A market-based algorithm for predicting soccer outcomes

Víctor Hernández García†

Universidad Complutense de Madrid & FEDEA‡

June 2017

Abstract

The main objective of this study is to develop a market-based forecasting algorithm for predicting soccer outcomes. A football model with defensive and offensive parameters is estimated using information from prediction markets (Betfair Exchange Market). Prediction markets are an interesting testing ground to test the validity of new statistical techniques, since all participating agents observe the relevant information without asymmetries, the events have a clear completion date and the final outcome is not influenced by market activity. A time-homogeneous independent Poisson model is estimated applying the following techniques: OLS, Weighted OLS and hierarchical bayesian methods. The prediction capability of the proposed methodology is tested using “De Finetti” measure, which is defined as the Euclidean distance between the prediction and the realisation. The results obtained indicate that the models have good predictive properties, being able to improve market predictions, except in the last weeks of the season. So, it can be concluded that the use of information from prediction markets can lead to more accurate estimates of the probability distribution of future events.

Keywords: Forecasting, Prediction Markets, Bayesian Methods

JEL classification: C5 C6 D4 D8

*Preliminary draft. Please do not quote or cite without permission from the author. Comments welcome.
†Please direct correspondence to victorhzgarcia@gmail.com.
‡I thank José Ignacio Conde Ruiz and Marco Celentani for their help and support. I gratefully acknowledge the support from the Spanish Ministry of Science and Technology from grants ECO2011-30323-C03-01 & BES-2012-061597.
1 Introduction

The main objective of this study is to propose a market-based methodology for predicting the outcome of future events. High-frequency data from prediction markets are used as primary information. A prediction market is created for the purpose of trading the outcome of events, so equilibrium prices can be interpreted as the market belief about the probability of the different events. In these markets, the payoff is a binary option, in other words, agents get a fixed amount of money or nothing at all. Some examples of prediction markets are: futures markets, stock markets, political markets and (sport) betting markets.

Our initial hypothesis is that the use of information from prediction markets can improve the prediction capability of the models. Then we will examine whether prediction markets data can provide useful information about the probability distribution of future events.

In this work, we focus on sport betting markets instead of financial markets mainly for the following reasons. Firstly, the events have a clear completion date and the outcome is perfectly observed by all agents. Unlike financial markets, in betting markets the final outcome is not influenced by market activity and all participating agents observe the relevant information without asymmetries. Finally, betting markets are easily accessible and have a large volume of transactions. For the reasons given before, betting markets are an interesting testing ground to test the validity of new models and statistical techniques.

To the best of our knowledge, none of the models proposed so far in the sport economics literature have used information from these markets.

While we will use the Spanish Football League (LIGA) data as an illustrative example, the main goal of this paper is to lay out the methodology to systematically carry out inference from predictive markets data. This framework could be applied to a variety of structural models, sports and data sets.

The remainder of this paper is organized as follows. Firstly, in Section 2, a brief review of the sport economics literature is presented. Then, in Section 3, we describe the content of our database, which include high-frequency betting data from Betfair Exchange. In section 4, we introduce and justify the five assumptions we will use to develop the models. On the one hand, we just incorporate some of the classical assumptions in the literature, such as time-homogeneous Poisson and independence assumptions. On the other hand, we include some additional refinements, proposing an hypothetical partitioning of goal scoring intensity and allowing for fluctuations in team performance over the season. Section 5 describes the methodology and the forecasting models. Several estimation methods are proposed: (1) (weighted) ordinary least squares, (2) hierarchical bayesian model with normal random shocks and (3) hierarchical bayesian model with autorregresive shocks. In Section 6, this methods are applied to calculate predictions for single matches and to simulate Liga 2013-2014. The predictive power of the models, relative to Market predictions, is tested using De Finetti measure. Finally, in Section 7, conclusions are presented.
2 Literature review

In order to summarize the state of the art on sport economics, a selection of articles is presented:

Most studies on "modelling soccer data" are based only on "historical outcomes". Maher (1982) uses a poisson model for predicting football scores using past performance. An iterative ML (Newton Raphson) is used to estimate the following parameters: home defense, home attack, away defense and away attack. Dixon & Coles (1997) propose a poisson dynamic model that accounts for fluctuations in performance of individual teams. The model gives relatively more weight to the most recent results. Karlis & Ntzoufras (2003) relax the independent assumption and use a bivariate poisson distribution to predict soccer and water-polo scores using a expectation-maximization algorithm. Brillinger (2008) apply a trinomial model in order to model win, draw and loss probabilities using the outcome results. Everson & Goldsmith-Pinkham (2008) propose a bayesian hierarchical framework with Poisson additivity and home-field advantages to predict goal scoring. Finally, Karlis & Ntzoufras (2009) has employed the Skellam’s distribution to model the goal difference between home and away teams.

