UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA IN

STATISTICA E INFORMATICA

TESI DI LAUREA

STATISTICS AND GAMBLING: THE NFL PREDICTIONS

RELATORE: Ch. mo Prof. STUART COLES

CORRELATORE: Ch. mo Prof. GUIDO MASAROTTO

LAUREANDO: NICOLA NOVELLO

ANNO ACCADEMICO 2009/2010

To my granny Rosa

INTRODUCTION

In the last years, the gambling market has been growing exponentially in any kind of settings like casino, poker (in particular the Texas Hold'em, but many other versions are going to emerge), gossip, but absolutely the main branch of this special market consists in sports.

I mentioned the casino and the poker because actually the probability is involved in almost all those games, but certainly, because of their particular structures, it is not possible to implement valid models to predict such results (for example, the prediction of the *roulette*'s numbers).

On the contrary, the probability and consequently the *Statistics* may be useful to predict the outcome of a sport match: this is actually the main role of the *bookmaker*.

The wide diffusion of the Internet has helped significantly all the firms that provide this type of supply; actually, in those web sites, it is possible to gamble 24 hours a day and 7 days out of 7 (the only constraint is that the customer must have at least the legal age in the jurisdiction of his/her state), with the natural consequence that a lot of people have started to play comfortably from their houses (throughout the use of the *PC*).

In the recent months, my passion for all the sports and the large recent diffusion of the betting market that concerns on sports, even in my country (*Italy*) interested me so much; this fact brought me to implement statistic methods to forecast sport outcomes.

By the way, prediction problems in many fields (e. g.: finance, political elections, sports etc.) are complicated by the presence of several sources of variation on which a predictive model must account as a valid tool to summarise the reality in exam.

In detail, this thesis develops models that should be able to forecast the correct results of the NFL (National Football League) matches. This professional league is the most important championship of the American Football.

NFL is one of the 4 main national sports in the USA, with NBA (National Basketball Association), NHL (National Hockey League) and MLB (Major League Baseball).

For the NFL games, team abilities may vary year to year due to changes in personnel and to the whole strategy. In addition, team performance may also vary depending on the site of the game; that's the reason why my analysis is particularly focused on the study of the *home effect.*

This kind of predictions exploits many and different explicative variables in order to achieve the main goal: the correct prediction of the American football outcomes.

Furthermore, this thesis discusses different approaches for modeling NFL scores using a Normal linear state-space model that accounts for these important sources of variability.

First of all, my mission is to introduce the sources and discuss about the relation between the *Statistics* and the betting world; after that, I'm going to give a short description of the NFL, with its rules, its scoring mechanisms and I'll have a glance on the legislation that concerns the betting market in the USA.

The main part of this thesis provides an overview of the statistic methods used to predict the NFL outcomes with the relative statistic outputs. Subsequently, I compared such statistic results to the real outcomes in order to evaluate the quality of the estimates via *MSE (Mean Square Error)*.

At the end there is a chapter that includes the conclusions and the personal comments.

Furthermore, I decided to include an appendix where I show the R code associated to the main functions that I had implemented to achieve the results.

All the analyses on the data have been possible thanks to the *open source* programme R 2.10.1 and thanks to Microsoft Excel 2007 (Office 2007 *package*).

INDEX

CHAPTER ONE: STATISTICS AND GAMBLING

This first chapter aims to point out the strong relation between the gambling world and the *Statistics*; in fact, *Statistics* has a central role on this particular betting market, almost exclusively from the bookmaker point of view (even if there are many expert gamblers that throughout the use of some statistic methods aim to win money).

Nowadays, with the large amount of data available in electronic files, especially about descriptive statistics of sports (just browsing web pages it is possible to find many databases or more specific *datamarts** suitable for the study of every kind of sport's data), it is possible to study many factors that are potential valid variables to determine the correct outcomes of the matches.

1.1: RANDOMNESS IN GAMES

Statistics has surely an important role in all this kind of predictions that concern games, because this kind of study is closely connected with the probability involved in such events; these probabilities can be estimated throughout the results of some specific statistic models.

This affirmation means that throughout the study of the data in the past and thanks to the probability, it's possible to predict the outcomes of the future matches.

Since the 90's, with the rapid development of the *computer science*, it has been possible to analyse billions of data (or even more) and easily compare statistic models based on different approaches. I would like to emphasize that before this "computer assisted age" it was almost impossible to think about such complex statistic models because it was impossible (or at least really difficult) to achieve the main aim of the analyses: the results.

This is one of the main reason why *Statistics* has been successful in the last years thanks to the fundamental help of the new technologies (in terms of *Informatics* and consequently statistic software, such as ).

In addition, the possibility to implement algorithms based on many sources of variability provides models that fit better the reality in exam (because the analysis is consequently more complex), if compared to the "simpler" statistic models that the experts of *Statistics* had used in the past (approximately before the 90's).

1.2: STATISTICS AS A TOOL FOR BETTING

In the betting world, there are 2 main antagonists: the *bookmaker* and the gambler; the first one provides the prices related to every kind of bet (obviously with the statistical support), the second one evaluates the bookmaker's supply and decide how it is more convenient to bet (most of them without the statistical supports).

Probably, this is the main reason why, most sports bettors are overall losers as the bookmakers *odds** are fairly efficient; as a consequence of this, the gambling world (in terms of a real market) has developed so rapidly in the recent years.

The general role of the *bookmaker* is to act as a market maker for sports wagers, most of which have a binary outcome: a team either wins or loses. The bookmaker accepts both wagers, and maintains a spread point which will ensure a profit regardless of the outcome of the wager.

The "war" between the bookmakers and the gamblers results a bit unfair; in fact the "bookies" (obviously before the games) have always a systematic advantage on the bettors because the sum of the probabilities given by the *odds* never sums to 1 (as it should be for a fair "war"), but it reaches a total probability around 90%; the 10% *gap* represents the systematic advantage of the market makers.

One important application which has emerged, is the use of the statistical models by bookmakers to determine the *odds* (practically the probabilities), and by expert gamblers aiming to win money throughout the use of statistical models to identify advantageous bets given by the market prices.

The largest sporting market for gambling is currently the traditional football; in literature, there are many assays that talk about traditional football bets rather than NFL (American football) bets. This is another reason why I decided to implement something related to this second sport.

In any case, it's necessary a dynamic modeling approach, to evaluate the performances of each team over the time and to take care about the site of the game (effectively, in every sport, the *home effect* results a relevant factor to establish correctly the final outcome of a match).

In addition, the study of this kind of data in the past gives important information relatively to the features (characteristics described from different variables) of the teams involved in the competition that I'm studying: by the way, this is a good starting point to analyze every phenomenon throughout the study of the *time series*.

In specific, the NFL web sites offer an enormous amount of such descriptive statistics related to any aspect of this game (this fact is due to an exaggerate passion of the Americans for this kind of information); these well-equipped databases provide a lot of information that the statistician might use in the right way to achieve reassuring results.

1.3: STATISTICS OF AMERICAN FOOTBALL

At the beginning of this work, I browsed the literature about the statistics studies that deal with NFL matches to have an general idea of the reality that I'm going to study, but actually I didn't find so many articles that write about this kind of predictions.

By the way, I want to underline that this thesis follows in some parts (specifically in the chapter 4) the approach introduced by Mark E. Glickman (Assistant professor, Department of Mathematics – Boston University) and Hal S. Stern (Professor, Department of Statistics – Iowa Sate University) in their article "A state space model for National Football League scores" published in March 1998 in the Journal of the American Statistical Association, that talk in specific about the NFL games.

To analyze this type of data I decided to use a data window of 8 years, sufficient, in my opinion, to highlight an eventual specific pattern or seasonality in these sport results over the time.

In specific, this thesis develops some predictive models for NFL game scores using data from the period 2002-2009.

The main aim of this work is to reduce, as much as possible, the *gap* between my predictions and the *Las Vegas line** (bookmakers' predictions).

The kinds of bet that this thesis analyzes are the under/over bets inherent in the difference of the points and the under/over bets inherent in the sum of the points.

The utopia would be to "beat" the *Las Vegas line*; anyway, the results will show us that the results of my models are really closed to such line, but still slightly less reliable.

Moreover, I may consider the results well achieved because my models are surely less complex than those ones used by the bookmakers, but they are still able to reach the prefixed goals (in terms of quality of the estimates, expressed via Mean Square Error index).

Furthermore, since I'm not a NFL *tipper*, I decided to analyse such data just from the statistical point of view, without taking care about the specific dynamics related to this sport.

For dynamics in this specific case, I mean changes in the short period (for example changing in staff, managing injuries with specific players' variables, etc).

Talking about the relevance of the injuries, the most important player of a team is the quarterback, and that's why the *bookmakers* take really care about the physical condition of this specific player (and obviously if he's playing or not), because this information obviously moves the spread points of the *Las Vegas line*.

In other words, this approach consecrates the power of *Statistics*, the ability to analyse something and achieve the prefixed aims, exclusively with the study of the "raw" data (subsequently suited to the analysis), even without having an excellent knowledge of the reality in exam.

1.4: DIFFERENCES BETWEEN TRADITIONAL FOOTBALL AND NFL

As I said in the previous paragraph, most of the statistical applications (articles, assays, thesis) about sport betting are focused on the traditional football outcomes (Serie A, Spanish Liga, etc.).

In this specific case the statistician analyses the number of goals scored by each team, aiming to determine the correct outcomes of the traditional football games involved in the analyses.

Unlike the traditional football case, the NFL predictions consider the number of points scored by each team, aiming to define the correct outcomes of the games involved in the analyses.

The main difference is due to the scoring mechanism and consequently to the different scale of points between the NFL and the traditional football results.

The traditional football goals are described by the *Poisson*'s distribution: it represents a discrete probability that describes the number of events occurred in a known average rate, independently of the time since the last event.

In detail, if $X \sim$ Poisson (λ), thus, $E[X] = VAR[X] = \lambda$

On the contrary, the distribution adopted for the NFL results is the *Normal* (or *Gaussian*) one: it's an absolutely continuous probability distribution where it is necessary to specify the mean (μ) and the variance ($\boldsymbol{\sigma}^2$).

In detail, if $Y \sim N(\mu, \boldsymbol{\sigma}^2)$, thus, $E[Y] = \mu$ and $VAR[Y] = \boldsymbol{\sigma}^2$

Obviously, the Normal approximation can't be a perfect approximation for the NFL scores because of the nature of the response variable: the variable that represents the difference (or the sum) of the points scored (between the home and the away team) is a discrete data and the Normal distribution is, clearly, absolutely continuous.

This general comparison between traditional and American football underlines that the predictions related to these 2 sports exploit different modeling structures and as a consequence of this, the approaches are totally different.

The main advantage of the *Poisson* approach (for the traditional football) is that throughout the manipulation of the only and unique parameter λ, it is possible to control, at the same time, expectation and variance.

In the NFL predictions, with the *Normal* approach I have to manage 2 different parameters; specifically, it is necessary to decide how to handle the variance ($\boldsymbol{\sigma}^2$).

Actually there's no reason to expect that the finite variance changes over the time and among different matches; as a consequence of this, in the following analyses I assumed homoscedasticity in all the *Normal* models considered.

1.5: DETAILS (*)

DATA MART

A *data mart* is a subset of an organizational data store, usually oriented to a specific purpose or major data subject, that may be distributed to support business needs.

They are analytical data stores designed to focus on specific business functions for a specific community within an organization.

Data marts are often derived from subsets of data in a *data warehouse*, though in the bottom-up *data warehouse* design methodology; it is created from the union of organizational *data marts*.

ODDS

In gambling, the *odds* do not represent the true chances that the event will occur, but they are the amounts that the *bookmaker* will pay out on winning bets.

In formulating his *odds* to display the *bookmaker* will have included a profit margin which effectively means that the payout to a successful bettor is less than that represented by the true chance of the event occurring.

Profiting in gambling involves predicting the relationship of the true probabilities to the payout *odds*. Sports information services are often used by professional and semi-professional sports bettors to help achieving this goal.

LAS VEGAS LINE

This imaginary line contains the predictions of the *bookmakers*; in specific for NFL, the market makers provide the difference of the points (between the home team and the away team) and the sum of the points (between the home team and the away team).

This is useful to assign the *odds* related to these 2 kinds of bet, from the *bookmaker* point of view.

In addition, this information is fundamental to evaluate the efficiency of the models that I carried out in this thesis.

Note: In my analysis, these 2 lines are available for the years 2006, 2007 and 2008; as a consequence of this, I underline once again that the main aim of my models is to predict the sum and the difference of the points scored between the home team and the away team.

CHAPTER TWO: HOW DOES NFL WORK?

The NFL is the highest level of professional American football. The league currently consists of 32 teams from the United States of America. The league is divided evenly into two conferences — the American Football Conference (AFC) and National Football Conference (NFC), and each conference has 4 divisions that have 4 teams each. The NFL is organized as an unincorporated association of its 32 teams. The NFL is by far the most attended domestic sports league in the world by average attendance per game (with 67,509 fans per game in the regular season 2009).

All the information contained in this second chapter is necessary to understand what I have considered in my analysis and which is the meaning of the variables in exam.

2.1: THE STRUCTURE

Since 2002, The NFL season features the following schedule:

- a 4-game exhibition season (or preseason) running from early August to early September that I don't take into consideration in the analysis because these matches are not official (they are just some friendly games to introduce the teams to the championship);
- a 16-game, 17-week regular season running from September to December or early January that is the main part of the data;
- a 12-team Single-elimination playoff beginning in January, culminating in the Super Bowl (the last match of the season and one of the most important sport event of the year in USA) in early February.

One of the main reason why I analyse this data is that all the teams remain the same over the years (no relegations nor promotions); this special characteristic of this American sport, is surely a point of strength to analise the *time series* associated to such matches (in detail I study the data from 2002 to 2009).

2.2: THE TEAMS

The 32 teams involved in the competition divided by conference (A is AFC, N is NFC) and zone are:

East

- Buffalo Bills
- Miami Dolphins
- New England Patriots
- New York Jets

North

- Baltimore Ravens
- Cincinnati Bengals
- Cleveland Browns
- Pittsburgh Steelers

South

- Houston Texans
- Indianapolis Colts
- Jacksonville Jaguars
- Tennessee Titans

West

- Denver Broncos
- Kansas City Chiefs
- Oakland Raiders
- San Diego Chargers

East

- Dallas Cowboys
- New York Giants
- Philadelphia Eagles
- Washington Redskins

North

- Chicago Bears
- Detroit Lions
- Green Bay Packers
- Minnesota Vikings

South

- Atlanta Falcons
- Carolina Panthers
- New Orleans Saints
- Tampa Bay Buccaneers

West

- Arizona Cardinals
- San Francisco 49ers
- Seattle Seahawks
- St. Louis Rams

2.3: HALL OF FAME

The following table shows the winners of the Super Bowl in the 8 years in analysis.

Table 2.1

Years	Team
2002	Tampa Bay Buccaneers
2003	New England Patriots
2004	New England Patriots
2005	Pittsburgh Steelers
2006	Indianapolis Colts
2007	New York Giants
2008	Pittsburgh Steelers
2009	New Orleans Saints

In detail, in the next table I show the Super Bowl matches played in those 8 years, with the associated points:

Table 2.2

Years	Teams	Points	Gap
2002	Oakland Riders	21	
	Tampa Bay Buccaneers	48	+27
2003	Carolina Panthers	29	
	New England Patriots	32	+3
2004	New England Patriots	24	+3
	Philandelphia Eagles	21	
2005	Seattle Seahawks	10	
	Pittsburgh Steelers	21	+11
2006	Indianapolis Colts	29	+12
	Chicago Bears	17	
2007	New York Giants	17	+3
	New England Patriots	14	
2008	Pittsburgh Steelers	27	+4
	Arizona Cardinals	23	
2009	New Orleans Saints	31	+14
	Indianapolis Colts	17	

From these data, I can conclude that New England Patriots, Pittsburgh Steelers and Indianapolis Colts look like the best teams in those years analysed.

2.4: RULES AND SCORING MECHANISMS

2.4.1: THE FIELD

The field measures 100 yards long and 53 yards wide. Little white markings on the field called yard markers help the players, officials, and the fans to keep track of the ball.

Probably the most important part of the field is the end zone. It's an additional 10 yards on each end of the field. This is where the points add up. When the team with the possession of the ball gets it into the opponent's end zone, it scores points.

Obviously, the team with the most points at the end of the game wins. So the offense tries to get as many touchdowns (points) as they can and the defense tries to stop them. Games are divided into 4 15-minute quarters.

This information about the field is important to interpret properly the variables available, related to the yards gained from the home and the away team.

The next paragraph underlines this aspect; obviously, the more yards are gained, higher is the probability to score touchdowns.

2.4.2: THE “YARDAGE”

All progress in a American football game is measured in yards. The offensive team tries to get as much "yardage" as it can, to move closer to the opponent's end zone. Each time the offense gets the ball, it has four downs, or chances, in which it may gain 10 yards. If the offensive team successfully moves the ball 10 or more yards, it earns a first down, and another set of four downs. If the offense fails to gain 10 yards, it loses possession of the ball.

The defense tries to prevent the offense not only from scoring, but also from gaining the 10 yards needed for a first down. If the offense reaches the fourth down, it usually punts the ball (kicks it away). This forces the other team to begin its drive further down the field.

2.4.3: THE UNITS

Each team has 3 separate units:

- the offense, those players who are on the field when the team has possession of the ball;
- the defense players who line up to stop the other team's offense;
- the special teams that only come in on kicking situations (punts, field goals, and kickoffs).

2.4.4: THE TURNOVERS

While trying to bring the ball to the end zone, the offense may accidentally turn the ball over to the defense. In other words, this important index (available as variable) represents how many times per match the offense loses the ball when it is attacking.

2.4.5: THE OFFENSE

Whichever team has possession of the ball is the offense. It is the quarterback, who is the leader of the team and the playmaker. In fact, he's the most talented player, not only he throws the ball, but also he outlines each play to his team.

The offense uses many different formations to set their players to start a play. The quarterback calls plays and leads the team. When a play is run all players work together to move the ball.

The formation can tell you what side of the field the play might be run. Offenses use extra wide receivers when they need to pass the ball, or extra blockers and running backs when they need to run the ball.

Teams use different strategies to score. Some teams like to run the ball more than pass. Some teams prefer to pass rather than run. All this depends on the type of players a team has and the style of play the coach likes to use.

2.4.6: THE DEFENSE

The job of the defense is to stop the offense. The 11 men on the defensive team work all together to keep the offense from advancing toward the defense's end zone.

The defense uses different players to help to stop the offense; in specific, it uses ends, tackles and nose guards to stop the run.

The defensive formation usually matches up with the offense. But the defense can try to do different strategies to stop the offense. The defense can move players around to cause the offense problems.

Note: I introduced such specific information because to predict sport outcomes, it is necessary to know as much as possible about such sport; this chapter wants to give an huge overview of the reality in exam.

2.5: BETTING MARKET AND LAW IN THE USA

2.5.1: THE "SHADOW" OF THE *BLACK MARKET*

American sports are emerging as important new market for gambling and it's fundamental to give a glance on the legislation that turns around this special market.

The aim of this paragraph is to explain what is the "shadow" of the *black market* on sports betting and why this is a topic that I need to cite in this thesis.

The NFL claims that sports betting would irreparably harm professional and amateur sports by fostering suspicion and skepticism about final scores of games that may have been influenced by factors other than honest athletic competition.

The actions from the NFL that the league has implemented in opposing this new law have escalated into a nationwide debate. Many people question why the NFL would spend millions of dollars tackling an issue as large as sports betting, when preventing betting on the sport would jeopardize their popularity among fans.

Other sports writers and fellow bloggers across the web have voiced their displeasure in the NFL's actions, stating that the league's actions are extremely hypocritical and the true motivations behind this fight are being concealed from the public.

Over the past few years there have been cases where players, coaches, and referees have wagered on sports, this means that their actions may have threatened the true outcome of the game.

However, those instances where unfair favoritism has been questioned in sports have been very rare occurrences and most of those instances were related to illegal sports betting operations rather than legal ones.

By the way, this thesis is not an attempt to debate whether sports betting is morally right or wrong. However, it highlights how important it is to restrict the betting market to the people who are closely connected with the reality of this sport.

