Figure 5.10). In this Figure we can appreciate how for players who are not so good is hard to determine an accurate position in the ranking. This is because they have played less matches than the ones who have better ranking, as they should get to quarters/semi-finals/finals while the others do not.

Another way to see whether our model predicts the outcomes of the events rightly is with the histogram in Figure 5.11. It is a normalized histogram

[3]The ATP ranking not necessarily the correct one. Indeed our ranking achieved with the parameter estimated could be a better approximation to the real quality of the players.

31.12.2012 ▾ Top 100 ▾ Todos los Países ▾ Ir

Ranking, Nombre, País	Puntos	Movimiento	Torneos Jugados
1 Djokovic, Novak (SRB)	12.920	0	18
2 Federer, Roger (SUI)	10.265	0	21
3 Murray, Andy (GBR)	8.000	0	20
4 Nadal, Rafael (ESP)	6.690	0	18
5 Ferrer, David (ESP)	6.505	0	25
6 Berdych, Tomas (CZE)	4.680	0	24
7 del Potro, Juan Martin (ARG)	4.480	0	23
8 Tsonga, Jo-Wilfried (FRA)	3.490	0	26
9 Tipsarevic, Janko (SRB)	2.990	0	28
10 Gasquet, Richard (FRA)	2.515	0	23

Figure 5.8: Official Ranking Top 10 players-2012

30.12.2013 ▾ Top 100 ▾ Todos los Países ▾ Ir

Ranking, Nombre, País	Puntos	Movimiento	Torneos Jugados
1 Nadal, Rafael (ESP)	13.030	0	20
2 Djokovic, Novak (SRB)	12.260	0	18
3 Ferrer, David (ESP)	5.800	0	24
4 Murray, Andy (GBR)	5.790	0	19
5 del Potro, Juan Martin (ARG)	5.255	0	21
6 Federer, Roger (SUI)	4.205	0	19
7 Berdych, Tomas (CZE)	4.180	0	24
8 Wawrinka, Stan (SUI)	3.730	0	25
9 Gasquet, Richard (FRA)	3.300	0	25
10 Tsonga, Jo-Wilfried (FRA)	3.065	0	21

Figure 5.9: Official Ranking Top 10 players-2013

corresponding to the proportion of observed events for which the predicted probabilities are comprised between 0 and 1, every 0.1 (xlabel), across the seasons for all the implementations of the model done in this project.

What corresponds to Bet365 will be the model of a bookmaker, were all the predicted probabilities are evenly distributed, being the ones in the range of 0.9, for example, the complementary of the ones in the range of 0.1, that is that the sum of this proportions will give us 1. Taking this as a reference we can see how our model closely resembles the model of the bookmaker, specially season 2012-2013. The mainly difference are in the lowest probabilities, range 0.1-0.2, where our model provides a greater