Some studies incorporate more information to the models, apart from the scores of matches already played. Dyte & Clark (2000) presents a log-linear Poisson for predicting the distribution of scores in international soccer matches taking the FIFA ratings as covariates. Bueno et al. (2010) proposes a bayesian methodology for predicting soccer outcomes based on experts’ opinions and past performance. Langseth (2013) extends the classical soccer models by taking a vast amount of data into account, including fired shots, shots on target, etc. Then a data intensive forecasting model is used in order to beat the bookie.

Several articles refer to betting markets when modelling soccer data. Dixon & Robinson (1998) propose a two-dimensional birth process to investigate setting prices in the spread betting market. Their model exploits only each teams past goal scores and the goal times within each match. Kuypers (2010) introduce a model of bookmaker behavior and look for profitable opportunities in soccer betting markets using an ordered probit.

However, to the best of our knowledge, none of the proposed models use predictive markets (betting data) as input to estimate the models. When modelling the stochastic processes that determine the goals scored by the different teams, we think that competitive markets provide a more complete information than the final scores that, after all, are a simple realisation of each stochastic process. The outcome of a particular match may depend greatly on the luck and not really on the quality of both teams. Therefore, the use of high-frequency information from predictive markets is the main contribution of this study to the literature.

Finally, this article is also very related to Prediction Markets literature (See Wolfers & Zitzewitz, 2006 for a review), which has grown significatively over the last years taking advantage of the availability of detailed data sets.
### 3 Data.

Betting markets have experienced a remarkably growth during the last decades principally due to deregulation, globalization and the emergence of online gambling. At present, according to the price setting mechanism, two types of betting markets can be distinguished: betting exchanges and fixed-odds markets.

Functioning of betting exchanges is quite similar to financial markets. In these markets, agents exchange odds in real time on sporting events and such transactions generate a competitive price at any given time. It is possible to buy and to sell an outcome, allowing backing and laying strategies. Bookmakers are simply intermediaries who fulfill the mission of bringing the different agents into contact and they generate revenue charging commissions on net winnings. In some cases, bookmakers use betting exchanges to hedge liabilities.

At the same time, in fixed-odds markets, bookmakers set prices to maximize profits and to achieve a balanced book. In this case, bookmakers offer odds and usually generate revenue by setting unfair odds (the sum of the implied probabilities is always greater than 1). In this market, punters have to bet against the bookmaker.

In this work, we only focus on betting Exchange for several reasons. Firstly, we are interested in competitive equilibrium prices because we are trying to take advantage of the market consensus. Secondly, odds are not altered by commissions as in the fixed-odds market case. Thirdly, it is possible to monitor the behavior of both sides of the market (back and lay). Finally, in betting exchanges there are no maximum bet limits.

Below is a summary of the data (Table 1). Our dataset include fully time-stamped historical Betfair Exchange Market price data\(^1\). The database covers all 380 matches of the Liga 2013-2014 and includes pre-play and in-play data for several markets (Match Odds\(^2\), Correct Score\(^3\), Over-Under betting, Half Time Score, Next Goal...). The Liga 2013-2014 is the latest available, as in November 2014 Betfair updated the application for downloading data\(^4\). The frequency is one observation per second and the following information is included for each selection:

- Best three back odds and its volume.
- Best three lay odds and its volume.
- Last price matched.
- Total matched (including both back and lay transactions).
- The status of the market (active or suspended).
- Timestamp.
- Pre-play or in-play

---

1. The data was downloaded from www.fracsoft.com
2. Match odds is a 3-way market where punters can bet either home, away or draw.
3. Correct Score is a 17-way market where punters can bet on the correct score (0-0,1-0,0-1...)
4. Betfair API 6.0
Table 1: Descriptive Statistics (Betfair Exchange)

<table>
<thead>
<tr>
<th></th>
<th>Match Odds</th>
<th>Correct score</th>
<th>Other markets a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>Observations (Total) b</td>
<td>22.8</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Observations (Pre-play)</td>
<td>15.5</td>
<td>0.4</td>
<td>0.31</td>
</tr>
<tr>
<td>Observations (In-play)</td>
<td>7.3</td>
<td>0.2</td>
<td>0.15</td>
</tr>
<tr>
<td>Observations (Active)</td>
<td>20.1</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td>Total Matched c</td>
<td>802</td>
<td>2.1</td>
<td>0.5</td>
</tr>
<tr>
<td>No. of matches</td>
<td>380</td>
<td>380</td>
<td>380</td>
</tr>
</tbody>
</table>

a They include: half time correct score, over/under markets, correct score away, correct score home, next goal and both teams to score.
b The data are expressed in millions of observations.
c The data are expressed in millions of euros.