2.5.2: GAMBLING ON SPORT IN THE USA

In the USA, the Professional and Amateur Sports Protection Act of 1992 makes it illegal to operate sport bets except for a few states; making a comparison, in many European nations bookmaking (the profession of accepting sports wagers) is regulated but not criminalized.

One of the main debates regards the betting market that turn around the colleges' tournaments, where some American states don't allow to gamble on such sport competitions among students.

The NCAA (National Collegiate Athletic Association) has threatened to ban all playoff games in Delaware if the state allows betting on college sports. New Jersey, which is also interested in this betting field, has been similarly threatened. Proponents of legalized sports betting generally regard it as a hobby for sports fans that increases their interest especially in sport events.

Opponents fear that, over and above the general ramifications of gambling, it threatens the integrity of amateur and professional sports.

The history of this *black market* includes numerous attempts by sports gamblers to fix matches, although proponents counter that legitimate bookmakers will invariably fight corruption just as fiercely as governing bodies and law enforcement do.

However, there are professional sports bettors, that make a good income betting sports, many of which use sports information services (some relevant information on which the bettor shouldn't be allowed to access, for a fair betting market).

In areas where sports betting is illegal, bettors usually make their sports wagers with illicit *bookmakers* (known colloquially as "bookies") and on the Internet, where there are thousands of online *bookmakers* who accept wagers on sporting events around the world.

2.5.3: THE DIFFERENT KINDS OF BET

This thesis focuses its interest on the simple predictions related to the difference and to the sum of the points, but in the betting market there are many different ways to gamble.

After this clarification, to give an exhaustive framework of the market in exam, I'm going to show the main bets (with the relative descriptions) about the NFL outcomes.

The betting market in USA allows these different ways of betting:

- *Proposition bet:* these are wagers made on a very specific outcome of a match; examples include guessing the number of goals each team scores in a match or betting whether a player will score in a game.

- *Parlays:* a parlay involves multiple bets (usually up to 12, but even more) and rewards successful bettors with a large payout. For example, a bettor could include four different wagers in a four-team parlay, whereby he is wagering that all four bets will win. If any of the four bets fails to cover, the bettor loses the parlay, but if all four bets win, the bettor receives a substantially higher payout than if he made the four wagers separately.

- *Teasers:* a teaser allows the bettor to combine his bets on two or more different games. The bettor can adjust the point spreads for the two games, but realizes a lower return on the bets in the event of a win.

CHAPTER THREE: EXPLORATIVE ANALYSIS

First of all, to introduce the reader to the statistical analysis, I need to present some graphics that describe the "raw" variables (or transformations of them) available for this study.

This kind of explorative study offers the possibility to discover in which variables is more convenient to orientate the statistical models.

This general presentation (descriptive statistics) wants to highlight the importance of the *home effect* (the site of the game) in the prediction of the American football outcomes.

3.1: DATASET STRUCTURE

3.1.1: GENERAL INFORMATION

The dataset used for the elaborations has 2136 rows that correspond to the 2136 matches played among the 32 teams during the period 2002-2009.

For each of the 8 seasons the dataset has 267 matches and every season has 21 weeks of games.

As I explained in the previous chapter, these are the 21 weeks of matches that I consider in the analysis for every year:

- 17 weeks standard (in the most of the cases with 16 matches, but sometimes less)
- 1 week Wild Card Round (4 matches)
- 1 week Divisional Round (2 matches)
- 1 week Conference Championships (2 matches)
- 1 week Super Bowl (1 match)

Observation: At a first moment I thought to consider the "standard" 17 weeks separated from the other 4 "special" weeks (playoffs) but after few analyses I understood that it is more convenient to consider all the 21 weeks together, because of the high incidence of the playoffs on the season.

3.1.2: TRANSFORMATION AND DESCRIPTION OF THE VARIABLES

As a consequence of what I said in the observation of the previous paragraph, the variable Week has been transformed into a numeric variable with a scale of number from 1 to 21, where 21 indicates the last mach of every season: the Super Bowl.

As I said before, the Week will be the reference unit of time, so I created a new variable, exploiting the numeric variable just created, called W_tot that is the cumulative sum of the weeks in the 8 years analysed; in this way it is possible to move easily the data windows in order to predict and to test the results of the football games.

The final dataset used for the analyses contains these following variables:

- Season: (numeric) indicates the year, from 2002 to 2009
- Week: (text) indicates the number of the week, from 1 to 17 and then text to describe the 4 different phases of the playoffs
- W_tot: (numeric) represents the cumulative sum of the Weeks, from 1 to 168
- Day: (text) indicates the day of the week
- Date: (text) describes the date
- Hpts: (numeric) indicates the points made from the team that played at home
- Apts: (numeric) indicates the points made from the team that played away
- Hyds: (numeric) indicates the yards gained from the team that played at home
- Hyds: (numeric) indicates the yards gained from the team that played away
- Hto: (numeric) indicates the turnover index for the team that played at home
- Ato: (numeric) indicates the turnover index for the team that played away
- **D**: (numeric) represents the difference between the score of the team that played at home and the score of the team that played away
- **S**: (numeric) represents the sum between the score of the team that played at home and the score of the team that played away
- y_diff: (numeric) represents the difference between the yards gained from the team that played at home and the yards gained from the team that played away
- y_sum: (numeric) represents the sum between the yards gained from the team that played at home and the yards gained from the team that played away
- dto: (numeric) represents the difference between the turnover index of the team that played away and the turnover index of the team that played at home (in this chapter I will explain why it's not the difference between the home and the away)
- sto: (numeric) represents the sum between the turnover index of the team that played away and the turnover index of the team that played at home

Note: I denoted with the **bold** the main response variables of the analyses.

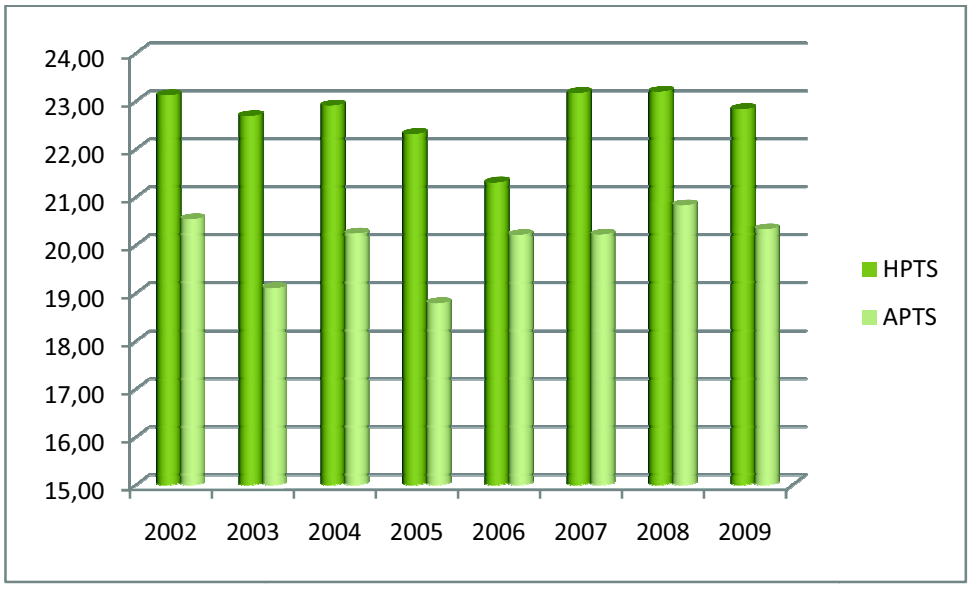
3.2: THE *HOME EFFECT*

3.2.1: THE POINTS

To forecast this kind of results, after browsing web pages about NFL matches and sites where it is possible to gamble, I choose to consider, as a valid response variable, the difference of the points between the home team and the away team, because in this sport, the home advantage looks approximately significant over the time.

This graphic below (where HPTS stands for the mean of the home points and APTS stands for the mean of the away points) confirms the systematic advantage of the teams that play in its stadium in the 8 years in analysis.

Graphic 3.1



As we expected, the site of the game tends to be determinant on how many points a team scores in a game; this is confirmed in all the 8 years in analysis.

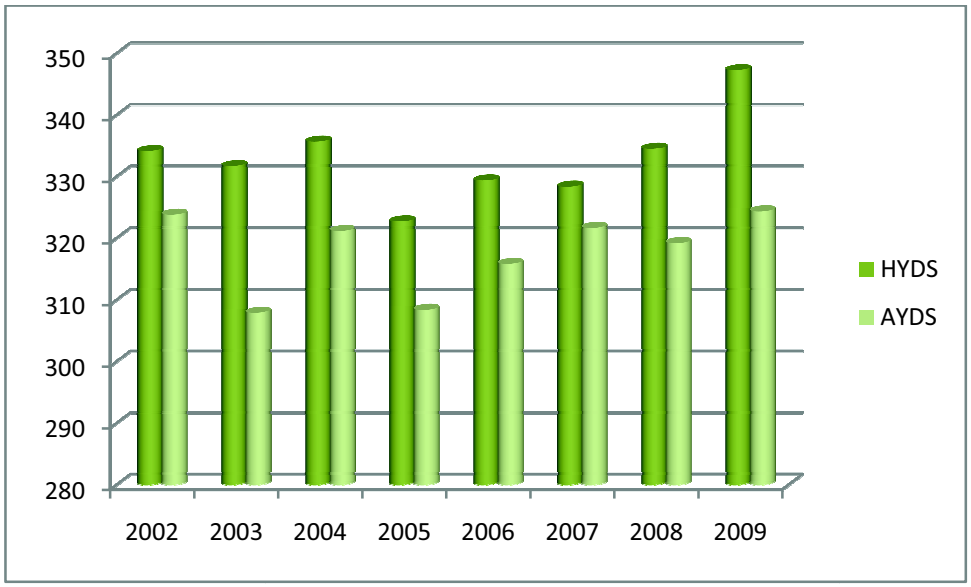
We can easily deduct that the *gap* between the home and the away points is significant, about 3 points on average (except for the 2006 where the advantage had been lighter, around 1 point).

3.2.2: THE YARDS

For the same reasons as before, but also for the high correlation (around the 65%) between the points and the yards (the more the team goes towards the opponent's end zone, the more probable it is to score points) I decide to consider also the difference of the yards between the home team and the away team, such a relevant variable to forecast the results of the NFL games.

This graphic below (where HYDS stands for the mean of the home yards and AYDS stands for the mean of the away yards) confirms, that also for this variable, the systematic advantage of the teams that play in their stadium in the 8 years in analysis.

Graphic 3.2



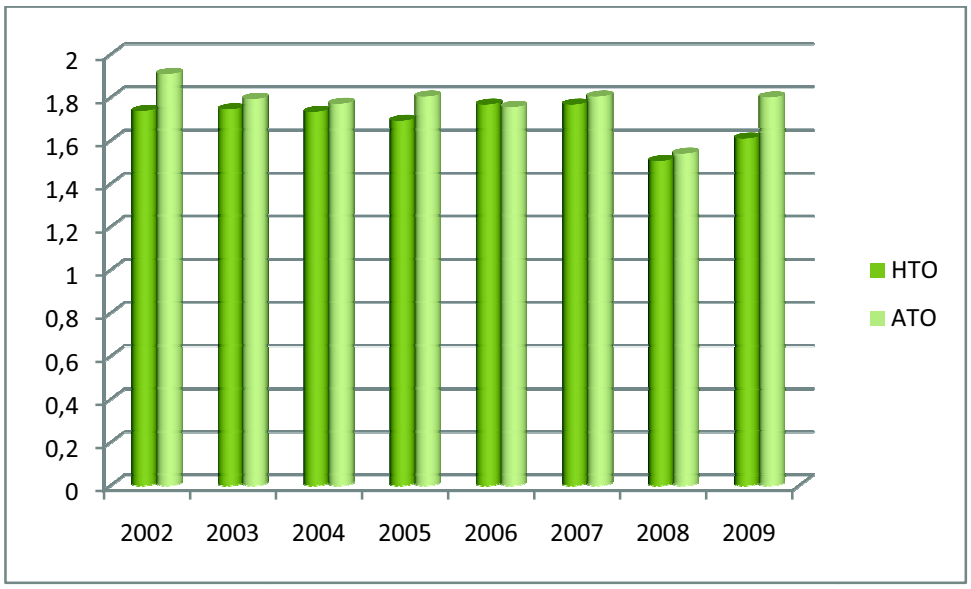
As we expected, the site of the game is determinant also on how many yards a team gains in a game, and the similarity with the previous graph confirms the correlation between points scored and yards gained.

We can easily deduct that the *gap* between the home and the away yards is significant, about 20 yards on average (except for the 2007 where the advantage had been lighter, less than 10 yards).

3.2.3: THE TURNOVER INDEXES

For the turnover indexes, the *home effect* is working in the opposite way. Actually, because of the nature of this descriptive statistic index, it is reasonable to attempt that a team when play away tends to lose more balls in offense (in other words, the turnover index tends to be higher) than when a team play in its own stadium.

Graphic 3.3



This graphic above shows that the teams globally tends to have an higher turnover index when they are playing away; this means that who plays at home tends to lose less balls because of the stadium effect (and consequently has a stronger defense).

I have to highlight that in the season 2006 the teams tended to lose more balls in offense when they played in their own stadium (practically, the *home effect* is not respected, just for this year)

That's why I said that the turnover index works in the opposite way, but actually it is correct that the *home effect* "behaves" like this for the particular construction and meaning of such index.

3.3: THE CANDIDATES AS RESPONSE VARIABLES

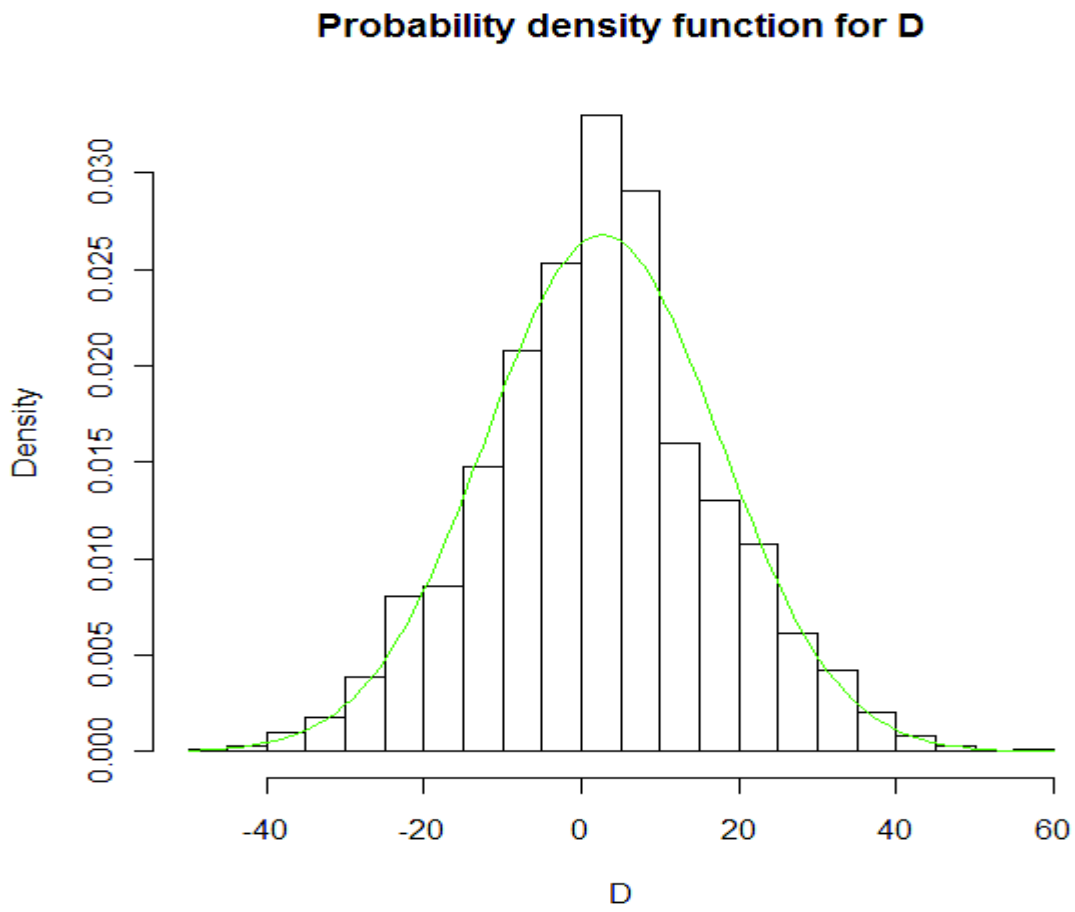
3.3.1: THE DIFFERENCE OF THE POINTS SCORED

In the statistic models used in this thesis, the 2 main response variables are the difference and respectively the sum, between the points scored by the home team and the points scored by the away team.

These variables indicate the 2 main borders where the betting market is oriented; in fact, all the bets turn around these bookmakers' predictions (they are actually the elements of the *Las Vegas line*'s vectors).

The graphic below shows the probability density function for D, that has a mean equals to 2,66.

Graphic 3.4

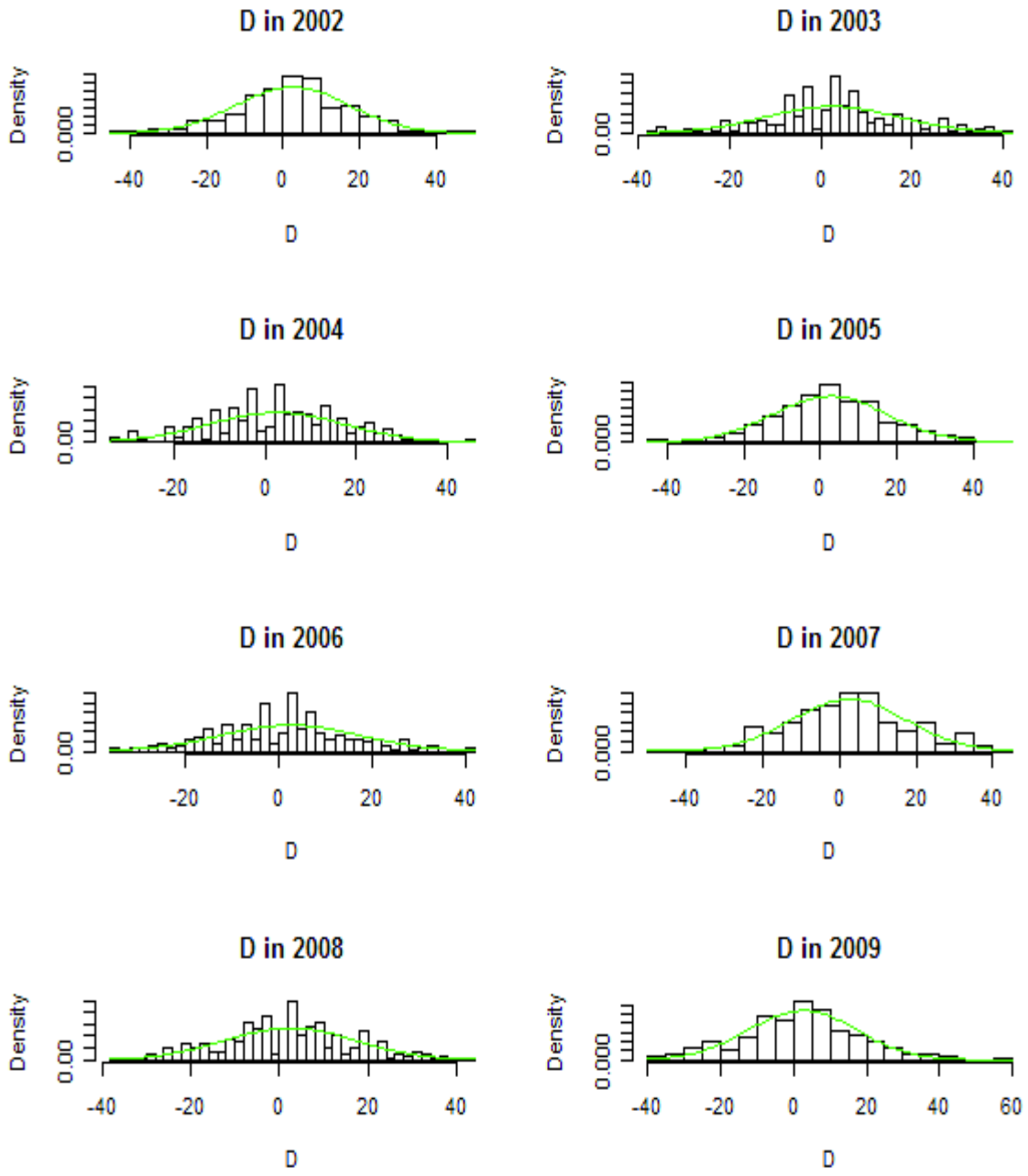


As we can see from the graphic above, this variable is approximately distributed as a $N(\mu,\boldsymbol{\sigma}^2)$ with μ that tends to be more than 0, confirming once again the clear *home effect.*

Note: In specific the green curve is the probability function associated to a N(E[D],VAR[D]).