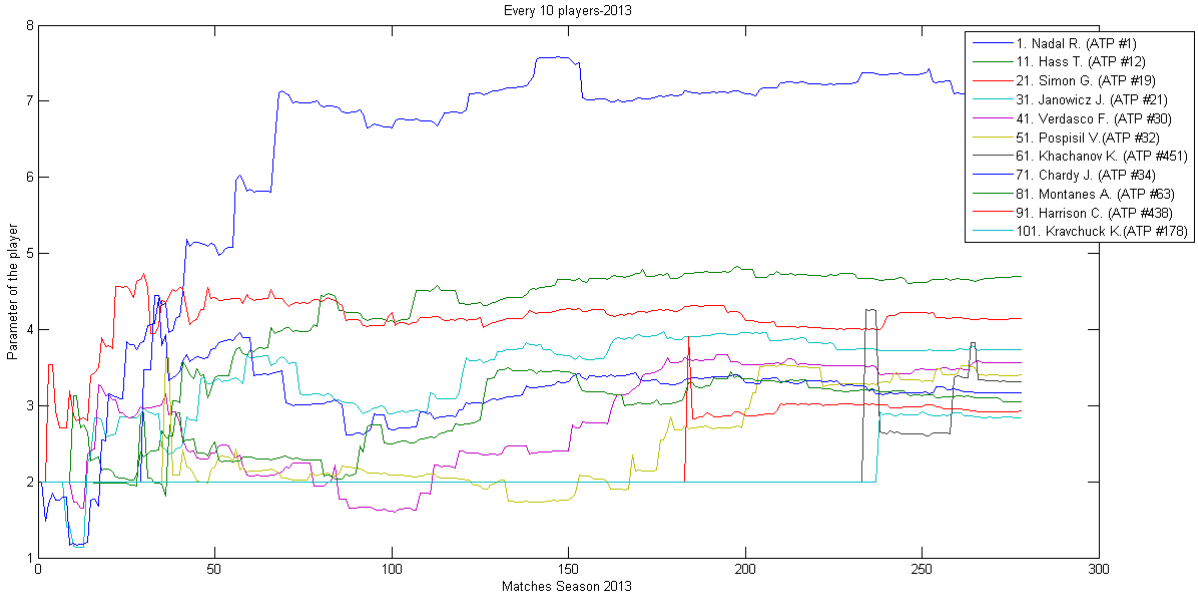

Figure 5.10: Every 10 players-2013.

number of estimated probabilities.

We can also evaluate the probabilities estimated with the histograms in Figure 5.12, Figure 5.13 and Figure 5.14, which represent the probabilities calculated by a bookmaker, Bet365, and the ones we have done. When we first begin the season, we do not know anything about the players, so most of our probabilities will be around 0.5, but as we go through the matches and the season, our model learns, and by the end of it we will have more variability in values and less mass in the 0.5 area.

We can see how Bet365 histogram (Figure 5.12) has less mass in the center area, as we explain before, while ours has more especially the histogram that contains the whole season (Figure 5.13). As we said, by the end of the season (Figure 5.14) we know much more about the players and our probabilities get better.

If we had obtained a histogram with a lot of data in the middle area, around 0.5, it might been caused because of the prior. The prior would be very restrictive and unless we had many events of the players, their probability will be around 0.5. A restrictive prior has the following shape:

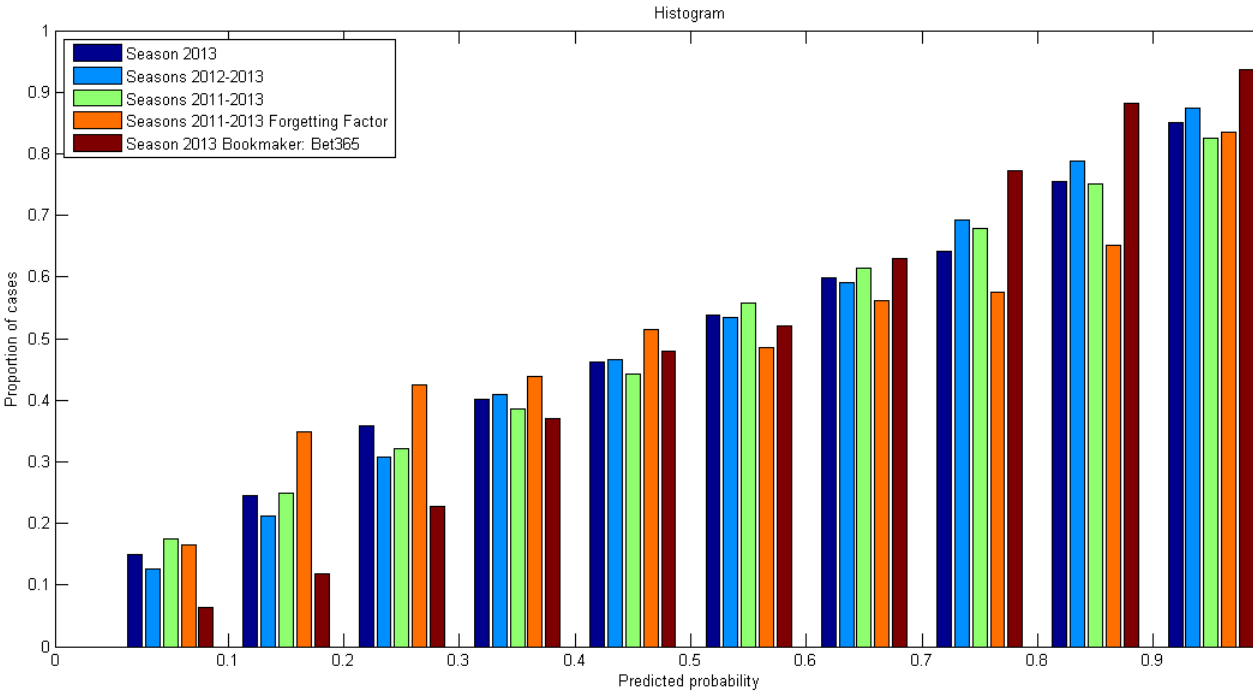

Figure 5.11: Histogram

While a prior with less restrictions will have a smoother shape and will led us estimate probabilities more varied. This means that although we do not have many events for all the players, with the results we obtain of the matches we can approximate our probabilities to ones calculated with a model that contains more matches to estimate this probabilities.