It can be observed that match odds markets are the most liquid ones, matching 2.1 millions of euros per match on average. At the same time, correct score markets rank in the second place, according to the number of bets placed, with 0.28 millions of euros per match on average. Finally, the rest of the markets are much less liquid, so in these markets competitive prices do not reveal much information.

4 Model Assumptions.

The objective of this section is to introduce the five assumptions we will used to develop the forecasting models. Before that, some notation is required. In a match between teams indexed i (home team) and j (away team), we define the following variables:

- **Let** $X_{i,j}^H$ represents the goals scored by team i against team j, or equivalently, the goals conceded by team j against team i.
- **Let** $X_{i,j}^A$ represents the goals conceded by team i against team j, or equivalently, the goals scored by team j against team i.

4.1 The Poisson Assumption:

Assuming Poisson distributions for home and away goals is a classical assumption in the literature for predicting football scores (Everson & Goldsmith-Pinkham, 2008, Bueno et al., 2010). The Poisson distribution is a member of the exponential family and is actually a limiting case of a Binomial distribution when the number of trials grows and the success probability tends to zero. In a Poisson distribution, interarrival times follow a exponential distribution. Poisson distribution has a number of advantages which explain its wide use in empirical works, among them its simplicity (it is specified only for one parameter) and analytical tractability.

Before assuming the Poisson Assumption, some tests are carried out to check whether our data is well-approximated by a simple Poisson distribution.
Testing the Poisson Assumption:

i) **Variance/mean ratio test.** One of the most important characteristic of the Poisson distribution is that the variance equals the mean. We can use this property to check the validity of the Poisson assumption. According to (Simonoff, 2013), a comparison of the sample mean and variance would provide a valid way to check the validity of the Poisson assumption, given the inflation of variance of the negative binomial and other mixed-Poisson distributions. Following the methodology described in Karlis & Ntzoufras (2000), we calculate the variance to mean ratio\(^5\) for all 20 Liga teams (Figure 1). Note that 90-degrees line correspond to Poisson distribution, while points above the line are indications of overdispersion relative to the Poisson distribution. It can be observed that 12 of the 20 points (60%) are above the line, as their variance to mean ratio is greater than one\(^6\). Graphically, it is not possible to detect any pattern of deviation. For instance, under a negative binomial distribution, the variance is a linear function of the mean. In the case of the Poisson Inverse Gaussian, the variance is a quadratic function of mean.

![Figure 1: Index of Dispersion](image)

Figure 1: Index of Dispersion (goal variance / goal mean)

Now, using the estimations of the variance to mean ratios, we perform a test to check if there are evidence against the Poisson distribution. Under the null hypothesis (i.e. the data follow a Poisson distribution) we expect that half of the teams present underdispersion and the other half present overdispersion. Then we can compute a classical Proportion Test\(^7\) using this information. The confident interval includes 0.5 and p-value equals 0.37, so we can not reject the Poisson Assumption.

---

5 Also known as index of dispersion.
6 Karlis & Ntzoufras (2000) obtained a similar proportion. In their study the 58.1% of the teams showed overdispersion.
7 In Stata, the command is prtest.
ii) **Rootogram.** Everson & Goldsmith-Pinkham (2008) propose an alternative approach to check the validity of the Poisson Assumption. Following the methodology described in their article, we develop a Rootogram of Liga 2013-2014 goal scoring. Firstly, we compute the square roots of the counts of individual goals for the 380 matches played during Liga 2013-2014 (760 observations). Secondly, we compute the square roots of the expected Poisson counts for 760 draws from a theoretical Poisson distribution (with parameter equal to 1.357, i.e., the goal average during the season). Results are displayed using a Rootogram (Figure 2)\(^8\). In general, we can conclude that our soccer data are consistent with a Poisson distribution, although there are slight deviations between both rootograms. On the one hand, there are slight evidence that the true goal distribution has more density at zero. Several studies (See Karlis & Ntzoufras (2003)) obtained the same conclusion, proposing the use of a zero-inflated Poisson distributions instead of a Poisson distribution. At the same time, the actual count for seven goals is greater than the theoretical expected count. In both cases, deviations do not seem significant.