The difference of points between the home and the away team is one of the 2 main variables, that's why I show also the distribution per year, in detail for the 8 years in analysis.

Graphic 3.5



The previous graphics show that the *home effect* is systematic and has almost the same trend for all the 8 years that I took into consideration in this thesis. Approximately all these distributions follow the shape of the Normal distribution described before.

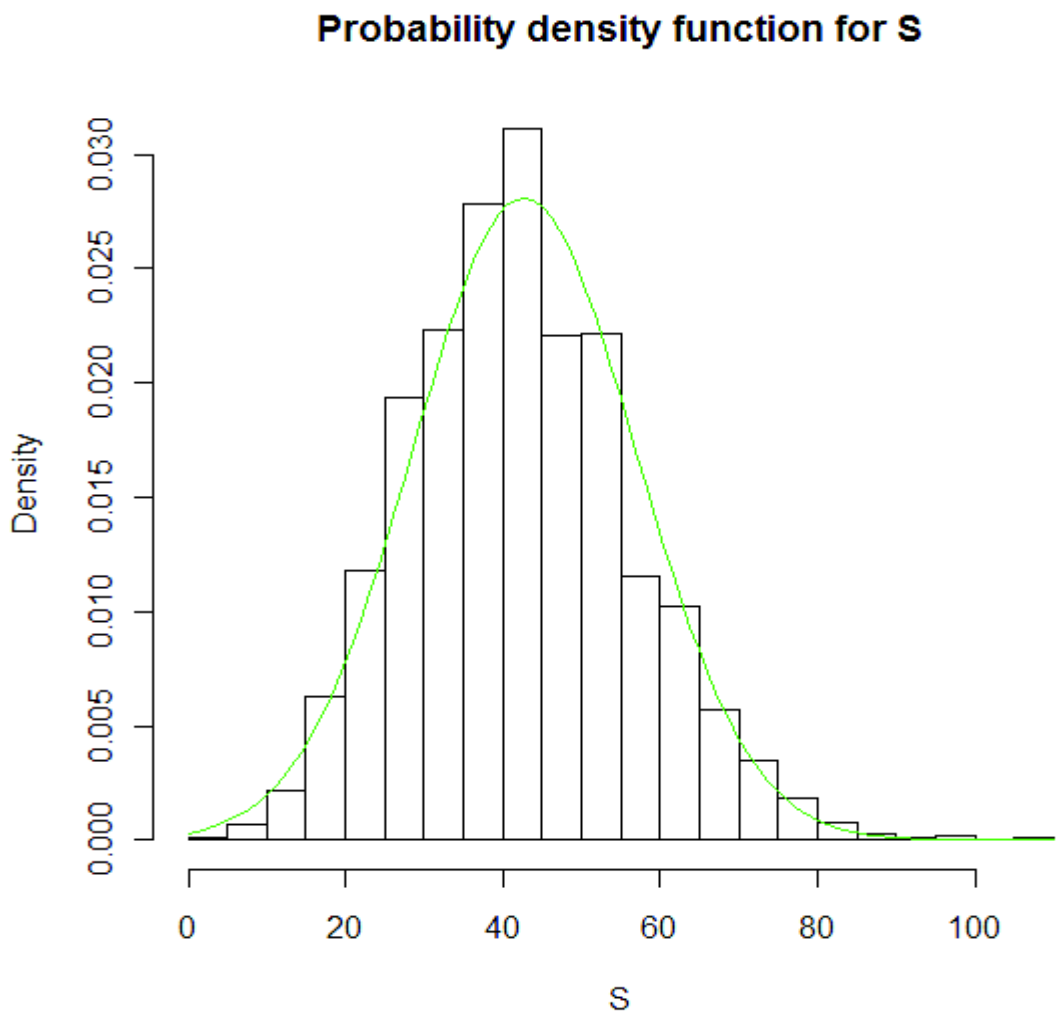
3.3.2: THE SUM OF THE POINTS SCORED

In this paragraph, I'm going to analyse the sum of the points of the teams to manage a typical bet that is really spread also in the traditional football: the under/over bet on the total of the points scored.

This special kind of bet fixes a border and the gambler must choose if the sum of the points of the 2 teams involved in the game will be higher or lower than the border fixed by the bookmakers.

The graphic below shows the distribution of S (sum) that has a mean equals to 42,72.

Graphic 3.6

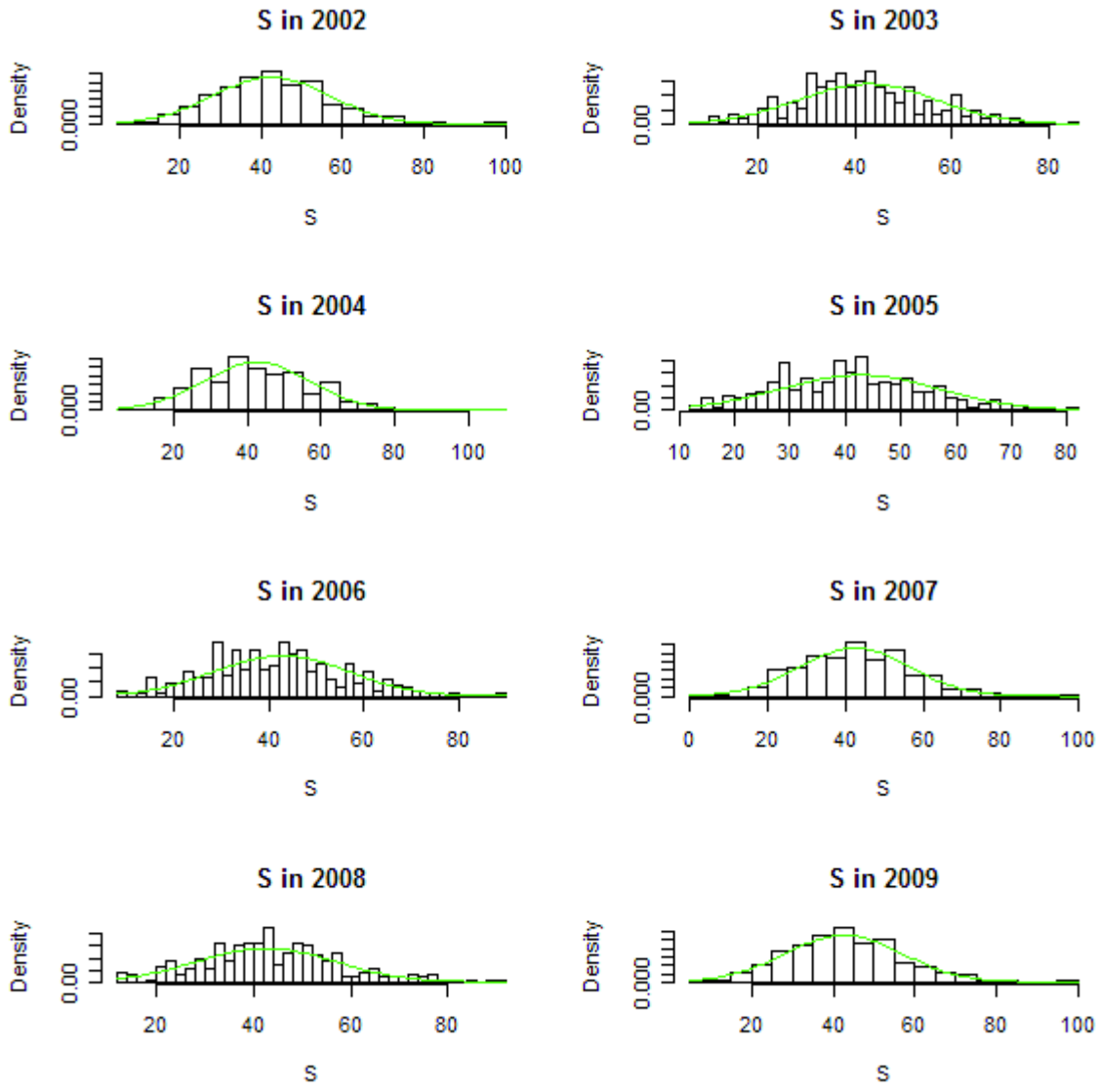


This distribution can be considered approximately Normal as the difference of the points (even if is less perfect) and I may expect the distributions per year with approximately the same characteristics.

Note: In specific the green curve is the probability function associated to a N(E[S],VAR[S]).

The graphics below show the distribution per year of the sum of the points between the home team and the away team.

Graphic 3.7



Even if the trend is not exactly the same over the years, in terms of frequency and in terms of shape of the distribution, we can assume that all these distributions are approximately Normal.

Note: anyway, the approximation is surely less perfect than before and it's possible to identify a particular distribution regarding the year 2005.

3.4: THE SUPPORT VARIABLES

3.4.1: THE DIFFERENCE OF THE YARDS GAINED

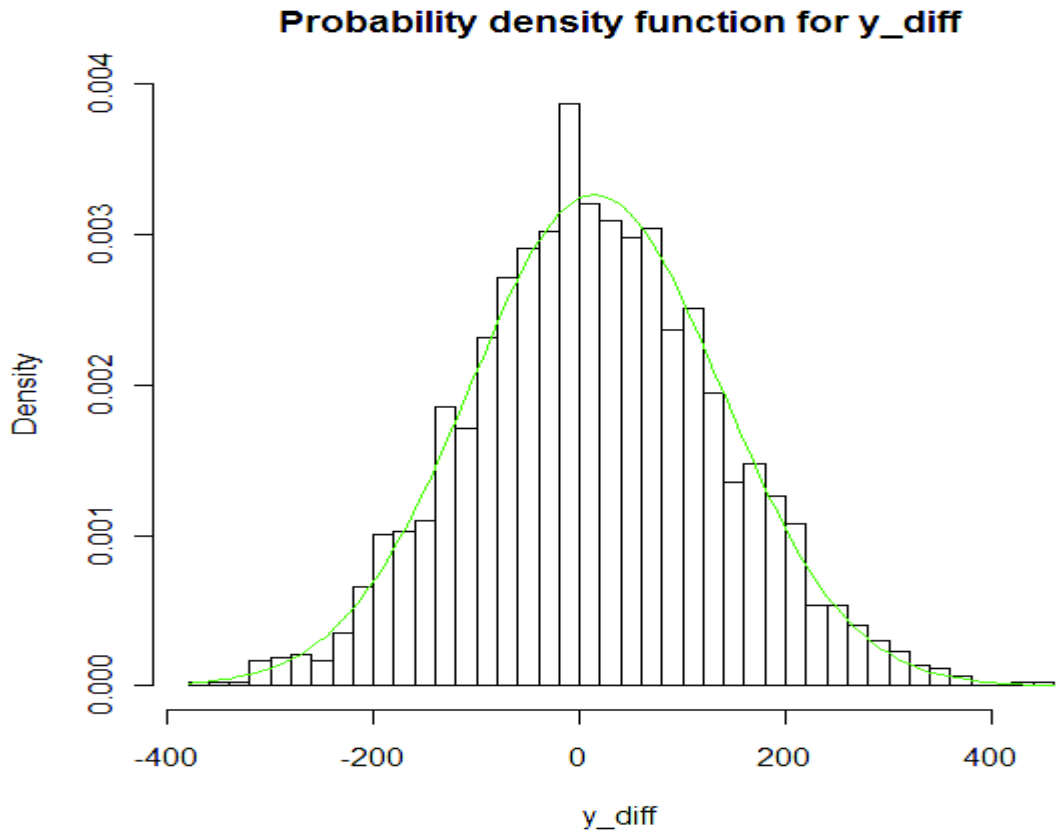
After the parallel analyses carried out in the subchapter 3.2 between the difference of points and the sum of the points, it is natural to set also the difference of the yards (and subsequently the sum) as valid variables to help the analyst to predict correctly the result of the game.

In the next analyses, it will be clear that there's no motivation to use such variables as response variables, because we are interested in predicting the difference of the points.

Note: In specific the green curve is the probability function associated to a N(E[y_diff],VAR[y_diff]).

The graphic below shows the probability function for this variable that has a mean equals to 15,17.

Graphic 3.8



As we can see, even in this case the analogism between the difference of the points and the difference of the yards is stated; this variable is approximately distributed as a $N(\mu,\sigma^2)$ with μ that tends to be more than 0 (even if the higher column of the graph is negative, the pyramidal shape tends to the positive side rather than the negative one), denoting once again the obvious *home effect* present in this kind of variables.

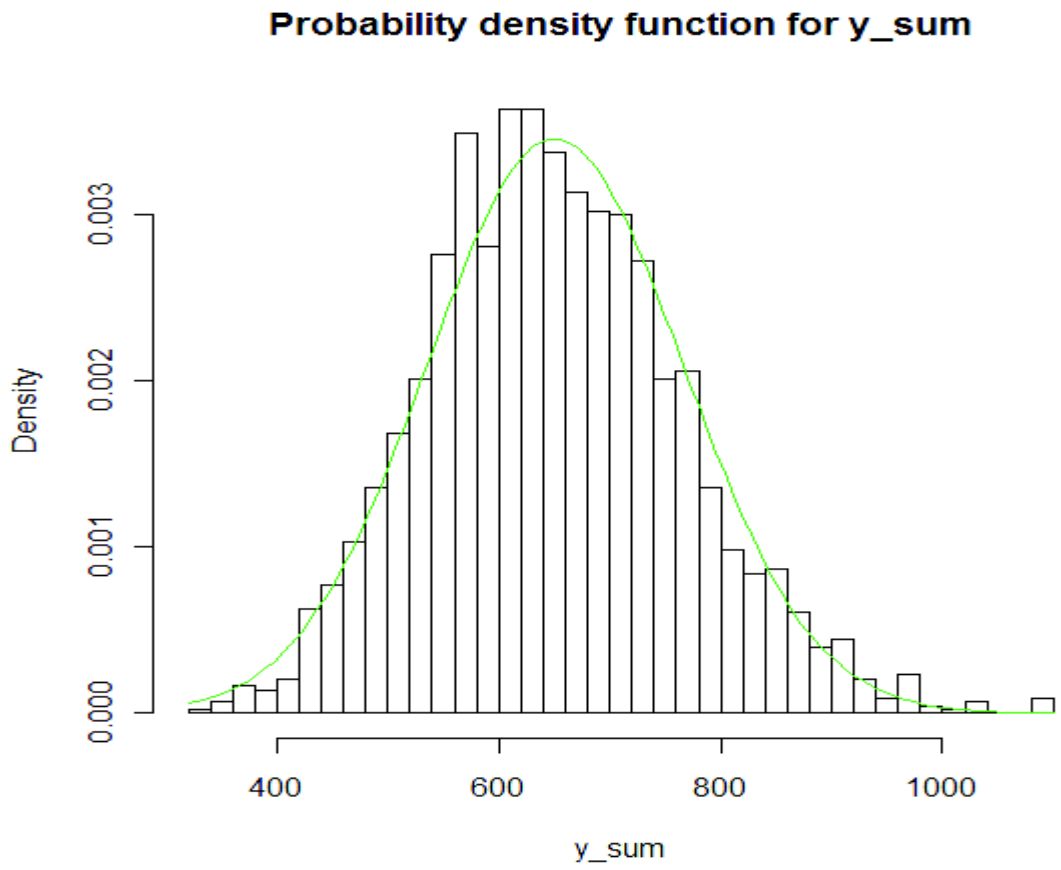
3.4.2: THE SUM OF THE YARDS GAINED

Probably, as a consequence of the high correlation between scored points and gained yards, it's possible that the sum of the yards can be useful to predict the sum of the points (for the under/over bets relative to the sum of the points).

Note: In specific the green curve is the probability function associated to a N(E[y_sum],VAR[y_sum]).

This following graphic shows the sum between the yards gained at home and away.

Graphic 3.9



As we can see, even in this case the analogism between the sum of the points and the sum of the yards is confirmed; also this variable is approximately distributed as a $N(\mu,\sigma^2)$ with μ equals to 650,74.

Nevertheless, It's important to underline that for values of y_sum > E[y_sum] the queue of the distribution tends to be longer.

Note: the yards and the turnover will be useful in the Chapter 5 with a model that exploits these 2 information to forecast the NFL outcomes.

3.4.3: THE DIFFERENCE OF THE TURNOVER INDEXES

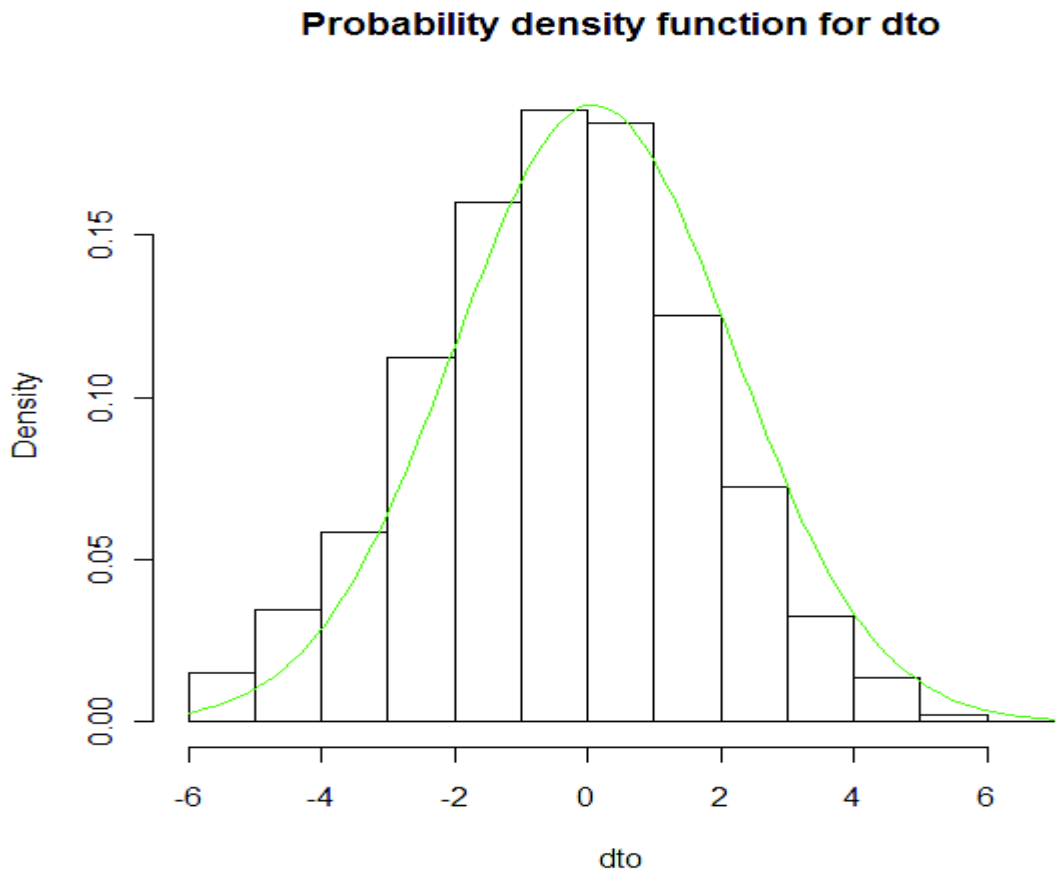
The turnover index is an important variable to fix the football outcomes, that's why I'm going to introduce also a graphic that shows the difference between the turnover index of the team that played away and the turnover index of the team that played at home.

Effectively, in this index the *home effect* works in the opposite way, because this specific variable describes how many times the offense loses a ball when it is attacking; it's natural to expect that when a team is playing away tends to lose more balls than who's playing at home.

Also for this variable (like for the yards), it will be clear that there's no reason to use directly this variable as a response variable, because we are interested in predicting the difference of the points (and the sum of the points).

Note: In specific the green curve is the probability function associated to a N(E[dto],VAR[dto]).