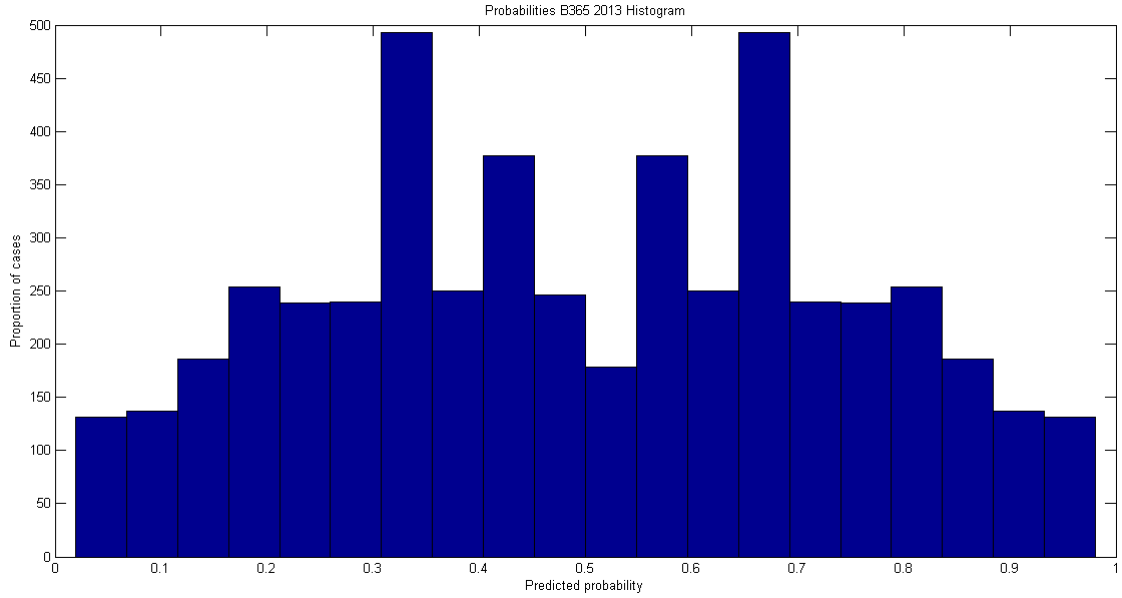

Figure 5.12: Histogram b365

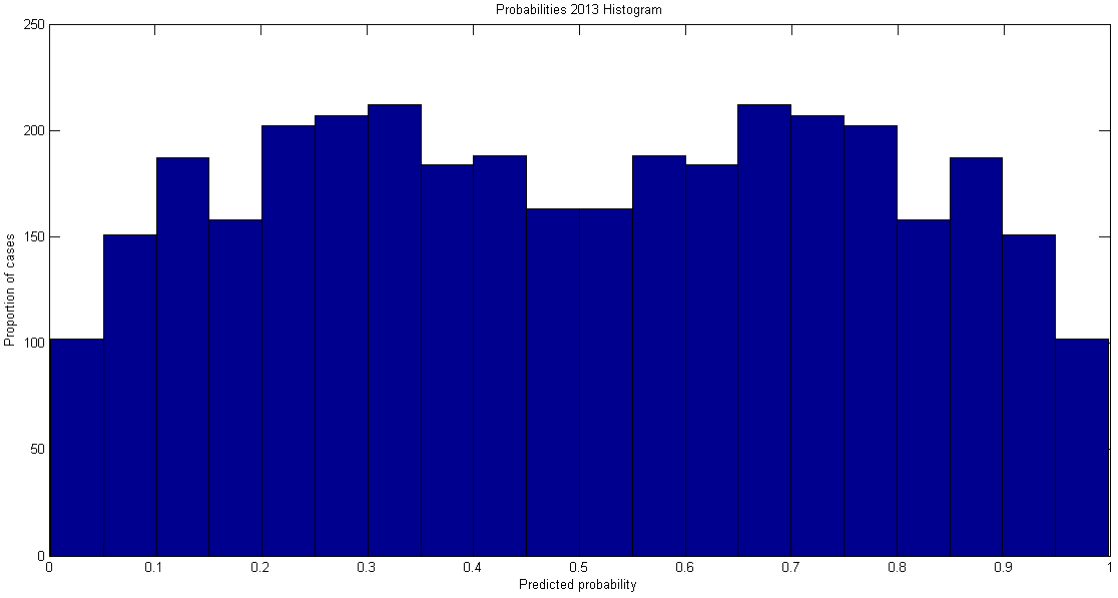

Figure 5.13: Histogram model

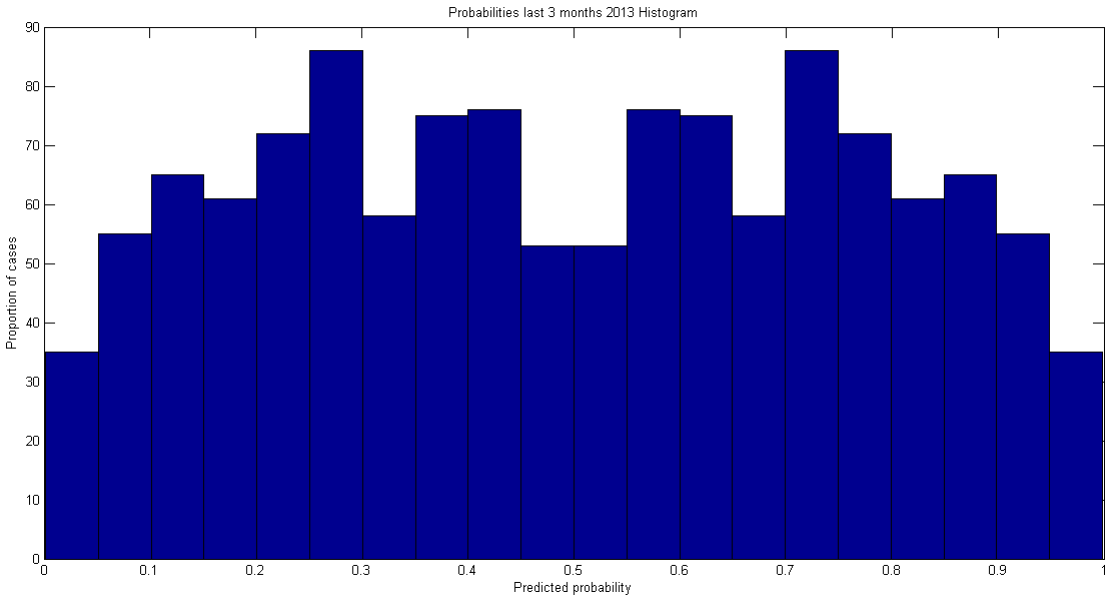

Figure 5.14: Histogram 3 months

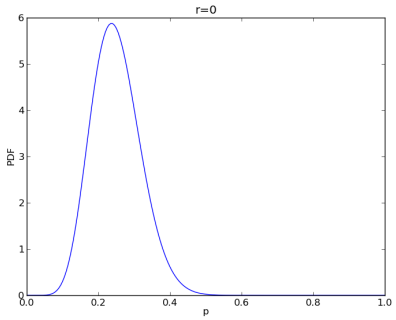

Figure 5.15: Prior

Chapter 6

Conclusions and future work

In this chapter the main conclusions reach at the end of this project, as well as the future developments are collected.