![Figure 2: Rootogram](image)

After ruling several tests, it seems that the Poisson distribution is appropriate and fits the data well. So, henceforth we assume that in each match home-team and away-team goals follow a Poisson distribution with parameters \(\lambda_{H,i,j}^T\) and \(\lambda_{A,i,j}^T\), respectively.

\[
\text{Assumption 1: } X_{H,i,j}^T \sim \text{Poisson}(\lambda_{H,i,j}^T \ast t) \quad ; \quad X_{A,i,j}^T \sim \text{Poisson}(\lambda_{A,i,j}^T \ast t) \quad (1)
\]

In the above expression \(t\) stands for \"match time\". This variable measure the percentage of the time\(^9\) elapsed from the beginning of the match and ranges from 0 (the start of the match) to 1 (the end of the match). Then in the middle of a match, \(t\ equals 0.5.

---

\(^8\)In the appendix, results are shown using home and away team goal distribution (figures A.1 and A.2), obtaining similar conclusions.

\(^9\)Given as a fraction of unity.
4.2 The Independence Assumption:

The inclusion of this assumption greatly reduces the complexity of the models and reduces the computation times of the inference process. In the literature, the decision of including the assumption of independence when modelling soccer data is widely debated. On the one hand, some authors include this assumption (Keller (1994), Ridder et al. (1994), Everson & Goldsmith-Pinkham (2008), Suzuki et al. (2010)). On the other hand, other authors suggest the use of more flexible models that allow to model possible dependencies. Maher (1982) suggests the Bivariate Poisson distribution. Dixon & Coles (1997) obtained that the assumption of independence between scores is reasonable (except for the scores 0-0, 1-0, 0-1 and 1-1) and proposed a bivariate inflation distribution. Karlis & Ntzoufras (2000) found evidence in favour of the dependence of the two variables, although in magnitude is very small, and propose a Negative Binomial model.

The "Pearson Chi-square" independence test is then carried out using the final scores of the 380 matches corresponding to the League 2013-2014, in order to identify the existence of possible dependencies and to assess whether the assumption of independence is appropriate. The Pearson Chi-square test is able to check the association between two nominal (categorical) variables. The null hypothesis states that knowing the level of Variable X does not help you predict the level of Variable Y. That is, the variables are independent. At the same time, the alternative hypothesis is that knowing the level of Variable X can help you predict the level of Variable Y. Table 2 presents a two-way contingency table for scores and the "Pearson Chi-square" independence test:

Table 2: Pearson Chi-square test for independence

<table>
<thead>
<tr>
<th>Home_team</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>32</td>
<td>33</td>
<td>19</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>96</td>
</tr>
<tr>
<td>1</td>
<td>41</td>
<td>31</td>
<td>27</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>109</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>26</td>
<td>22</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>85</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>17</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>11</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>137</td>
<td>122</td>
<td>80</td>
<td>27</td>
<td>8</td>
<td>6</td>
<td>380</td>
</tr>
</tbody>
</table>

Pearson chi2(35) = 25.7251  Pr = 0.874

The test finds no evidence against the assumption of independence between both variables (p-value is much greater than 0.05). In view of the results, from now on we apply the Independence Assumption:

\begin{align*}
\text{Assumption 2:} \quad & X_{i,j}^H(t) \perp X_{i,j}^A(t) \quad i \neq j
\end{align*}

(2)
4.3 Additive Stochastic Process Assumption: Composite Poisson.

Additive Property of Poisson Distributions: Let $N_1(t)$ and $N_2(t)$ be two independent Poisson processes with parameters $\lambda_1$ and $\lambda_2$, respectively, and let $N(t) = N_1(t) + N_2(t)$. It follows that $N(t)$ is a Poisson process with parameters $\lambda_1 + \lambda_2$.