Graphic 3.10



Actually the *home effect,* it's not so marked as for the previous variables. Anyway I can consider also this distribution, approximately as a $N(\mu,\sigma^2)$.

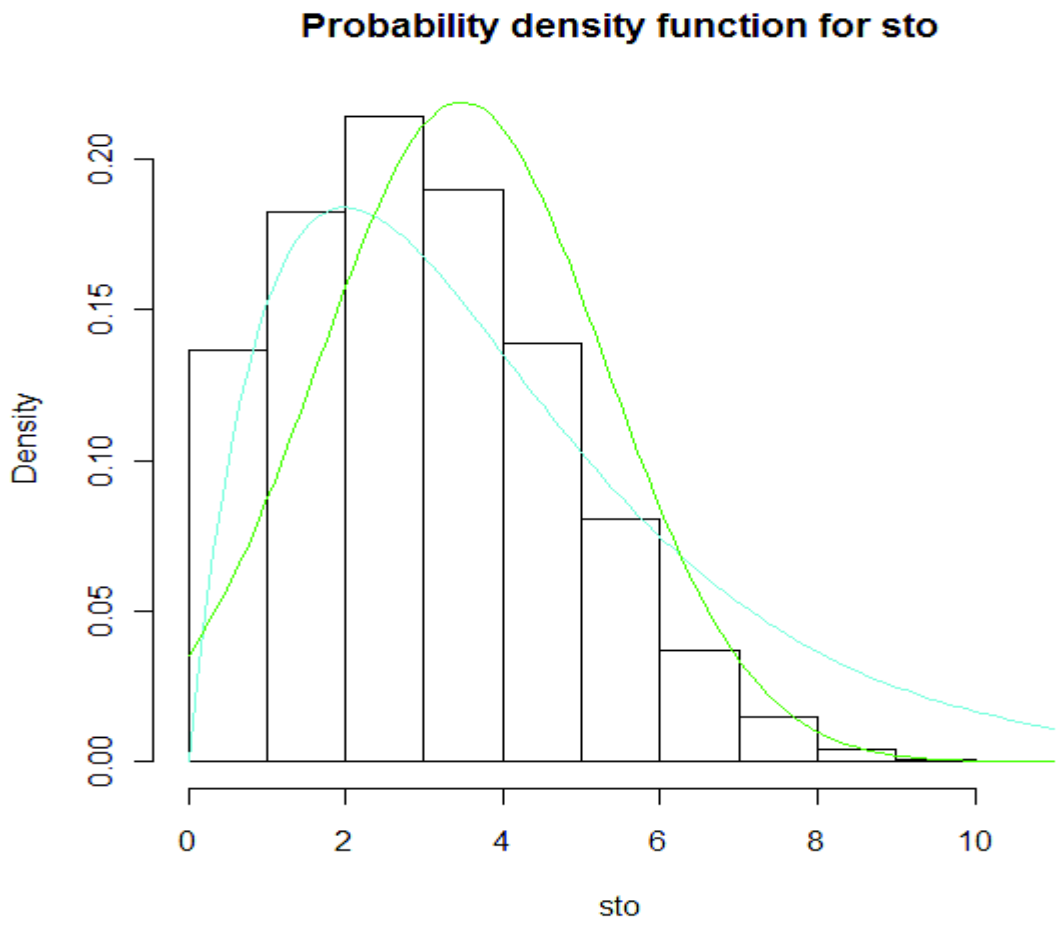
3.4.4: THE SUM OF THE TURNOVER INDEXES

For the same reasons discussed for the sum of the yards, even the sum of the turnover indexes could be interested to the main aim of this thesis (in order to fix the correct outcome of the NFL matches). Certainly it will be useful to predict the sum of the points as response variable.

Note: In specific the green curve is the probability function associated to a N(E[sto],VAR[sto]).

This following graphic shows the sum between the turnover index of the home team and the turnover index for the away team.

Graphic 3.11



This variable has a mean equals to 3,44 and it is the worst approximation to the Normal belle curve. To consider the model approximately Normal, it should be: E[sto] ≅ Me[sto]

In this case the median is equal to 3, and the *gap* is quite marked for the scale of this variable.

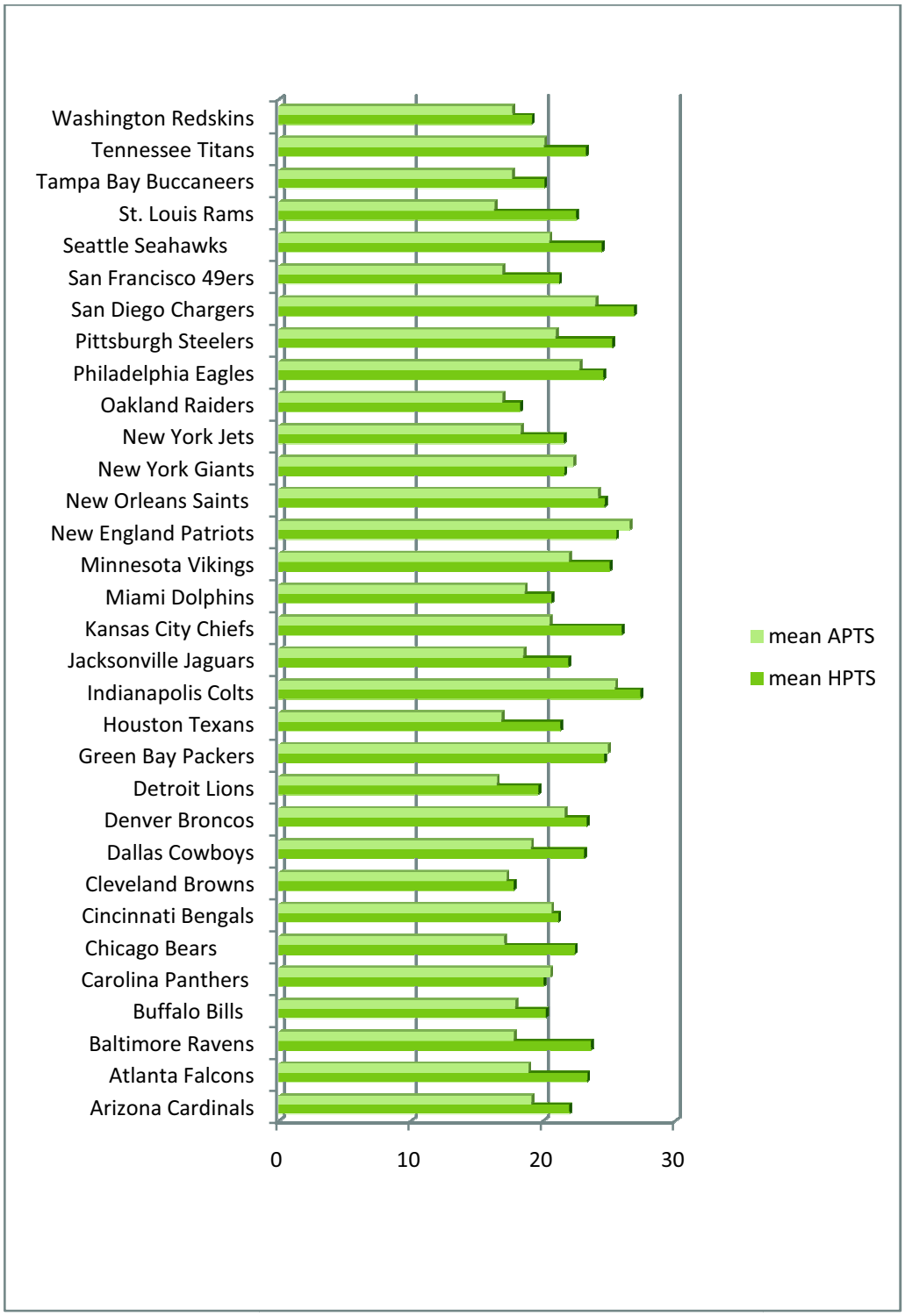
Actually the shape reminds a χ^2 distribution, specifically with 4 degrees of freedom, as I highlight in light blue in the graphic (but it can't be exploit in this analysis). Anyway, we can assume that this distribution is approximated Normal (even if is the worst approximation).

3.5: THE STRENGHTS OF THE TEAMS

3.5.1: GLOBAL OVERVIEW

At a first moment I wanted to show the distributions of the means of the points per year, but after this kind of diagnostics I denoted just a general increasing of the points scored over the time.

Graphic 3.12



The previous graphic principally shows 2 things: the differences that describe the global strength of each team and then it underlines the home effect that is included in almost all the teams.

To confirm this second observation, I study also the table with the data relative to the previous graph to show the relevance of the site of the game on the teams' results.

Table 3.1

TEAMS	HPTS	APTS	DIFF
Arizona Cardinals	21,95625	19,0924	2,8639
Atlanta Falcons	23,29514	18,8410	4,4542
Baltimore Ravens	23,60243	17,7651	5,8374
Buffalo Bills	20,17188	17,8594	2,3125
Carolina Panthers	20,01389	20,5028	**-0,4889**
Chicago Bears	22,34722	17,0330	5,3142
Cincinnati Bengals	21,09722	20,5781	0,5191
Cleveland Browns	17,73438	17,1858	0,5486
Dallas Cowboys	23,0625	19,0486	4,0139
Denver Broncos	23,25625	21,5851	1,6712
Detroit Lions	19,59375	16,4531	3,1406
Green Bay Packers	24,60347	24,8750	**-0,2715**
Houston Texans	21,26562	16,8125	4,4531
Indianapolis Colts	27,34126	25,4118	1,9295
Jacksonville Jaguars	21,89062	18,4719	3,4187
Kansas City Chiefs	25,92361	20,4566	5,4670
Miami Dolphins	20,58507	18,5625	2,0226
Minnesota Vikings	25,00347	21,9278	3,0757
New England Patriots	25,48403	26,5184	**-1,0344**
New Orleans Saints	24,62311	24,1129	0,5103
New York Giants	21,56597	22,2412	**-0,6752**
New York Jets	21,54167	18,2754	3,2663
Oakland Raiders	18,21875	16,8767	1,3420
Philadelphia Eagles	24,53958	22,7186	1,8210
Pittsburgh Steelers	25,18946	20,9070	4,2824
San Diego Chargers	26,85938	23,9167	2,9427
San Francisco 49ers	21,1684	16,8663	4,3021
Seattle Seahawks	24,4066	20,4444	3,9622
St. Louis Rams	22,47743	16,2813	6,1962
Tampa Bay Buccaneers	20,06076	17,6129	2,4479
Tennessee Titans	23,21354	20,0365	3,1771
Washington Redskins	19,09375	17,6208	1,4729

Note: I denoted in red the teams where the *home effect* seems absent; in fact, those 4 teams tend to score more points away. Furthermore, I didn't consider also the same distribution for the yards because the results are similar and bring to the same conclusions obtained with this table.

3.5.2: FEW EXAMPLES OF THE PERFORMANCES OF THE TEAMS OVER THE YEARS

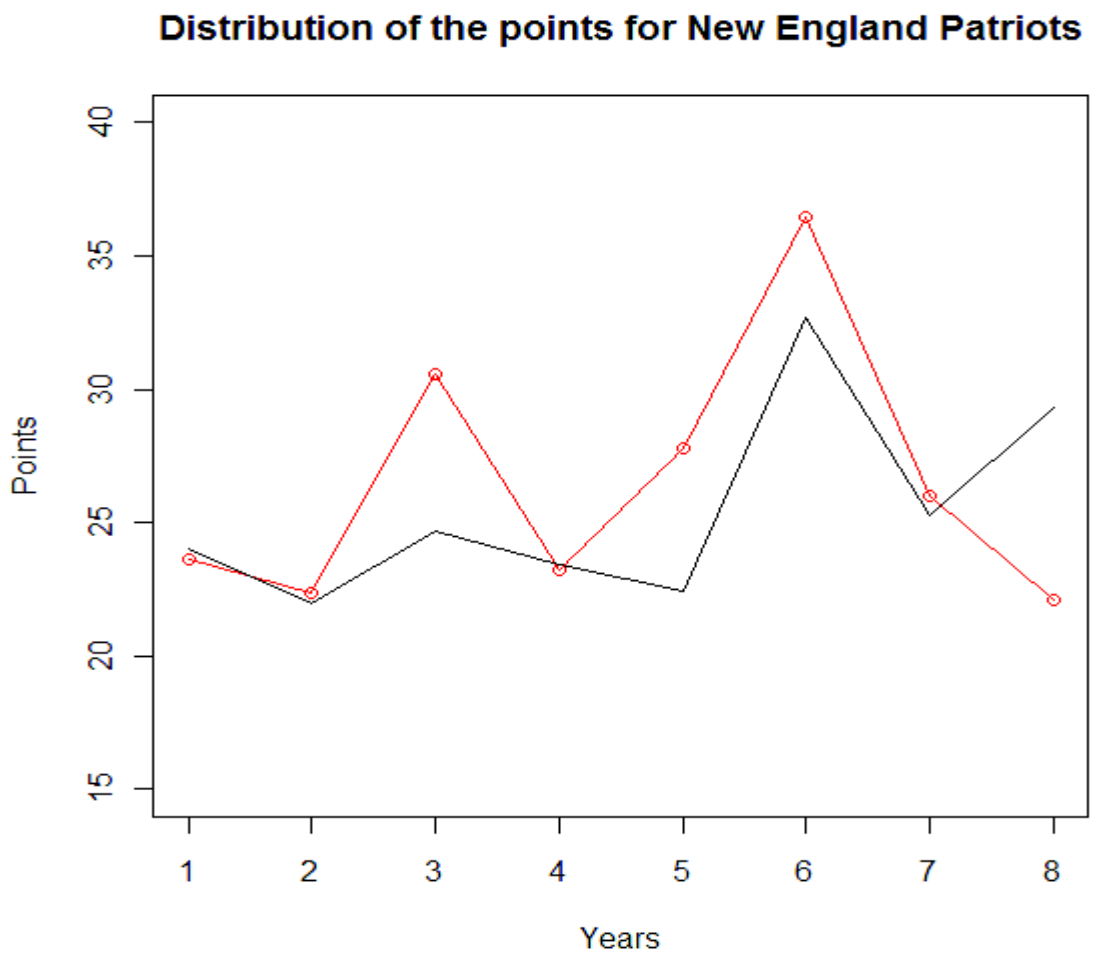
A bookmaking approach would present singularly all the performances of the teams involved in the championship and surely it would be possible to indentify some trends for some specific teams; certainly the bookmaker's models take care about such kind of information.

By the way, for the aims of my thesis, it is not necessary to show all these specific distributions, but I prefer to introduce few interesting examples linked to 3 of the best teams in these 8 years of analysis (see 2.3: HALL OF FAME):

a) NEW ENGLAND PATRIOTS

The New England Patriots is one of the best team of the AFC conference and it's a special case because it tends to score more points when it isn't playing in their own stadium.

Graphic 3.13

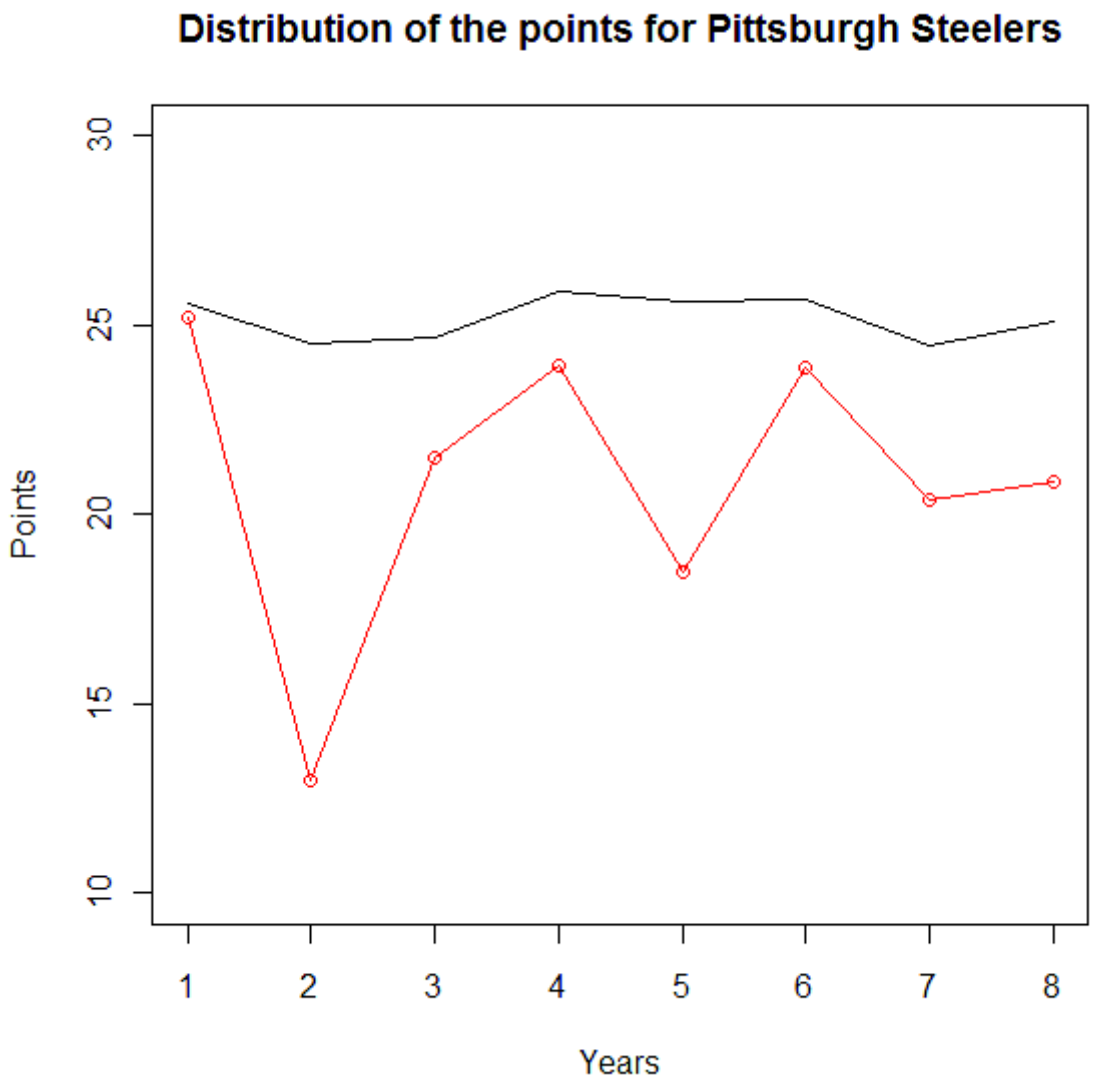


Note: In black, I denoted the line of the means of the points scored by the team when it is playing at home (per year); in red I denoted the line of the means of the points scored by the team when it is playing away (per year).

b) PITTSBURGH STEELERS

Another good team in these 8 years, from the AFC conference as well, is the Pittburgh Steelers; this team is peculiar because of its constancy when it is playing at home, in fact in the 8 years the mean of the points scored at home is always around 25.