6.1 Conclusions.

The project was focused on the prediction of the results of a sporting event. More specifically what was attempted was to predict the winner of each match in the tennis Men's Singles category.

The main conclusions that can be drawn after assessing the project carried out are:

- Our main goal for this project was to get a better system for estimating the odds than the bookmakers for tennis matches, but through the calculation of probabilities it is not possible to affirm that our system is better.
- Even though there were many parameters we did not consider, as the statistics of each player in their play, their personal situation, the physical condition, we obtained a good model that learns long the seasons and by the end of them the predictions are really good. We are also able to evaluate in a good way the players, as we have seen in the

results we hit the top 10 players although not all of them in the same position.

- The matches evaluated were played by only two players, of whom we just know which two players were going to play that event, but not many other factors, so we ended up with a model that assumes a dimensional hierarchy, that is only one degree of freedom. This is why our model might not be so precise. What this dimensional hierarchy means is that in our model we cannot contemplate the idea of a player with a position 2 in the ranking,for example, winning the first player in the ranking, as happened some time when R. Nadal was the number 2 and most of the times won R. Federer, the number 1 in the ATP ranking.

- Noting the results of our model we can see that it is possible to obtain a yield introducing us to the world of gambling. The main key to this success is to get a more or less successful winning percentage, then compare it not only with the odds of a bookmaker but with the most representative ones, and bet with the one that can get us more profit. This gives us a large margin for profit especially in the long term.

- Since the beginning of the project was known about the difficulty that would be getting a model that gained very high percentages of correct answers. However, we get good results by the end of the seasons, as we have seen in the BNP Paribas Master or even the Master Cup in London. But to be able to achieve higher percentages as the bookmakers we would need a complex model, for example, that each player instead of having one parameter associated, had a vector of parameters, and this way we might get closer to the bookmakers.

The hardest task of the project was to create the database necessary to implement our model, those are the parameters. It took to long to obtain then not only because of the difficulty that was to develop the algorithm to do it, but because of the time that took Matlab to execute it.

Finally, note that the objectives explained at the beginning of the project were achieved. A database was created with the selected key variables to be able to obtain a high amount of information, about the matches, the players and the bookmakers, to work with. After we created our algorithm to develop our model with all the variations and extensions that we could do, taking into account the limitations of the project. We develop several results taking

three tennis seasons separately, changing the budget to determine if we could obtain more benefits, and computing the forgetting factor to obtain always the latest and updated information of the players. Moreover, using several rates of the bookmakers and the control and management of our bench by the algorithm of Kelly, it was possible to improve the final benefit over all the matches.

6.2 Future work.

From the fundamental purpose used in this project we could raise other options that expand the original idea or towards other goals from expanding our horizons.

Here are some ideas that could be done:

- The first idea will be to extend this same model to other modes of tennis, such as women's tennis, tennis doubles.
- Use the official ranking given by the ATP World Tour, as information to infer the parameters that we calculate in our model.
- Expand the number of data to use in the model, not only how good or bad they are, but maybe the surface, as we tried to do, or instead of predicting the final result do it by sets. This expansion in the number of parameters will make the model more complex and the predictions will be more accurate. A complex model will led us do a study to predict interactively, that is to detect during the course of the game at what time the critical moments occur and appropriate odds to bet. In this way we can bet, in real time, when it was more profitable for a player or another.
- In the same model line expansion try to predict the winner of the match but go farther than that and try to predict the number of sets, winner per set, winner of the tie break, duration of the match, etc.
- Continue to study in the field of other sports that may be beneficial or have more room for profit in our bets.