Graphic 3.14

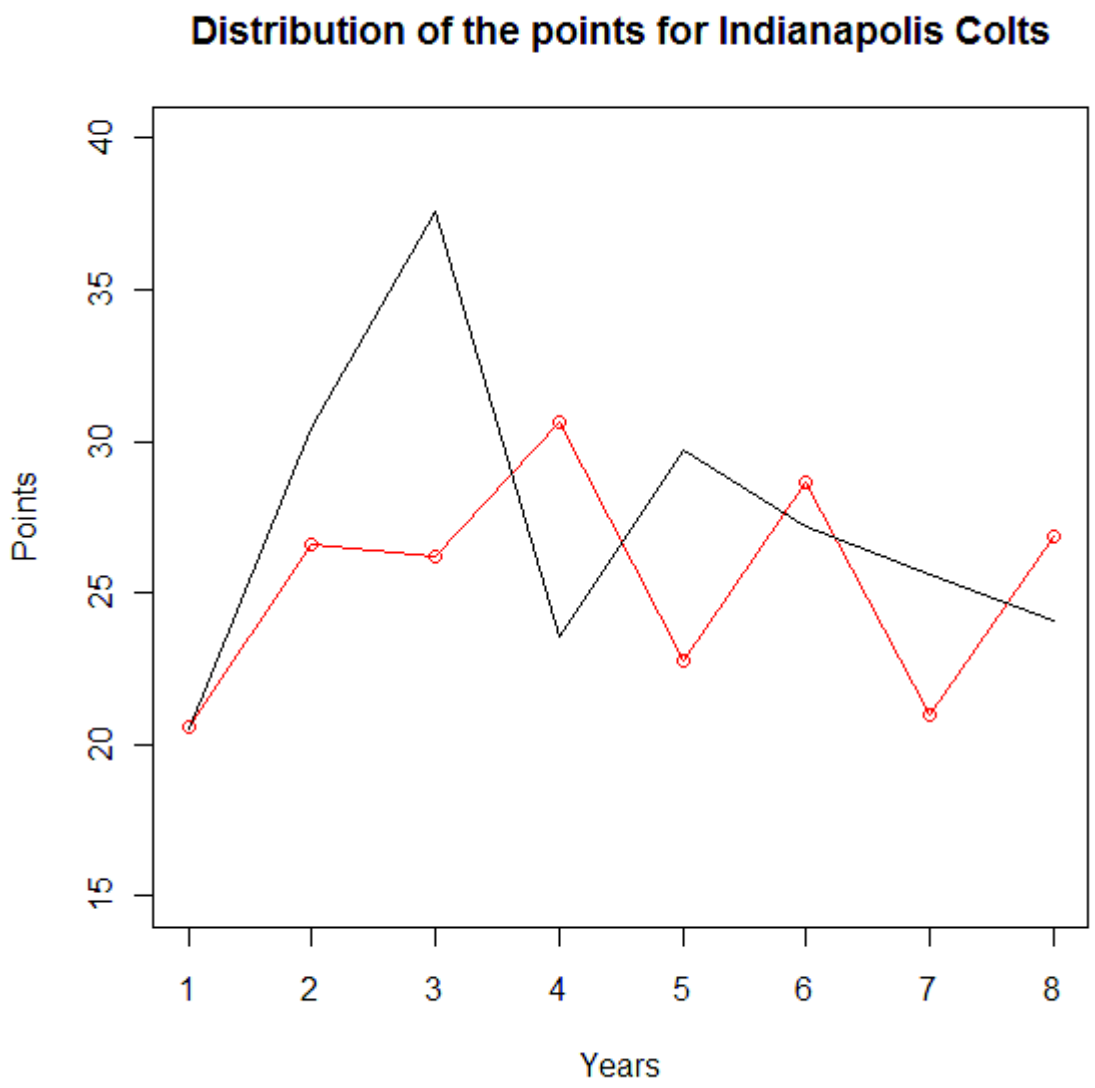


Note: In black, I denoted the line of the means of the points scored by the team when it is playing at home (per year); in red I denoted the line of the means of the points scored by the team when it is playing away (per year).

c) INDIANAPOLIS COLTS

The Indianapolis Colts, that arrived at the Super Bowl in 2006 (when it won) and in the 2009 (when it lost) has the following distribution; even in this case the home effect is not particularly marked (except for 2004), in fact just 4 times out of 8, the team had scored more points at home.

Graphic 3.15



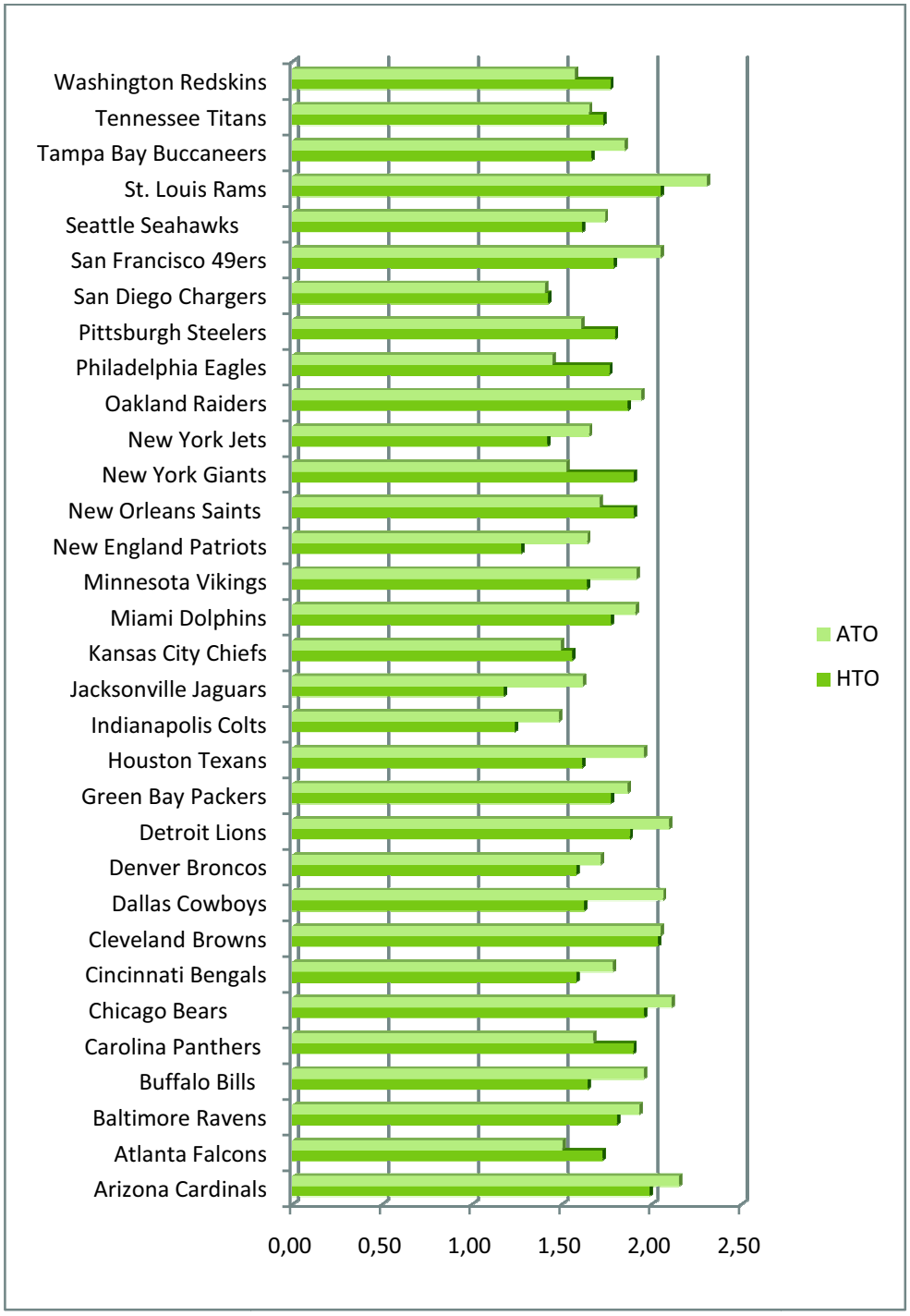
Note: In black, I denoted the line of the means of the points scored by the team when it is playing at home (per year); in red I denoted the line of the means of the points scored by the team when it is playing away (per year).

3.6: FOCUS ON THE TURNOVER

In the general overview of the NFL game I underlined the turnover index, that somehow represents the efficacy of the offense of each team (and at the same time the defense of each team).

This variable represents how many times the offense loses the ball when it is attacking:

Graphic 3.16



In the last graphic it's clear that the turnover index tends to be higher when the teams are playing away (like the conclusions for the subchapter 3.4). Theoretically, when you are playing away it is more probable that the defense of who is playing at home would tend to be stronger (for the *home effect*); as a consequence of this, the attack of the away team tends to lose more balls.

Table 3.2

TEAMS	HTO	ATO	DIFF
Arizona Cardinals	1,985075	2,149254	-0,164179
Atlanta Falcons	1,723077	1,500000	**0,223077**
Baltimore Ravens	1,803030	1,927536	-0,124506
Buffalo Bills	1,640625	1,953125	-0,312500
Carolina Panthers	1,893939	1,671429	**0,222510**
Chicago Bears	1,955224	2,107692	-0,152468
Cincinnati Bengals	1,575758	1,78125	-0,205492
Cleveland Browns	2,031250	2,046154	-0,014904
Dallas Cowboys	1,621212	2,059701	-0,438489
Denver Broncos	1,575758	1,712121	-0,136363
Detroit Lions	1,875000	2,09375	-0,218750
Green Bay Packers	1,768116	1,863636	-0,095520
Houston Texans	1,609375	1,953125	-0,343750
Indianapolis Colts	1,232877	1,478873	-0,245996
Jacksonville Jaguars	1,171875	1,61194	-0,440065
Kansas City Chiefs	1,553846	1,492308	**0,061538**
Miami Dolphins	1,769231	1,90625	-0,137019
Minnesota Vikings	1,636364	1,910448	-0,274084
New England Patriots	1,270270	1,637681	-0,367411
New Orleans Saints	1,897059	1,707692	**0,189367**
New York Giants	1,895522	1,521739	**0,373783**
New York Jets	1,415385	1,647887	-0,232502
Oakland Raiders	1,863636	1,938462	-0,074826
Philadelphia Eagles	1,760563	1,442857	**0,317706**
Pittsburgh Steelers	1,791667	1,602941	**0,188726**
San Diego Chargers	1,420290	1,402985	**0,017305**
San Francisco 49ers	1,784615	2,046154	-0,261539
Seattle Seahawks	1,608696	1,735294	-0,126598
St. Louis Rams	2,046154	2,30303	-0,256876
Tampa Bay Buccaneers	1,661765	1,846154	-0,184389
Tennessee Titans	1,727273	1,647059	**0,080214**
Washington Redskins	1,765625	1,567164	**0,198461**

Note: I denoted in red (bold) the positive differences between the turnover indexes, this evidence that those teams tend to lose more balls (in offense) when they are playing in their own stadium and this particularity is weird and in contrast with the *home effect*.

CHAPTER FOUR: STANDARD MODELLING TECHNIQUES

In this chapter, my mission is to show the main features, the assumptions and the results achieved with the *standard* models that I carried out to predict the football outcomes.

In these analyses, to predict my results, I exploited the linear model methodology throughout the least squares method (weighted and un-weighted) that, under the assumption of a linear *Gaussian* model, is the same as the maximum likelihood estimate methodology.

These results, in a first instance are compared to the real results of the games, and thanks to the *Mean Square Error (MSE)*, it's possible to evaluate the quality of the estimates.

In detail, my MSE is a measure that is provided by the comparison between the results of my predictions and the real results of the games involved in the analysis.

After that, in a second instance I calculated the bookmakers' MSEs (which measures the gap between the *Las Vegas line* and the real results of the matches) and thus I could compare my outcomes with the expert's results (via MSEs comparison).

Furthermore, I created a numeric tool that can help the analyst (statistician) to evaluate the efficiency of the model in terms of betting power: it's the *Correct Classification Index* (*CCI)*.

This index represents the percentage of the matches that my model classified correctly (under or over the bookmaker's line) throughout a specific comparison among my predictions, the *Las Vegas line* and the real results of the games.

Note: There are no motivations to check the correct classification using the 0 like a spread point, because the betting market is orientated around the Las Vegas line, and it is particularly focused on the under/over bets instead of the win/lose bets (diffused in traditional football).

I classify correctly when my and the real results (obviously of the same match) are either both under or both over the *Las Vegas line*. In other words, it means the percentage of how many times I would win the bet (for a specific year).

However, I want to underline, once again, that my main aim isn't to "beat" the MSEs of the bookmakers, but it's to reduce as much as possible (from my "rookie" point of view about NFL) the *gap* between my line and the *Las Vegas* one.

4.1: EXPLORATIVE MODELS

These following 2 models exploits the standard linear modeling methodology, using all the data available in this study (all the 8 years in analysis from 2002 to 2009, without adopting any kind of weighted approach).

4.1.1: EXPLORING THE DIFFERENCE OF THE POINTS

The first model aims to predict the difference of the points in function of some explicative variables available in the *dataset* and in function of the relative transformations of the best predictors (yards and turnover indexes) to have a general overview of the reality in exam.

The natural transformations of such important variables, used to predict the difference of the points, are the difference of the yards (between the home and the away team) and the difference of the turnover indexes (between the away and the home team respectively).

Under the assumption that:

$$D_i \sim N(\mu_i, \sigma^2)$$

The multiple regression that I thought for this first explorative step is:

$$\mu_i = \beta_0 + \beta_1(day)_i + \beta_2(week)_i + \beta_3(hteam)_i + \beta_4(ateam)_i + \beta_6(dto)_i + \beta_7(y\_diff)_i$$

I want to remind, once again, that under the assumption of Normality (shown in the Chapter 3 for the response variable: D), the estimation throughout the use of the least square is the same as the maximum likelihood estimation.

The model just described provides good values, in terms of observed α, for the estimates relative to the difference of the turnover indexes (dto) and the difference of the yards (y_diff), consecrating that these 2 variables are closely connected to the NFL outcomes (in detail the difference of the points).

These results are really reassuring, but this model cannot be used for the aim of this thesis, considering that obviously the yards and the turnover indexes are available just at the end of the games (in fact these 2 variables are specific statistics about a match has already been played)

Furthermore this model ratifies the Week as the most important reference unit of time (some weeks result significant, in terms of observed α, on the contrary the variable day cannot help the statistician for the analyses because it looks a “non-sense” variable).

In addition, for the analises it is more convenient to consider every week as an unique block, with the relative number of matches. From the *Informatics* point of view, the transformation of such variable in W_tot simplifies the operations of selection in the 168 weeks available.

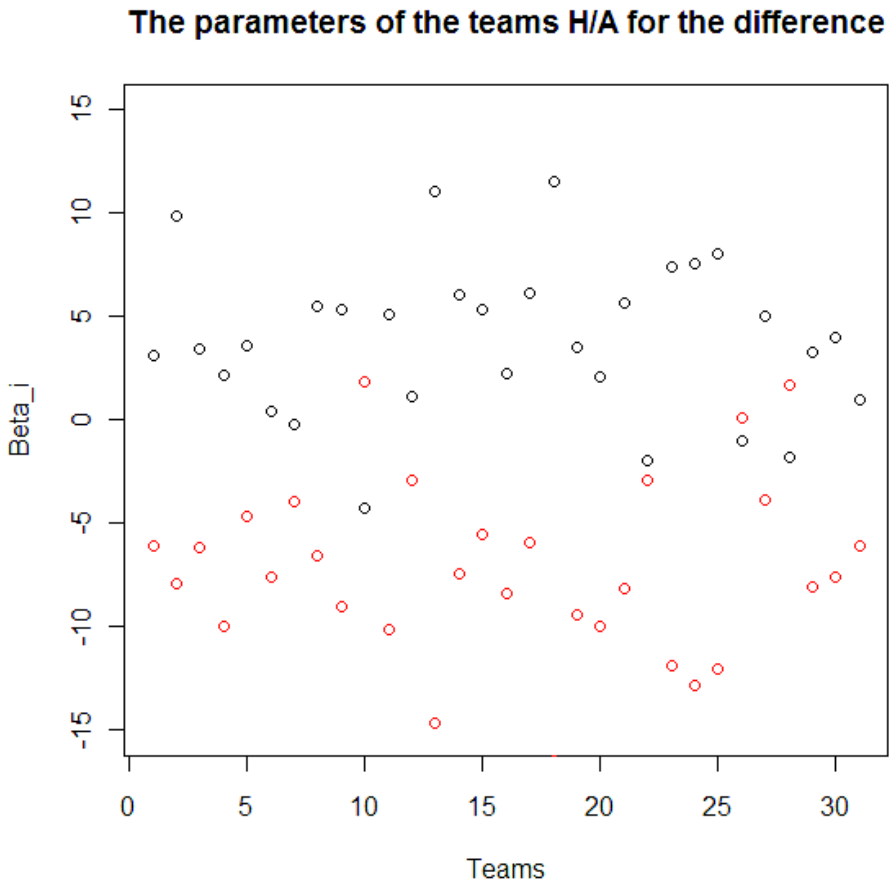
After the presentation of the points' mean at home and away (in the previous chapter) it looks interesting to reduce the previous model with just the parameters of the teams at home and away (because it is not relevant to consider in the model the yards and the turnover in this way, given that we don't have this information before the game).

The regression used to estimate the follow parameters reduce the main parameter μ to:

$$\mu_i = \beta_0 + \beta_1(hteam)_i + \beta_2(ateam)_i$$

That's why, I exploit this linear modeling approach to present the estimates of the parameters related to the 32 teams (obviously, the parameters available are just for 31 teams in order to avoid an over-parametrisation in the model: Arizona Cardinals has been omitted).

Graphic 4.1



The black points denote the parameters when the team i (in alphabetical order from Atlanta Falcons that correspond to i = 1) is playing at home, on the contrary the red points show the parameters when the team i is playing away.

4.1.2: EXPLORING THE SUM OF THE POINTS

This second model aims to predict the sum of the points in function of some explicative variables available in the *dataset*.

To have a general overview of the reality in exam, let's take into account the relative transformations of the best predictors (yards and turnover).

The natural transformations of such important variables, used to predict the sum of the points, are the sum of the yards (between the home and the away team) and the sum of the turnover indexes (between the away and the home team respectively).

Under the assumption that:

$$S_i \sim N(\mu_i, \sigma^2)$$

The multiple regression that I thought for this second explorative step is:

$$\mu_i = \beta_0 + \beta_1(day)_i + \beta_2(week)_i + \beta_3(hteam)_i + \beta_4(ateam)_i + \beta_6(sto)_i + \beta_7(y\_sum)_i$$

The model just described provides good values for the observed α (as we expected), for the estimates related to the sum of the turnover indexes (sto) and to the sum of the yards (y_sum). This fact affirms that these 2 variables (the sum and the difference) are closely connected with the NFL outcomes; thus it is useful to find the right way to contrast the bookmakers' line.

Once again, the Week is a significant variable and it is considered as the main unit of the time in the analyses.

Substantially the conclusions are the same ones as the previous model, except for the *home effect* considerations: for the sum of the points the home effect results less emphasized (due to the stadium), but it's still possible to recognize a higher pattern associated with the sum of the points scored at home.

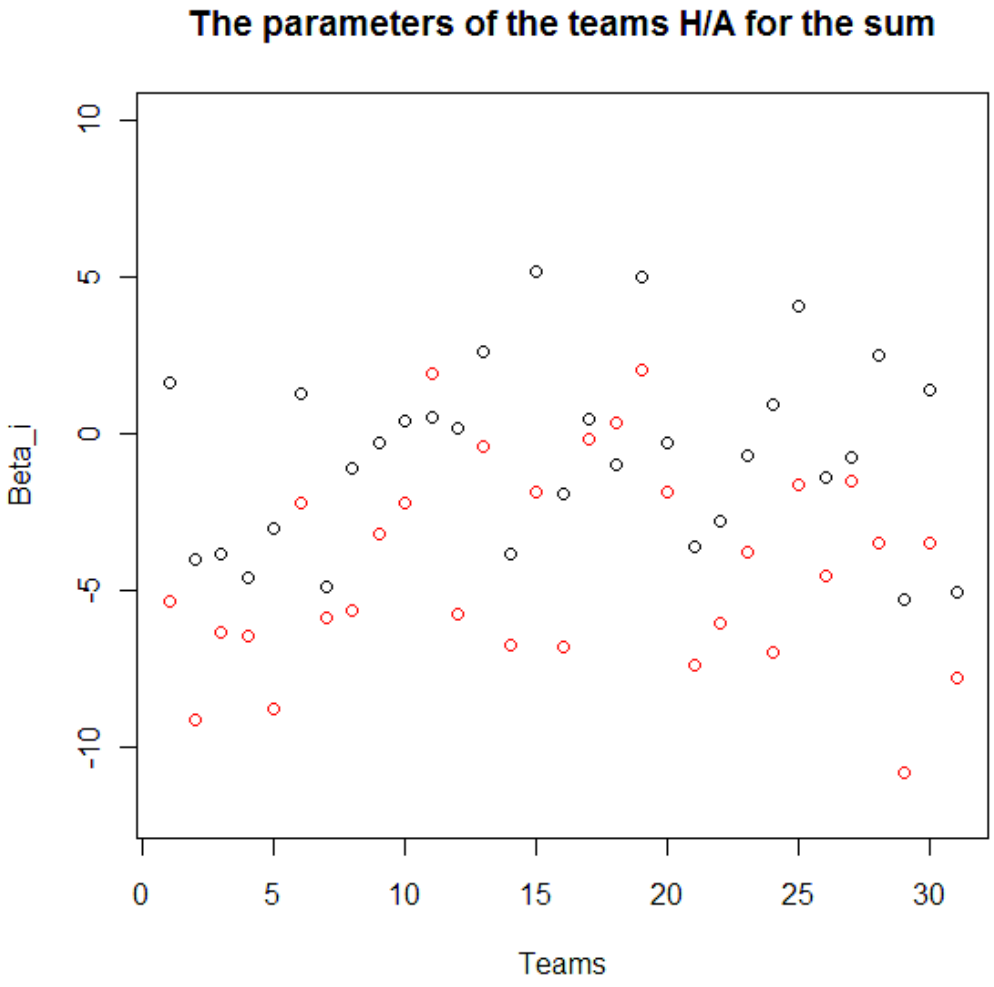
To highlight this previous consideration, also in this case I take a look to the parameters of the team used to predict the sum of the points.

In this specific case, the regression used to estimate the follow parameters reduce the main parameter μ to:

$$\mu_i = \beta_0 + \beta_1(hteam)_i + \beta_2(ateam)_i$$

I exploit this linear modeling approach to show the estimates of the parameters related to the 32 teams (obviously the parameter available are just for 31 teams to avoid an over-parametrisation in the model: Arizona Cardinals is omitted).

Graphic 4.2



Note: As we expected the general parameters for the teams (home and away) to determine the sum of the points respect the *home effect*.

I must underline that the *gap* between the parameter of the team i at home and the parameter of team i away is less marked than the parameters linked to the prediction of the difference of the points.

4.2: THE DESIGN MATRIX OF THE EXPLICATIVES: X

It's necessary to find another way to manage the teams that are playing at home and away since I've seen that this factor is absolutely crucial to the aims of this thesis.

Substantially, I need another approach that involves at least half of the parameters involved in the previous statistic models (63 parameters) and enable to predict correctly the NFL outcomes.

4.2.1: DETAILS OF CONSTRUCTION

To manage the problem of the *home effect* I followed the approach of the article written by Mark E. Glickman and Hal S. Stern: "A state space model for National Football League scores" published in March 1998 in the Journal of the American Statistical Association; this methodology follows the rules described from the following 2 notes.

Note: Difference of the points

This matrix has 2136 rows as the number of the matches involved in the analysis; each row must contains a 1 and a -1, and the matrix is populated following this simple rule for each row: fill 1 if the team of that respective column plays at home, vice versa fill -1 if the team of the respective column plays away, finally, fill 0 in all the rest columns.

Note: Sum of the points

This matrix has 2136 rows as the number of the matches involved in the analysis; each row must contains two 1s, and the matrix is populated following this simple rule for each row: fill 1 if the team of that respective column plays at home, and likely fill 1 if the team of the respective column plays away, finally, fill 0 in all the rest columns.

The 2 matrixes just created are suitable to front the modeling approach that I have just described; in other words we have the numeric tools to calculate the parameters μ_i for every games involved in the 8 years of the *time series* (i varies from 1 to 2136).

These models are certainly over-parametrized, that's why, to implement it I had to take out a column (like in the previous models), specifically, I deleted the first one; this modification means that the parameters that I estimate are in function of the team: Arizona Cardinals (omitted).

Now the new matrix with 31 columns represents the X matrix of a model correctly specified, where I obviously add the vector of 1s associated to the intercept (that I want to include in the model, because it will be the general parameter that describes the *home effect*: β_0).

Thus, the final matrix used to calculate the predictions with the least squares method (weighted and un-weighted) has 32 columns and obviously 2136 rows (one for each game).

These numeric objects allow to calculate the predictions for the difference (and for the sum) of the points between the team that plays at home and the team that plays away, with the well-known formula of the least squares.

4.2.2: THE SPECIFICATION OF THE MODELS

The model now has one parameter for every team (except for the first team – Arizona Cardinals that I omitted to avoid an over-parametrisation).

Finally the models has 31 parameters associated to the 31 teams (from Atlanta Falcons to Washington Redskins, in alphabetical order), plus the parameter related to the intercept.

The model that I assumed for the 2 response variables, as seen before, is $N(\mu, \boldsymbol{\sigma}^2)$, thus,

For the difference of the points (between the home team and the away team):

$$D_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

For the sum of the points (between the home team and the away team):

$$S_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

$\boldsymbol{\sigma}^2$, the variance, for both cases, is assumed constant over the time (homoscedasticity assumption).

4.3: *MOVING WINDOW*

4.3.1: THE WINDOW

As I said in the explorative analysis the variable that permits to the analyst to move comfortably in the 8 years in analysis is W_tot: the cumulative sum of the weeks.

The moving window permits to the analyst to compare predictions associated with some specific weeks, based on a different amplitude of the data window used to predict.

Usually, the bookmaker and the bettor are going to attend that more recent matches give more information (on a specific week to predict) than a match played several weeks before; that's true, but we need to evaluate the best amplitude of this window that contains the data used to predict a certain specific week.

4.3.2: THE SPREAD POINT FOR THE DIFFERENCE

The first idea that I had about such kind of time series like the NFL matches over the time, is that the more recent games should be more relevant to predict the specific outcomes for a certain week.

The first approach consists in the amplitude modulation of a window to understand in which grade the amplitude of the window can influence the effectiveness of my results.

The model assumed is:

$$D_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where the regression for the parameter μ is:

$$\mu_i = \beta_0 + \beta_1(hteam)_i - \beta_2(ateam)_i$$

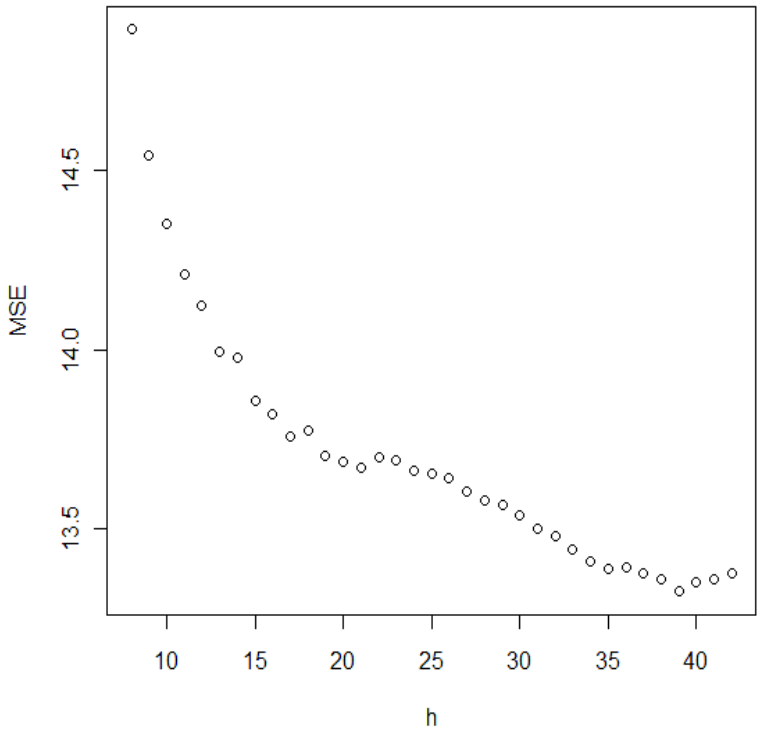
As I demonstrate in Chapter 3, it is correct to assume approximately *Gaussian* this model because of the normality of the response variable (difference of the points).

Furthermore, the variance ($\boldsymbol{\sigma}^2$) among the matches is assumed constant over time; what I'm modeling in detail is the parameter μ that explains the strength of the teams at home and away.

Note: I start to predict from the first week of season 2004, because I decide to move the window from 2 years (before the prediction) to 8 matches (actually 2 months before the week that I want to forecast); in this way I can easily evaluate which window size brings to the best tradeoff (between the MSE and the amplitude of the window) to forecast the results.

The graphic above shows the MSEs regarding the 128 weeks predicted (the weeks are 168 but the first 2 years aren't included) for every choice of the amplitude of the window h.

Graphic 4.1



The graphic is partially in contrast with the introduction of this subchapter; actually it shows that longer period in the data for estimating brings the analyst to predict better the real outcomes.

As a consequence of the previous graph, I choose the h equals to 39 (almost the maximum amplitude analysed) to calculate the relative MSEs per year.

Note: Actually, it is possible to deduce that every value of h more than 35 is a correct choice; in fact the function does appear to have *bottomed-out* for those particular values of h.

Anyway, there are no particular differences, in terms of MSE (it is the numeric tool used to establish the accuracy of the model), when the analyst uses a window width larger than 30 weeks.

As I said before, I show in the following table the MSEs concerning each year where is available the comparison with the *Las Vegas line*.

Note: In detail, in the following table (and in all the others tables with the same structure):

- The 1st column represents the seasons
- The 2nd column represents the MSE that I obtained comparing my model's results with the real results
- The 3rd column represents the MSE that I obtained comparing the results of the bookmakers' models with the real results
- The 4th column represents which is the *gap* between my predictions and the *Las Vegas line.*
- The 5th column represents the correct classification index of this specific model per year.

Table 4.1

Year	ME	BOOK	GAP	CCI
2006	13,84896	13,48817	0,36	46,81%
2007	14,10719	13,35387	0,75	49,81%
2008	14,31934	13,85654	0,46	52,43%

These results can be considered as a good beginning point for this analysis with an average *gap* that is around the 0,52 (compared with the bookmakers' predictions), in terms of MSE; actually, the *gaps* are not too marked even if the model is quite simple but particularly focused on the *home effect,* the main factor of this kind of predictions.

To evaluate the quality of the MSEs (obtained with the un-weighted model), it's natural to take a look to the main index of variability relative to D: the variance, it results equal to 221,95.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare the 2 indexes in the same scale:

$$SD(D) = \sqrt{VAR(D)} = 14,89$$

The gap between my average MSE (14,09) and the standard deviation is 0,80; my index is significantly lower than the SD(D) and it confirms that the model works properly.

4.3.3: THE SPREAD POINT FOR THE SUM

The *Las Vegas line* provides results even related to the sum of the points of every single match; these predictions are really important in this field to be able to manage the under/over bets about the sum of the points, really spread in the gambling market of the NFL sport.

That's why it is natural and useful, try also a model that involves the sum of the scores between the 2 teams and subsequently compare the results, as before, with the *Las Vegas line*.

The model assumed is:

$$S_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where:

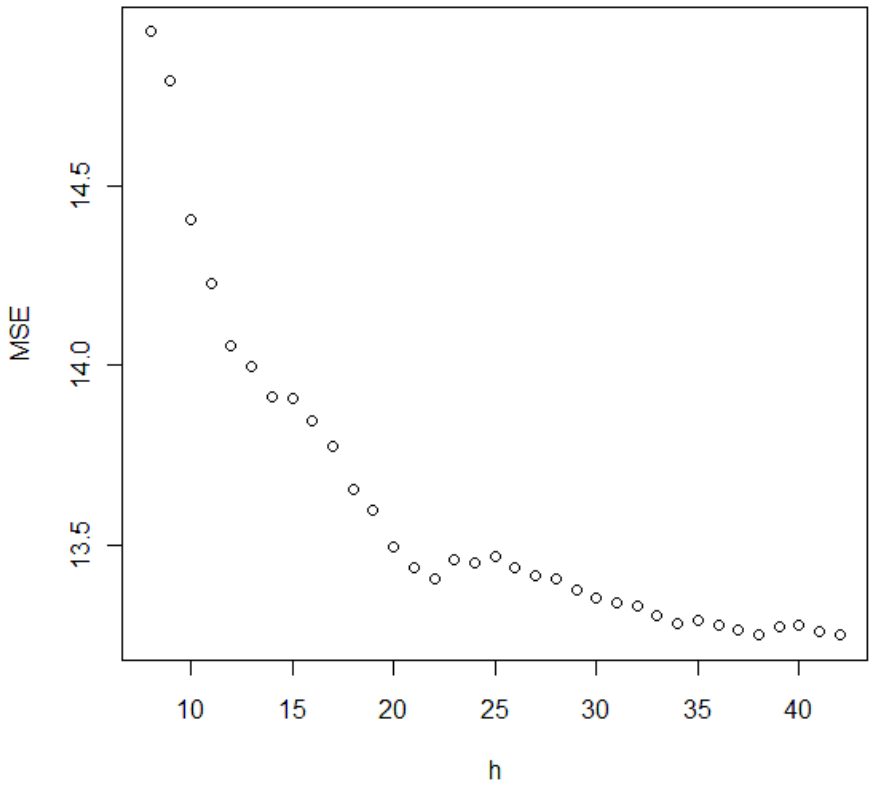
$$\mu_i = \beta_0 + \beta_1(hteam)_i + \beta_2(ateam)_i$$

As I demonstrate in the chapter 3 is correct to assume *Gaussian* this model because of the normality of the response variable (sum of the points).

Furthermore, the variance among the matches is assumed, as for the previous model, constant over time (anyway, the homoscedasticity is an assumption valid for all the analyses that I carried out in this thesis).

The graphic above shows the MSEs of the 128 weeks predicted (the weeks are 168 but the first 2 years aren't included) for every choice of the amplitude of the window h.

Graphic 4.2



The conclusions we drawn from this graphic are the same as the previous ones. In the analysis of the sum of the scores of the teams, larger windows provide inferior MSEs.

The graphic shows once again that is more convenient (in terms of precision of the estimates) to use a bandwidth that must be larger than 30 weeks (that actually it is almost one season and half of data before the week that is going to be predicted).

The fifth column of the following table represents the index of correct classification of this specific model to define the betting point of view of the analysis.

Table 4.2

Year	ME	BOOK	GAP	CCI
2006	14,24129	13,82403	0,41726	49,81%
2007	14,87784	14,13091	0,74693	46,81%
2008	13,80730	13,46114	0,34616	49,81%

This table doesn't denote a good fit of the model, even if the average gap is lower than before (around 0,50); this is confirmed from the CCI indexes, in fact none of them overcome the 50% of correct classification.

By the way, in this case the 2007 analysis brings to the worst results; anyway also the bookmakers' predictions are less reliable in this year, probably due to some unpredictable factors.

To evaluate the quality of the MSEs (obtained with the un-weighted models), it's natural to take a look to the main index of variability for S: the variance, it results equal to 201,90.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare the 2 indexes in the same scale:

$$SD(S) = \sqrt{VAR(S)} = 14{,}21$$

The gap between my average MSE (14,31) and the standard deviation is -0.1; this bad signal confirms the marked gaps and the low CCIs (in the 3 years) found using this un-weighted model to predict the sum of the points.

The results reached with the *moving window* approach establish that to have a reliable tool to predict such outcomes I have to consider a data window (on which the model is based for the prediction) bigger than one season and half (around 30 weeks).

That's why in the next Chapter, I introduce a weighted approach that exploit an infinite window into the past because also farer information looks significant.

4.4: WEIGHTED SOLUTION

4.4.1: THE INFINITE WINDOW INTO THE PAST

This kind of approach is based on an infinitive window into the past, that means that I use all the data in the past for every week that I predict.

This method assigns more relevance on the most recent matches of every team; these difference degree of relevance is given by a negative exponential function built on the data that I want to use to predict (all the matches available in the past, from the first week of 2002).

It should be clear that if I'm predicting the first week of 2004, I'm using a window that is large 2 years; if I'm forecasting the first week of 2008, I'm using a window that is large 6 years (in this last case obviously the first years available has just a small part of relevance in the further predictions).

In fact it's reasonable to think that a match played in the "far" 2002 can't be useful to predict a game played in 2009 for example (that is due to a "generation change" in almost all the teams approximately every 10 years, in terms of staff, as coaches, players and all the people involved in the teams).

4.4.2: THE WEIGHT'S FUNCTION

After all the considerations made up on the relevance of the games in past, the main conclusion is that the team parameters should be more influenced by the most recent games.

As a consequence of this, the natural weight's function is a negative exponential function built on all the previous weeks (as regards to the single week that I'm going to forecast).

To assign the weights to the weeks relative to every single prediction, I decide in a first instance to use a parameter (α) to manage the "speed" of the negative exponential function; the bigger α is chosen, the less weight is assigned to the farer matches.

In a second moment, I thought that I can't consider all the games as an unique block of 168 weeks, without taking care about the changes of the seasons; I mean, there isn't the same relation between the 5^{th} and the 6^{th} week, if compared to the relation between 21^{st} (Super Bowl of the first season, 2002) and 22^{nd} (first match of the second season in analysis, 2003).

That's why I introduced in this function a second parameter (β) able to handle with the problem of the change of season, establishing which "jump" is more suitable among the seasons.

The following formula is the weight's function that I adopted for this approach; it attributes different weights to every single week before the week w that I want to forecast:

$$\exp\{\alpha*((W\_tot=i)-w)\} + \beta*((season[W\_tot=i])-2002)$$

For i in 1…w-1, where w obviously indicates the week that I want to predict.

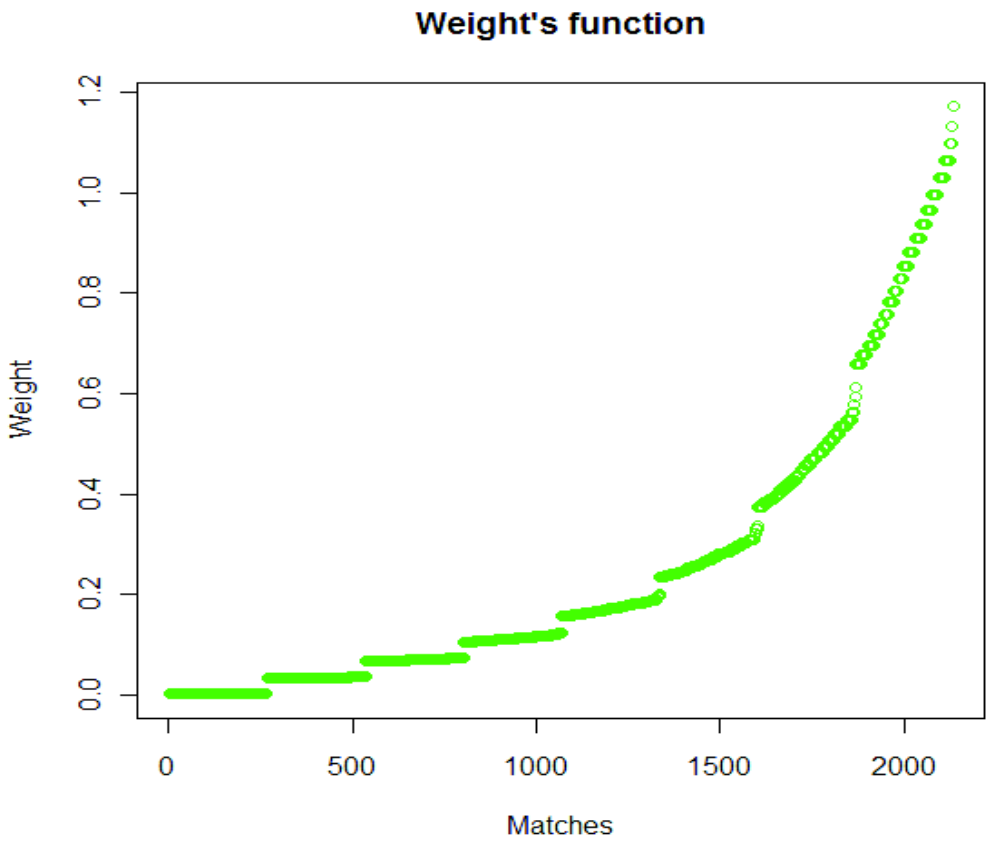
Note: the notation season[W_tot=i] stands for the season where W_tot corresponds to i.

In this way, I can build all the different weight's functions relative to every single week in the 6 seasons that I want to predict (from the 2004 to the 2009).

For a good interpretation of such function I purposed the weight's function used to predict the last week (168^{th}) to have the function spread in all the 8 years involved in the analysis

The distribution of the weight's function with α=0.04 and β=0.03, to predict the 168^{th} week is described in the following graph.

Graphic 4.3



In the graph of the previous page is possible to clearly recognize the 8 seasons involved in the analysis, and the 7 "jumps" due to the changes of the season.

The first parameter, α, sets the speed of the function to go down; the second parameter, β, sets the largeness of the "jumps" among the seasons.

My main aim, in this approach, is to look for the best combination of the 2 parameters involved in the function (α and β), finding the smallest margin of error (throughout the use of the MSE index).