Bibliography

[1] "Gambling History, from the beginning.". Gambling Info. Retrieved June 23, 2011.

[2] July 29, 2012 (2012-07-29). "title=The History of American Gambling". Casino.org. Retrieved 2012-09-22.

[3] Sports Betting: Past, Present and Future - Part 1 by Jeremy Martin.

[4] "A Look At Bookmakers Sponsoring Football Clubs". ReliableBookies.com. 2013-03-19. Retrieved 2013-03-19.

[5] Gambling Commission Gambling industry statistics April 2009 to September 2012

[6] McCullagh, Peter; Nelder, John (1989). Generalized Linear Models, Second Edition. Boca Raton: Chapman and Hall/CRC. ISBN 0-412-31760-5.

[7] Ley 13/2011, de 27 de mayo, de regulación del juego.

[8] Rebecca Goldin (2007). "Odds Ratios". George Mason University. Retrieved 11 July 2014.

[9] "Betting School: Understanding Fractional Decimal Betting Odds". Goal. 10 January 2011. Retrieved 27 March 2014.

[10] "Betting Odds Format". World Bet Exchange. Retrieved 27 March 2014

[11] "Fractional Odds". http://betstarter.com/. Retrieved 27 March 2014

[12] Fairleigh Dickinson University's PublicMind, (February 21, 2011). Sports Betting, Sure Thing; Internet Betting, Nyet!

[13] Betting in Tennis. Web-site: http://www.bettingexpert.com/es/academia/apostar-distintos-deportes/tenis

[14] Kelly´s Criterion. Kelly, J. L. (1956). "A New Interpretation of Information Rate". Bell System Technical Journal 35 (4): 917926. doi:10.1002/j.1538-7305.1956.tb03809.x

[15] J. L. Kelly, A new interpretation of information rate. IRE Transactions on Information Theory, 2(3): 185-189, 1956.

[16] Database. Web-site: http://www.tennis-data.co.uk/alldata.php

[17] David J.C. MacKay. Information Thoery, Interference, and Learning Algorithms. University of Cambridge. Version 7.2 2005.

[18] Carl Edward Rasmussen and Zoubin Ghahramani. Lecture 1 and 2. Probabilistic Regression. Machine Learning. University of Cambridge.

[19] Carl Edward Rasmussen and Zoubin Ghahramani. Lecture 10. Discrete Distributions. Machine Learning. University of Cambridge.

[20] M. J. Maher, Modelling association football scores, Statis. Neerland., 36: 109-118, 1982

[21] D. M. Blei, Build, compute, critique, repeat: data analysis with latent variable models, 2013.

[22] D. Borsboom, G. J. Mellenbergh and Jaap van Heerden, The theoretical status of latent variables. Psychological Review, 2003.

[23] G. Baio and M. A. Blangiardo, Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics, 37: 253-264, 2010.

[24] D. Karlis and I. Ntzoufras, Bayesian modelling of football outcomes: using the Skellams distribution for the goal difference, 2007

[25] J. G. Skellam, The frequency distribution of the difference between two Poisson variates belonging to different populations. Journal of the Royal Statistical Society. Vol 109, 1946.

[26] M. T. Boudjelkha, Extended Riemann Bessel functions. Discrete and continuous dynamical systems. 121-130, 2005.

[27] Rubin, Donald B.; Gelman, Andrew; John B. Carlin; Stern, Hal (2003). Bayesian Data Analysis (2nd ed.). Boca Raton: Chapman Hall/CRC. ISBN 1-58488-388-X. MR 2027492

[28] Jack David Blundell, Numerical Algorithms for Predicting Sports Results. School of Computing, Faculty of Engineering.

[29] A.P.Rotshtein, M.Posner, A.B. Rakityanskaya, Football prediction based on a Fuzzy model with genetic and neural tuning. Cybernetics and Systems Analysis, Vol. 41, No. 4, 2005.