4.4.3: THE WEIGHTED SPREAD POINT FOR THE DIFFERENCE

For this approach, I calculate once again the predictions restricted at the season 2006, 2007 and 2008, in order to have also in this case the possibility to compare my results with the bookmakers' line, in the case of the difference of the points, assuming the same model of the paragraph 4.3.2.

The model assumed is:

$$D_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where:

$$\mu_i = \mu + \beta_1(hteam)_i - \beta_2(ateam)_i$$

The following table contains the mean of MSEs obtained for different choices of the 2 parameters

Table 4.3

α / β	0,00	0,02	0,04	0,06	0,08
0,02	13,94922	13,95365	13,95710	13,95987	13,96214
0,04	13,85281	13,86407	13,87630	13,88712	13.89631
0,06	13,86636	13,84304	13,84919	13,86003	13,87078
0,08	13,92489	13,83692	13,83417	13,84456	13,85655
0,10	14,00529	13,83378	13,82409	13,83510	13,84864
0,12	14,09901	13,83182	13,81783	13,83020	13,84527
0,14	14,20191	13,83093	13,81481	13,82876	13,84513
0,16	14,31137	13,83108	**13,81432**	13,82979	13,84714
0,18	14,42554	13,83211	13,81568	13,83246	13,85047
0,20	14,54305	13,83380	13,81829	13,83616	13,85457

Note: I highlighted in red (bold) the best combination (in terms of MSE) between the 2 parameters involved in the function.

The weight's function with the parameter: α = 0,16 and β = 0,04 generates the best results in terms of reliability of the model.

Table 4.4

Year	ME	BOOK	GAP	CCI
2006	13,67743	13,48817	0,18	52,43%
2007	13,62160	13,35387	0,26	50,56%
2008	14,13811	13,85654	0,28	49,43%

As we can see from the table above, these predictions are closer than before (if compared with the un-weighted approach) to the *Las Vegas line*, especially in season 2006.

In addition, this table denotes a good fit of the model, the average gap (with the *bookmakers*' line) is around 0,25. However the CCI can't be consider satisfactory, even if is a bit higher than the results achieved with the un-weighted methodology.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare the 2 indexes in the same scale:

$$SD(D) = \sqrt{VAR(D)} = 14,89$$

The gap between my average MSE (13,81238) and the standard deviation is 1,08567; this is good signal and it confirms the advantages in terms of gaps and CCIs (in the 3 years) obtained using this weighted model to predict the difference of the points.

4.4.4: THE WEIGHTED SPREAD POINT FOR THE SUM

The parallel analysis between the difference of the points and the sum of the points continues also in the weighted approach, where clearly the weights applied to the design matrix are different, if compared to the previous choices for α and β.

The model assumed is:

$$S_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where:

$$\mu_i = \beta_0 + \beta_1(hteam)_i + \beta_2(ateam)_i$$

As for the prediction of the difference of the points, I present a table that cross α and β, for the prediction of the sum.

In specific the table contains the mean of MSEs obtained for different choices of the 2 parameters involved in the weight's function:

Table 4.5

α / β	0,00	0,02	0,04	0,06	0,08
0,02	15,04559	14,23096	14,18755	14,19411	14,20664
0,04	14,16842	14,17771	14,19018	14,20193	14,21226
0,06	14,19100	**14,15983**	14,16491	14,17646	14,18839
0,08	14,27395	14,16902	14,16186	14,17177	14,18315
0,10	14,38429	14,18430	14,16542	14,17274	14,18457
0,12	14,50792	14,19891	14,17053	14,17660	14,18840

Note: I highlighted in red the best combination (in terms of MSE) between the 2 parameters involved in the function. In addition, on the table just presented I include also the column relative to β=0 to emphasize the relevance of this second parameter; it's clear that it has been worth introducing β in the weight's function.

The weight's function with the parameters: α = 0,06 and β = 0,02 generates the best results in terms of reliability of the model.

Table 4.6

Year	ME	BOOK	GAP	CCI
2006	14,13227	13,82742	0,30485	47,56%
2007	14,60504	14,13091	0,47413	49,06%
2008	13,72858	13,46114	0,26744	51,68%

Also in this situation the weighted approach works better than the un-weighted one, however the CCI continues to be not satisfactory.

The average *gap* among the years in analysis is 0,34, that confirms the relevance due to the use of the weight's function (includes the comparisons between the 2 parameters involved).

To evaluate the quality of the MSE (obtained with the weighted model), it's natural to take a look to the main index of variability: the variance, it results equal to 201,90.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare the 2 indexes in the same scale:

$$SD(S) = \sqrt{VAR(S)} = 14{,}20$$

The gap between my average MSE (14.15) and the standard deviation is 0.05; at least with the weighted approach my results (in mean) are smaller than the SD(S).

By the way the weighted method brings improvements on this simple model that I have implemented in the first part of this thesis.

CHAPTER FIVE: THE MODEL INVOLVES THE YARDS AND THE TURNOVER

The turnover and the yards are the most interesting indexes (thus the most important explicative variables) available for every NFL match, after the consideration of the main information connected with to the site of the game (the study of the *home effect*).

At a first moment to understand the phenomenon (the relationship among the response variables and the transformations of yards and turnover indexes), I tried these following simple models (similarly to the first 2 explorative models purposed at the beginning of the Chapter 4):

For the difference of the points, I assumed:

$$D_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

where

$$\mu_i = \mu + \beta_1(y\_diff)_i + \beta_2(dto)_i$$

Where y_diff is the difference between the yards gained at home and the yards gained away and vice versa dto is the difference between the turnover index away and the turnover index at home.

Equivalently, the same approach will be followed also for the model that predicts the sum of the points, substituting the differences of the yards and the differences of the turnover indexes with the respective sums, where I assumed:

$$S_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

where

$$\mu_i = \beta_0 + \beta_1(y\_sum)_i + \beta_2(sto)_i$$

Where y_sum is the sum between the yards gained at home and the yards gained away and sto is the sum between the turnover index away and the turnover index at home.

These models fit really good the data in exam, and this is confirmed from the adjusted R squared of the model that in both cases is more than 60%.

Obviously, these models cannot be applied directly because we don't know how many yards a team will gain and which will be the turnover index of a future match; that's why we need to predict the future yards and the future turnover indexes to apply such models.

Substantially, my idea is to predict separately the differences of the yards (between the home team and the away team) and the differences of the turnover indexes (between the away team and the home team), with a weighted model that exploits, as for the previous weighted models, the variables relative to the home team and to the away team (focus on the *home effect*).

Essentially, for the first model I thought that throughout the *home effect* study, extended to the prediction of yards and the turnover indexes, I could define somehow the difference of the points between the home and the away team (spread point for the difference).

Naturally, the analysis has been carried out also for the sum, using respectively the sum of the points as response variable, extending once again, the lighter *home effect* to the predictions of yards and turnover indexes.

Finally the idea is to create 2 models (one for the sum and one for the difference) based on the predictions of the difference of the yards (and respectively on the sum of the yards) and on the difference of the turnover indexes (and respectively on the sum of the turnover indexes).

5.1: THE DIFFERENCE OF THE POINTS

5.1.1: THE YARDS PREDICTION

The model that I assumed to predict the difference of the yards between the home team and the away team is, following the same approach of the chapter 4:

The model assumed is:

$$Y\_diff_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where:

$$\mu_i = \beta_0 + \beta_1(hteam)_i - \beta_2(ateam)_i$$

using the same technique as before, with the design matrix with 1 and -1 and the matrix of the weights (with the best choices for α and β for that particular prediction).

Thanks to this kind of prediction is not possible to compare any bookmaker's line because effectively it is not a real result of the game, but surely a significant statistic.

Table 5.1

α/β	0,00	0,02	0,04	0,06	0,08
0,02	112,4088	112,4088	112,4445	112,4571	112,4675
0,04	**111,6037**	111,6858	111,7756	111,8548	111,9220
0,06	111,9366	111,6913	111,6974	111,7521	111,8139

In this particular case the best function of the weights is given from: α=0,04 and β=0. It's evident that in this case it is better to use the weights without considering the parameter β relative to the "jumps" among the seasons.

To evaluate the quality of the MSE (obtained with the weighted model), it's natural to take a look to the main index of variability: the variance, it results equal to 14964,50.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare 2 indexes in the same scale:

$$\sqrt{VAR(y\_diff)} = 122{,}32$$

The gap between my best MSE (111,60) and the standard deviation is 10,72, thus I can consider this model a good predictor for the difference of the yards.

5.1.2: THE TURNOVER INDEXES PREDICTION

The model that I assumed to predict the difference of the turnover indexes between the away team and the home team is:

$$dto_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where:

$$\mu_i = \beta_0 + \beta_1(ateam)_i - \beta_2(hteam)_i$$

For this prediction I tried also the model with the weights but the un-weighted model results more efficient, that's why I use this approach to predict the difference of the turnover indexes (between the away team and the home team).

The MSE that results from the un-weighted model is: 2.07; I tried many other models with different weights and the best result that I achieved was around 2.08, that's why I choose the un-weighted approach for this specific prediction.

To evaluate the quality of the minimum MSE (obtained with the un-weighted model), it's natural to take a look to the main index of variability: the variance, it results equal to 4.41.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare 2 indexes in the same scale:

$$SD(dto) = \sqrt{VAR(dto)} = 2.10$$

The gap between my best MSE (2,073034) and the standard deviation is 0,03

5.1.3: THE FINAL MODEL

At this point, I have the best predictions for the differences of the yards and the differences of the turnover indexes, calculated for the 3 years available for the comparison with the *Las Vegas line*

The model aims to predict the difference of the points (between the home team and the away team) using the 2 explicative variables just forecasted.

The model that I assumed is:

$$D_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where all the values for μ are calculated using the following formula:

$$\mu_i = \beta_0 + \beta_1(y\_diff)_i + \beta_2(dto)_i$$

The model that I'm using for such analysis must avoid the weighted approach because it doesn't involve the teams but just 2 explicative variables that describe the difference of the yards and the difference of the turnover indexes respectively (there's no motivation to apply the weights).

In the models presented in the Chapter 4 the explicative variables were the home team and the away team, for which a dynamic model based on the weights was justified; in fact it's reasonable to think that some teams follow specific trends over the time (dynamic approach).

The un-weighted model based on yards and turnover brings to the best results in terms of *gap* and CCI, restricted to the season 2006 and 2008.

Table 5.2

Year	ME	BOOK	GAP	CCI
2006	14,03834	13,48817	0,55017	52,80%
2007	14,29009	13,35387	0,93622	49,06%
2008	14,03834	13,85654	0,1818	54,68%

The average gap is 0,55, but it's evident that the model forecasts good for 2006 and 2008 and works bad for the 2007.

By the way, the index of correct classification (specifically in 2006 and 2008) is definitely most significant than the previous models from a gambling point of view.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare 2 indexes in the same scale:

$$SD(D) = \sqrt{VAR(D)} = 14{,}89$$

The *gap* between my best MSE (14.12051) and the standard deviation is 0.78; even if this approach provides a gap (between my results and the SD(D)) smaller than the weighted solution that I showed in the paragraph 4.4.3, this model can be considered the most efficient exclusively referred to the 2006 and the 2008 (as it's clear from the table just presented).

5.2: THE SUM OF THE POINTS

5.2.1: THE YARDS PREDICTION

The model that I assumed to predict the sum of the yards between the home team and the away team is:

$$Y\_sum_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where:

$$\mu_i = \beta_0 + \beta_1 hteam + \beta_2 ateam$$

Using the same technique as before, with the design matrix with 1 (corresponding to the team that plays at home and to the team that plays away) and the matrix of the weights (with the best choices for α and β for this particular prediction).

Also in this kind of prediction it is not possible to compare any bookmaker's line because effectively is not a real result of the game, but surely a significant statistic.

Table 5.3

Alpha/beta	0,00	0,02	0,04	0,06	0,08
0,02	108,5913	108,6269	108,6557	108,6793	108,6992
0,04	107,1904	107,3486	107,5056	107,6422	107,7582
0,06	107,0710	106,9952	107,1402	107,3056	107,4557
0,08	107,5635	**106,9800**	107,0601	107,2260	107,3880
0,1	108,3548	107,0819	107,0894	107,2489	107,4145

In this particular case the best function of the weights is given setting: α=0,08 and β=0,02. It's clear that in this case, it is better to use the weights considering also the parameter β relative to the "jumps" among the seasons.

To evaluate the quality of the MSEs (obtained with the weighted model), it's natural to take a look to the main index of variability: the variance, it results equal to 13352,73.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare 2 indexes in the same scale:

$$\sqrt{VAR(y\_sum)} = 115{,}55$$

The gap between my best MSE (106,98) and the standard deviation is 8,57. The gap is significant and outlines a good prediction for the sum of the yards.

5.2.2: THE TURNOVER PREDICTION

The model that I assumed to predict the difference of the turnover indexes between the away team and the home team is:

The model assumed is:

$$sto_i \sim N(\mu_i, \boldsymbol{\sigma}^2)$$

Where:

$$\mu_i = \beta_0 + \beta_1 hteam + \beta_2 ateam$$

For this prediction I tried also the model with the weights but the un-weighted model results also in this situation more efficient; that's why I carried out the un-weighted approach to predict the sum of the turnover indexes (between the away team and the home team).

The output for the normal model (without weights) provides an MSE equals to 1,82 that I will compare with the standard deviation relative to the sum of the turnover indexes.

However I also tried the model with the weights but the output, in terms of MSE, was around 2,10, definitely less reliable than the model that I will use to predict the sum of the yards.

Once again, to evaluate the quality of the minimum MSE (obtained with the weighted model), it's natural to take a look to the main index of variability: the variance, it results equal to 3,32.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare 2 indexes in the same scale:

$$\sqrt{VAR(sto)} = 1{,}82$$

The gap between my best MSE (1,82) and the standard deviation is 0. Even if this difference results 0 (not such a good signal) I have to carry out the final model using these predictions because they represents the best results that I had achieved.

5.2.3: THE FINAL MODEL

At this point, I have the best predictions for the sums of the yards and the sums of the turnover indexes, calculated for the 3 years available for the comparison with the *Las Vegas line.*

The model aims to predict the sum of the points (between the home team and the away team) using the 2 explicative variables just forecasted.

The model implemented is described from the following formula:

$$\mu_i = \beta_0 + \beta_1(y\_sum)_i + \beta_2(sto)_i$$

For the same reasons explained in the paragraph 5.1.3, the model that I'm using for such analysis must avoid the weighted approach because it doesn't involve the teams (where there is an evolution over the time) but just 2 explicative variables that describe the difference of the yards and the difference of the turnover indexes respectively.

The un-weighted model based on yards and turnover brings to the following results:

Table 5.4

Year	ME	BOOK	GAP	CCI
2006	14,36473	13,82742	0,53	49,81%
2007	14,94017	14,13091	0,81	50,56%
2008	14,03717	13,46114	0,58	50,93%

The average gap is 0,64, but it's evident that the model forecasts better for 2006 and 2008 rather than 2007 (as before), in terms of gap (between my prediction and the Las Vegas line).

By the way the 50% of correct classification (2007 and 2008) is lightly more significant than 2006 from the gambling point of view.

Now for the comparison I consider the squared root of the variance (or standard deviation) to be able to compare 2 indexes in the same scale:

$$SD(S) = \sqrt{VAR(S)} = 14.20$$

The gap between my average MSE index (14.44) and the standard deviation is 0.24; this model can't be considered efficient as the model for the sum carried out in the previous paragraph.

5.3: HOW TO EVALUATE THE STRATEGY TO GAMBLE

In this paragraph my purpose is to focus on (the study of) a possible profitable strategy to gamble using the results that I have just predicted with the 2 previous models (5.1, 5.2).

The results of my models looks totally satisfactory just in terms of MSEs (but not in terms of CCIs), in fact, except for 2007 the *gaps* between my results and the bookmakers are really light.

On the other hand, the CCI indexes aren't so reliable for an smart betting strategy; the best performance I obtained is a correct classification to the 55% of the matches (with the model implemented in this Chapter, related to the prediction of the difference of the points).

By the way we can't consider it, as a reliable method to "make the war" to the bookmakers, because even if the result about MSEs are comforting, the "row" index CCI can't assure a profitable betting strategy over the time.

Having a look at the results (in particular the matrix that contains: my results, the real results and Las Vegas line), I noticed that sometimes, when my predictions are really close to the bookmakers' ones, the risk to classify wrong the result is higher than when my predictions are farer from the Las Vegas line.

This is the main reason why I defined "row" the CCI; in fact such index may bring to a non-correct classification for a match, even if the MSE is really low and consequently close to the *Las Vegas line*.

That's why the gaps between my MSEs and the bookmaker's MSEs are relatively low and at the same time the CCI indexes are not satisfactory at all.

By the way, my idea is to consider less matches in the betting strategy and try to improve the CCI index, that is the unique measure (index) that can assure profits to the gambler.

The new matrix for this new analysis contains: my results, the real results, the Las Vegas line and the difference between my results and the bookmakers' line.

This strategy wants to consider the advantages to exploit with more reliability the results of the model that provides an higher *bias* if compared with the bookmakers' wagers.

As a consequence of this I introduce a parameter, k, that describes the distance between the 2 quantities in analysis (my predictions and the *Las Vegas line*); I fixed that this new parameter may vary from 1 to 10 (for values bigger than 10 remain just few predictions).

Furthermore, I include also the percentage of matches that would be included in the betting strategy (obviously this percentage decreases when the distance increases), in a specific year.

5.3.1: THE STRATEGY FOR THE DIFFERENCE OF THE POINTS

In the following table I show the CCI indexes and the relative percentages of games involved in the bet, when the parameter k varies from 1 to 10.

Table 5.5

k	1	2	3	4	5	6	7	8	9	10
2006										
CCI	**52,45%**	**53,95%**	**57,52%**	**57,45%**	**56,06%**	**54,55%**	**60,00%**	**64,71%**	**63,64%**	**66,66%**
% match	76,00%	56,00%	42,00%	35,00%	24,00%	16,00%	9,00%	6,00%	4,00%	2,00%
2007										
CCI	**46,90%**	**47,54%**	**48,95%**	**50,94%**	**50,00%**	**49,09%**	**52,38%**	**54,84%**	**72,22%**	**61,53%**
% match	84,00%	68,00%	53,00%	39,00%	29,00%	20,00%	15,00%	11,00%	6,00%	4,00%
2008										
CCI	**55,61%**	**54,97%**	**54,69%**	**52,94%**	**51,85%**	**52,94%**	**50,00%**	**77,78%**	**66,66%**	**50,00%**
% match	83,00%	64,00%	47,00%	31,00%	20,00%	12,00%	6,00%	3,00%	1,00%	1,00%

Note: The % match is the ratio between how many games are involved for that specific choice of k and 267 (number of matches per year)

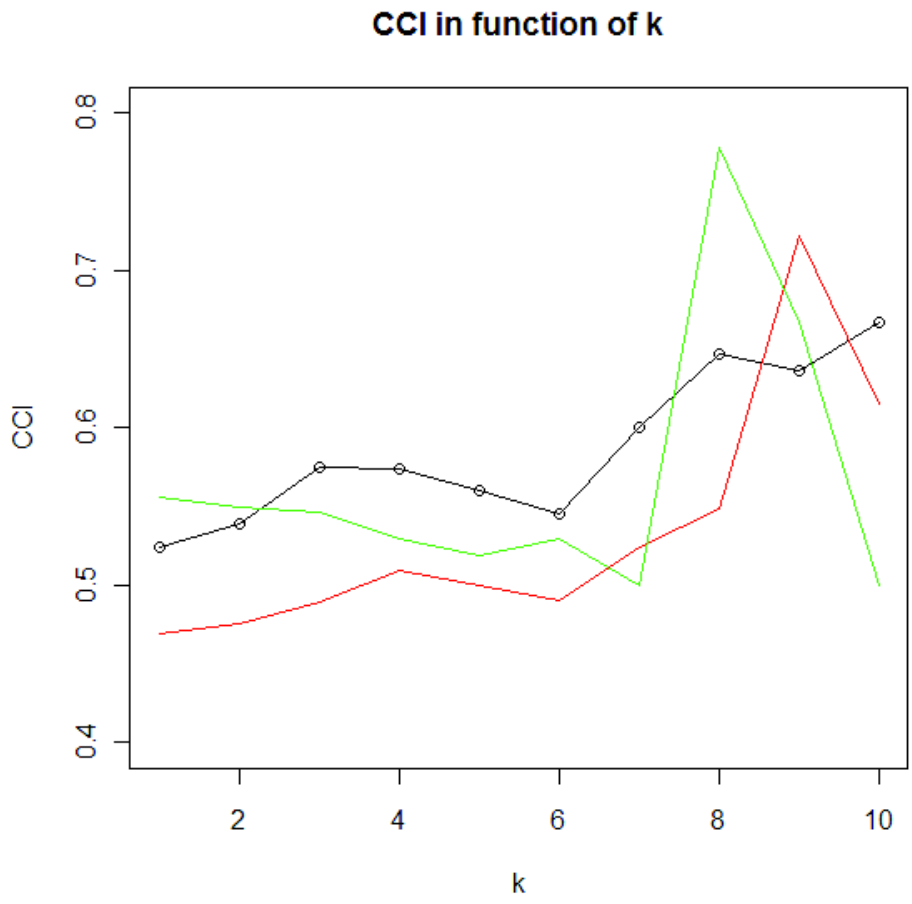
The percentage of the matches confirms that the advantageous bets are based on just few games per year and here the gambler must decide which amplitude of the distance k is the best tradeoff for a profitable betting strategy.

This kind of comparison permits to the bettor to identify in which matches is more convenient to orient the betting strategy; on consequence of this, the gambler can evaluate the respective *odds* (in such games) and draw up a profitable way to “make the war” to the “bookies”.

In the next page to evidence the importance of the parameter k, I built a graphic that may help the gambler in the decision of the best trade-off between the *CCI (Correct Classification Index)* and the percentage of the games that would be involved in the betting strategy.

The graphic confirms that when the distance (described from the parameter k) from my results and the bookmakers' ones tends to increase, the reliability of my predictions is higher.

Graphic 5.1



Note: The black line denotes the series of season 2006, the red line denotes the series of season 2007 and the green line denotes the series of the season 2008.

Finally the best choice for the parameter k seems to be k=8, that provides an average CCI equals to 65,78%, betting on 54 games out of 801 (total of the matches in the 3 years in analysis).

5.3.2: THE STRATEGY FOR THE SUM OF THE POINTS

In the following table I show the CCI indexes and the relative percentages of games involved in that particular bet, when the parameter k varies from 1 to 10.

Table 5.6

k	1	2	3	4	5	6	7	8	9	10
2006										
CCI	**50,00%**	**51,70%**	**52,51%**	**52,47%**	**48,00%**	**48,33%**	**50,00%**	**55,88%**	**48,14%**	**45,00%**
% match	82,00%	65,00%	52,00%	37,00%	28,00%	22,00%	16,00%	13,00%	10,00%	7,00%
2007										
CCI	**52,23%**	**51,87%**	**52,98%**	**55,00%**	**53,00%**	**50,74%**	**52,17%**	**55,26%**	**57,69%**	**56,52%**
% match	83,00%	70,00%	56,00%	44,00%	37,00%	25,00%	17,00%	14,00%	9,00%	8,00%
2008										
CCI	**49,54%**	**47,19%**	**45,11%**	**46,23%**	**45,94%**	**45,65%**	**42,42%**	**50,00%**	**61,53%**	**58,33%**
% match	81,00%	66,00%	49,00%	34,00%	27,00%	17,00%	12,00%	6,00%	4,00%	4,00%

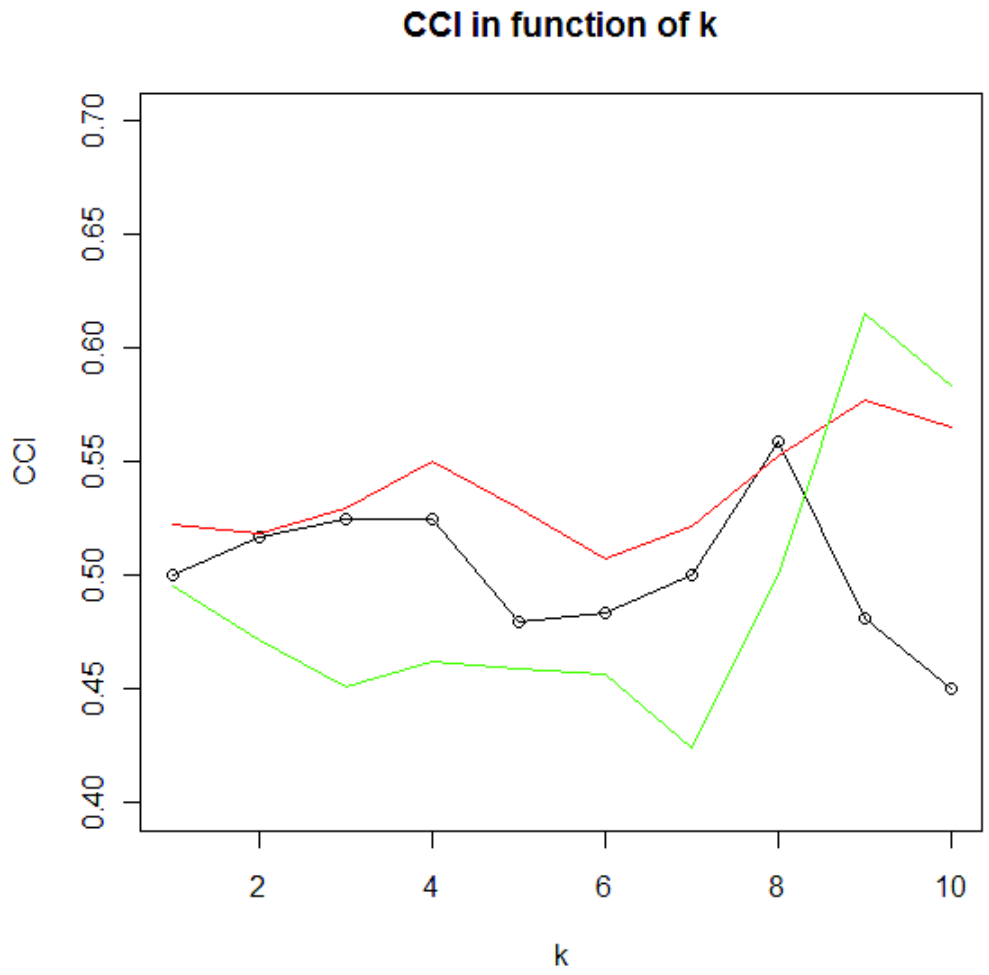
Note: The % match is the ratio between how many games are involved for that specific choice of k and 267 (number of matches per year)

In this specific prediction the advantage is lighter than before, even if also in this case, higher distances involved just few games per year. The role of the smart gambler, at this point of the analysis, is to decide which amplitude of the distance k is the best tradeoff for a profitable betting strategy (between k and the relative number of game involved).

Anyway, the model for the sum used for this predictions isn't satisfactory at all; that's why also the method that choose k, as the best trade-off in terms of CCI, doesn't bring to such brilliant results like for the prediction of the differences of the points.

The next graphic confirms that when the distance (described from the parameter k) from my results and the bookmakers' ones tends to increase, the reliability of my predictions is higher.

Graphic 5.2



Note: The black line denotes the series of the season 2006, the red line denotes the series of the season 2007 and the green line denotes the series of the season 2008.

The graphic shows that the model is working bad for 2008, however with a larger distance also this prediction gives an acceptable output in term of CCI.

Finally, the most convenient choice for the parameter k, also in this case, seems to be k=8, that provides an average CCI equals to 53,71%, betting on 89 games out of 801 (total of the matches in the 3 years in analysis).

CONCLUSIONS

COMPARISON AMONG THE MODELS

For the analyses that I carried out, I had considered 3 different approaches to forecast the difference of the points and the sum of the points between the team that plays at home and the team that plays away.

A) *MOVING WINDOW* APPROACH (see subchapter 4.3)

- THE DIFFERENCE OF THE POINTS

Table 4.1

Year	ME	BOOK	GAP	CCI
2006	13,84896	13,48817	0,36	46,81%
2007	14,10719	13,35387	0,75	49,81%
2008	14,31934	13,85654	0,46	**52,43%**

- THE SUM OF THE POINTS

Table 4.2

Year	ME	BOOK	GAP	CCI
2006	14,24129	13,82403	0,42	49,81%
2007	14,87784	14,13091	0,75	46,81%
2008	13,80730	13,46114	0,35	49,81%

The *moving window* approach, actually provides the worst results in terms of gap (from the bookmakers' line) and in terms of *CCI (Correct Classification Index)*.

However, this method highlights that to obtain a good prediction it is necessary to consider a data window (information in the past) larger than 30 weeks.

B) WEIGHTED APPROACH (see subchapter 4.4)

- THE DIFFERENCE OF THE POINTS

Table 4.4

Year	ME	BOOK	GAP	CCI
2006	13,67743	13,48817	0,19	**52,43%**
2007	13,62160	13,35387	0,27	**50,56%**
2008	14,13811	13,85654	0,28	49,43%

- THE SUM OF THE POINTS

Table 4.6

Year	ME	BOOK	GAP	CCI
2006	14,13227	13,82742	0,30	47,56%
2007	14,60504	14,13091	0,47	49,06%
2008	13,72858	13,46114	0,26	**51,68%**

The weighted model works better (as in the previous approach), predicting the difference of the points, where 2 times out of 3 the CCI results more than the 50%.

In addition in the model for the difference provides really small gaps, if compared with the *Las Vegas line*, as I have already evidenced in the Chapter 4.

Anyway, the results are not so comforting because the maximum value reached from the *Correct Classification Index* is 52,43%.

As discussed before, the results that derive from the CCI aren't a reliable tool to evaluate the accuracy of the modeling methodology adopted, because sometimes, when my prediction is really near to the bookmaker's one, the risk to classify wrongly the game is higher than when my prediction is farer from the *Las Vegas line*.

By the way, in terms of *gaps*, the results obtained for the weighted solution can be considered quite good, if compared with the *mowing window* approach.

C) YARDS AND TURNOVER APPROACH (see Chapter 5)

- THE DIFFERENCE OF THE POINTS

Table 5.2

Year	ME	BOOK	GAP	CCI
2006	14,03834	13,48817	0,55	**52,80%**
2007	14,29009	13,35387	0,93	49,06%
2008	14,03834	13,85654	0,18	**54,68%**

- THE SUM OF THE POINTS

Table 5.4

Year	ME	BOOK	GAP	CCI
2006	14,36473	13,82742	0,54	49,81%
2007	14,94017	14,13091	0,81	**50,56%**
2008	14,03717	13,46114	0,57	**50,93%**

As I evidenced in the Chapter 5, this method is the best way to forecast the difference of the points (particularly emphasizes in the subchapter 5.3).

Anyway, also the prediction for the sum of the points provides 2 out of 3 CCI indexes that result more than the 50%, and also the *gaps* are around the 0,50 (except for the 2007).

POINTS OF STRENGTH AND WEAKNESS

All the analyses carried out in this thesis give important information on the prediction of the National Football League results.

The un-weighted methodology underlines the importance of the historical data, consequently the weighted method improves the estimations (using all the data available into the past) throughout the use of a specific weight's function.

Finally the models based on the predictions of yards and turnovers offer the higher level of reliability to establish the correct outcomes of the American football.

The *Correct Classification Index* (CCI) is the main point of weakness, in fact it can't be consider a valid tool to establish the accuracy of a certain model in this particular contest.

On the contrary, the main point of strength of this kind of analysis is represented by the betting strategy explained in the subchapter 5.3, it exploits the CCI in order to define the most interesting bets (available in the years in analysis) for a possible profitable betting strategy.

By the way, to contrast properly the bookmakers I should study singularly every single team, observing its performances over the time (points, yards, turnovers) and taking care about the most important players of every single team.

APPENDIX

The main goal of this appendix is to complete the thesis showing the main functions that I exploited to achieve the results.

In specific I show the R code associated to the model introduced in the subchapter 5.1 because it contains all the main functions implemented in this thesis.

```
#HOW TO MANAGE THE VARIABLES WEEK AND DTO

data$hteam <- as.factor(data$hteam)

data$ateam <- as.factor(data$ateam)

dto <- data$ato-data$hto

data <- cbind(data,dto)

data$Week <- as.factor(data$Week)

W <-
ordered(data$Week,levels=c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","WildC
ard","Division","ConfChamp","SuperBowl"))

W <- as.numeric(W)

data <- cbind(data,W)

W_tot <- (data$season-2002)*21 + data$W

data <- cbind(data,W_tot)


#HOW TO MANAGE THE HOME EFFECT

X <- matrix(0,2136,32)

h <- as.integer(data$hteam)

for (i in 1:NROW(data)) X[i,h[i]] <- -1

h <- as.integer(data$ateam)

for (i in 1:NROW(data)) X[i,h[i]] <- 1

int <- rep(1,2136)

X <- X[,-1]

X <- cbind(int,X)

data <- cbind(data,X)
```

```
#LEAST SQUARES METHOD

MMQ <- function(data.stima,data.predict)

{

a <- as.matrix(data.stima[,(16:47)])

y <- as.vector(data.stima$dto)

n <- as.matrix(data.predict[,(16:47)])

b_hat <- (solve(t(a)%*%(a)))%*%(t(a)%*%(y))

b_hat <- as.vector(b_hat)

p <- n%*%b_hat

r <- data.predict$dto

cfr <- cbind(p,r)

cfr <- data.frame(cfr)

return(cfr)

}


#THE FOLLOWING FUNCTION RETURNS A MATRIX 63x2 WITH THE PREDICTIONS AND THE REAL RESULTS
#IN THE SEASONS 2006, 2007 AND 2008

cfrP <- function(d)

{

g <- NULL

for (i in (1:63))

{

g <- rbind(g,MMQ(d[d$W_tot>=1 & d$W_tot<(84+i),],d[d$W_tot==(84+i),]))

}

return(g)

}

pp_dto <- cfrP(data)


#I TAKE OUT THE 2nd COLUMN ASSOCIATED TO THE REAL RESULTS AND THE DESIGN MATRIX

pp_dto <- pp_dto[,-2]

data <- data[,1:15]
```

```
#THE WEIGHT'S FUNCTION

pesi_new <- function(d,w,alpha=0.04,beta=0)

{

peso <- NULL

for (i in (1:(w-1)))

  {

peso <- c(peso,(exp(alpha*((d$W_tot[d$W_tot==i]-w)))+beta*(d$season[d$W_tot==i]-2002)))

  }

return(data.frame(cbind(d[(d$W_tot>=1 & d$W_tot<=(w-1)),],peso)))

}


#THE NEW DESIGN MATRIX FOR THE PREDICTION OF THE YARDS

X <- matrix(0,2136,32)

h <- as.integer(data$hteam)

for (i in 1:NROW(data)) X[i,h[i]] <- 1

h <- as.integer(data$ateam)

for (i in 1:NROW(data)) X[i,h[i]] <- -1

int <- rep(1,2136)

X <- X[,-1]

X <- cbind(int,X)

data <- cbind(data,X)

ydiff <- data$hyds-data$ayds

data <- cbind(data,ydiff)


#WEIGHTED LEAST SQUARES METHOD

MMQP <- function(data.stima,data.predict)

{

a <- as.matrix(data.stima[,(16:47)])

y <- as.vector(data.stima$ydiff)

f <- data.stima[,49]

w <- matrix(diag(f),length(data.stima[,49]),length(data.stima[,49]))

n <- as.matrix(data.predict[,(16:47)])

b_hat <- (solve(t(a)%*%(w)%*%(a)))%*%(t(a)%*%(w)%*%(y))

b_hat <- as.vector(b_hat)
```

```
p <- n%*%b_hat

return(p)

}


#THE FOLLOWING FUNCTION RETURNS A VECTOR WITH THE PREDICTIONS ASSOCIATED TO THE SEASONS
#2006, 2007 AND 2008

cfrP <- function(d)

{

pp <- NULL

for (i in (1:63))

{

pp <- rbind(pp,MMQP(pesi_new(d,(84+i)),d[d$W_tot==(84+i),]))

}

return(pp)

}

pp_ydiff <- cfrP(data)


#THE FINAL MODEL

d <- data$hpts-data$apts

data <- cbind(data,d)

zeri <- rep(0,1068)

zerii <- rep (0,267)

pp_ydiff <- c(zeri,pp_ydiff,zerii)

pp_dto <- c(zeri,pp_dto,zerii)

data <- cbind(data,pp_ydiff,pp_dto)

fit <- function(data.stima,data.predict)

{

a <- data.stima$ydiff

b <- data.stima$dto

p <- data.stima$d

fit <- lm(p~a+b)

a <- data.predict$pp_ydiff

b <- data.predict$pp_dto

r <- data.predict$d
```

```
new <- data.frame(a,b)

pp <- predict(fit,new,type="response")

cfr <- cbind(pp,r)

cfr <- data.frame(cfr)

z <- cfr[,1]-cfr[,2]

MSE <- sqrt(sum(z^2)/nrow(cfr))

return(cfr)

}


#AS BEFORE, AS OUTPUT OF THE FOLLOWING FUNCTION A MATRIX 63x2 THAT CONTAINS THE
#PREDICTIONS AND THE REAL RESULTS FOR THE SEASONS: 2006, 2007, 2008

cfrN <- function(d)

{

g <- NULL

for (i in (1:63))

{

g <- rbind(g,fit(d[d$W_tot>=1 & d$W_tot<(84+i),],d[d$W_tot==(84+i),]))

}

return(g)

}


#THE LAS VEGAS LINE

line <- rbind(line.2006.mat,line.2007.mat,line.2008.mat)

line <- line[,-2]

line <- -line

a <- cfrN(data)

a <- cbind(a,line)


#2006, 2007, 2008

a1 <- a[1:267,]

a2 <- a[268:534,]

a3 <- a[535:801,]
```

```
#CORRECT CLASSIFICATION INDEX
CCI <- function(a)
{
p1 <- a[,1]-a[,3]
p2 <- a[,2]-a[,3]
pro <- p1*p2
b <- length(pro[pro>0])
R <- b/nrow(a)
mse <- sqrt(mean((a[,1]-a[,2])^2))
return(cbind(nrow(a)/267,mse,R))
}

#ff REPRESENTS THE GAP BETWEEN MY LINE AND THE LAS VEGAS' ONE
ff <- a[,1]-a[,3]
a <- cbind(a,ff)

#FUNCTION TO IDENTIFY THE BEST VALUE OF K FOR A PROFITABLE BET
Nov <- function(a)
{
gg <- NULL
for (K in (1:10))
{
gg <- rbind(gg,CCI(a[a[,4]>=K | a[,4]<=-K,]))
}
return(gg)
}

a1 <- a[1:267,]
a2 <- a[268:534,]
a3 <- a[535:801,]
res1 <- Nov(a1)
res2 <- Nov(a2)
res3 <- Nov(a3)
```

#GRAPHICS

```
plot(a1[,3])

lines(a1[,3])

lines(a2[,3],col=2)

plot(res1[,3],ylim=c(0.4,0.8),main="CCI in function of k",ylab="CCI",xlab="k")

lines(res1[,3])

lines(res2[,3],col=2)

lines(res3[,3],col=3)
```

BIBLIOGRAPHY

Glickman S. – Stern S. (1998), “A state space model for National Football League scores”, Journal of the American Statistical Association

Iacus S. – Masarotto G. (2007), “Laboratorio di Statistica con R”, Mc Graw Hill

Masarotto G., “Lucidi di Statistica Computazionale 1”

Coles S., “Lucidi di Statistica Computazionale 2”

WEBOGRAPHY

www.sportspunter.com

www.vegasinsider.com

www.wikipedia.it

SPECIAL THANKS

First of all, I have to underline the fundamental help of my parents Nico and Morena to achieve this degree, in terms of money but especially for the human support.

A special thank you to the teachers that helped me to write down this thesis, professor Stuart Coles and professor Guido Masarotto.

I have to highlight that the help of the professor Stuart Coles, with his work experience on the field of the sports betting, has been the best solution to analise and understand properly the analyses carried out in this thesis.

A special thank you to my girlfriend Alessandra and to her family for the constant support in all the years that I spent in the University.

A big thank you to all my friends, in the University life and in the “normal” life, because everybody has always trusted in me, giving to myself the strength to achieve this important goal.

Finally, a special thank you to my Erasmus friends that for sure I will never forget…