[30] A.Joseph, N.E.Fenton, M.Neil,Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems Volume 19, Issue 7, November 2006, Creative Systems.

[31] Log Loss. Web-site: https://www.kaggle.com/wiki/LogarithmicLoss

[32] Ranking. Web-site: http://es.atpworldtour.com/Rankings/Singles.aspx?d=31.12.2012r=1c=.

[33] Tetko, I. V.; Livingstone, D. J.; Luik, A. I. (1995). "Neural network studies. 1. Comparison of Overfitting and Overtraining". J. Chem. Inf. Comput. Sci. 35 (5): 826833. doi:10.1021/ci00027a006

Appendix A

Plan and Budget

In this chapter we are going to talk in detail about the budget calculated for this project. We will take int account the costs in personal level but not the material costs as since it is assumed that what is needed is the development application level, which only takes into account the cost per engineer.

A.1 Plan

First of all, we will detail the different parts in the project:

- Phase 1: Documentation: Study all the data given and search for information about matching learning and several probabilistic models in order to decide what to develop.
- Phase 2: Creating databases: Once the database is given with all the players, matches, tournament, outcomes, etc. We create our own database and develop a probabilistic model that let us calculate the parameters needed.
- Phase 3: Implementation of the model designed: With the model design we use it to calculate the parameters of the players in order to get the probability of winning/loosing in a occurring event.

- Phase 4: Extensions: After getting the basic model we develop some extensions to get a better performance of the model.
- Phase 5: Simulations and results: We made all the simulations with and without the extensions and obtained the results.
- Phase 6: Writing the memory of the project: Once we have all the work done we write the document.

Here is a Gantt Diagram to illustrate the work done and time spent in the project:

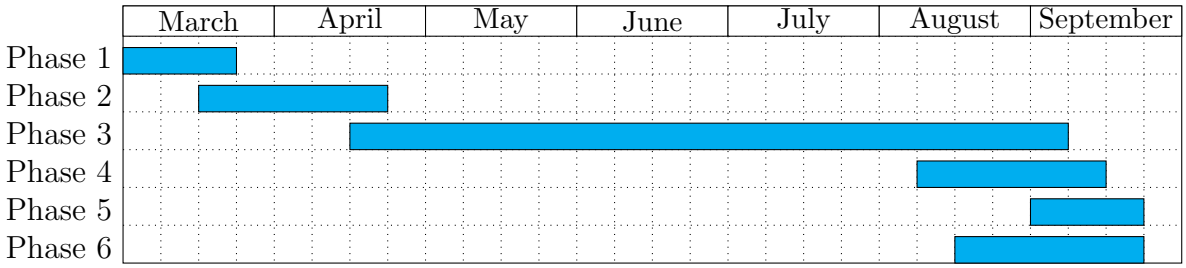

Figure A.1: Gantt Chart.

A.2 Budget

Having only one engineer working, knowing that I did not work all days and not every day the same hours, in average 4 hours per day.

Under this assumption, based on the values ??described in the report of COIT (Official College of Telecommunications Engineering) (31) we can establish that the average monthly salary of a Telecommunications Engineer, with less than 5 years of experience is about 2083.33 €.

1. Author:
   Lourdes Gómez González

---

2. Department:
   Communication Systems Engineering

---

3. Project Description:

Title:	Análisis de resultados para apuestas deportivas: tenis.
Duration (months):	7
Indirect Costs Rate:	21%

4. Project's Total Budget (values in Euros):
   14583.31 €

5. Budget Breakdown (Direct Costs):

STAFF						
Surname, Name	N.I.F.	Category	Dedication (Men Month)	Men Month Cost	Cost (Euros)	Signature
Gómez González, Lourdes	05326372D	Engineer	7	20833.33	14583.31	

6. Costs Summary:

Staff	14583.31
Indirect Costs (Euros)	3062.4951
Total (Euros)	17645.8051

This project's total budget add up to the amount of 17645.8051€.

Leganés, September 24th, 2014

Project's Engineer,

(Signature)

Fdo. Lourdes Gómez González.