Using the additive property, as in (Everson & Goldsmith-Pinkham, 2008), we propose a hypothetical partitioning (Assumption 3) to model both team total scores, $X_{i,j}^H(t)$ and $X_{i,j}^A(t)$, as the sum of two independent Poisson processes. Then we are assuming the following Composite Poisson model:

**Home team (i) goals:**

\[ X_{i,j}^H(t) = O_{i}^H(t) + D_{j}^A(t) \]  \hspace{1cm} (3)

- Offensive parameter (home team): $O_{i}^H(t) \sim \text{Poisson}(\theta_{i}^H \ast t)$
- Defensive parameter (away team): $D_{j}^A(t) \sim \text{Poisson}(\delta_{j}^A \ast t)$

**Away team (j) goals:**

\[ X_{j,i}^A(t) = O_{j}^A(t) + D_{i}^H(t) \]  \hspace{1cm} (4)

- Offensive parameter (away team): $O_{j}^A(t) \sim \text{Poisson}(\theta_{j}^A \ast t)$
- Defensive parameter (home team): $D_{i}^H(t) \sim \text{Poisson}(\delta_{i}^H \ast t)$

Note that it is required, $\theta_{i}^H > 0$, $\delta_{j}^A > 0$, $\theta_{j}^A > 0$ and $\delta_{i}^H > 0$.

We also assume:

\[ O_{i}^H(t) \perp D_{j}^A(t) \]  \hspace{1cm} (5)

\[ O_{j}^A(t) \perp D_{i}^H(t) \]  \hspace{1cm} (6)

Applying Assumption 2, we get:

\[ O_{i}^H(t) \perp D_{j}^A(t) \perp O_{j}^A(t) \perp D_{i}^H(t) \]  \hspace{1cm} (7)

Applying the additive property of Poisson distributions, we get the following equalities, which will greatly facilitate inference:

\[ \lambda_{i,j}^H = \theta_{i}^H + \delta_{j}^A \]  \hspace{1cm} (8)

\[ \lambda_{i,j}^A = \theta_{j}^A + \delta_{i}^H \]  \hspace{1cm} (9)
4.4 Time-homogeneous Poisson Process:

There is strong evidence that the probability of a goal being scored steadily increases over the course of the match, perhaps because of tiredness of players (Dixon & Robinson (1998)). Figure 1 shows the sample cumulative distribution of goals (Liga 2013-2014).


Figure 3: Sample cumulative % of goals by minute (Liga 2013-2014)

Nevertheless, in this study we assume, for simplicity, that the goal score intensity does not depend on the time\textsuperscript{10}.

Assumption 4: \[ \lambda_{i,j}^H(t) = \lambda_{i,j}^H \quad \forall t; \quad \lambda_{i,j}^A(t) = \lambda_{i,j}^A \quad \forall t \] (10)

4.5 Team quality:

In the literature, there is an intense debate about how to model team quality fluctuations. Some authors assume, generally for simplicity, that the team quality is constant over the season. At the same time, other studies allow quality fluctuations. Dixon & Coles (1997) accounting for fluctuations in performance of individual teams. Bueno et al. (2010) also assume that team quality is continually evolving. In our case, we will use both assumptions depending on the models:

Assumption 5. i: Parameters (offensive and defensive) are constant over the season. This assumption is less realistic but simplifies calculations and computation times. We will use this assumption in the models presented in section 5.2.1 (OLS estimation) and 5.2.2 (Hierarchical Bayesian Model with random shocks).

Assumption 5. ii: Parameters (offensive and defensive) change throughout the season. This assumption is more flexible and realistic, because it allows fluctuations in “the quality” of individual teams. This assumption will be used in section 5.2.3 (Hierarchical Bayesian Model with autorregresive shocks).

\textsuperscript{10}A refinement is introduced in Hernandez (2017b) to adjust match time according to historical goal distribution.
5 Methodology. Parameter inference.

In this section we define a methodology to estimate the offensive and defensive capacity (quality indicators) of the different teams using information from prediction markets. Remember that for each team we need to estimate a total of four parameters: home attack, home defense, away attack and away defense:

The inference process consists of two phases. In the first phase, we use information from the betting markets to estimate the parameters \( \lambda' \)s that characterise the Poisson processes in each of the 380 games of the season. Since there are two stochastic processes in each match, one for each team, we need to infer a total of 760 values. It should be noted that in the rest of studies, this phase does not call for any calculation, since historical results (or expert opinions) are directly used as primary information. In the second phase, the information collected in the first phase is used to estimate the offensive and defensive capacity of the teams \( (\theta' \)s and \( \delta' \)s).

5.1 Step 1: Obtaining input data \( (\hat{\lambda}^H_{i,j}, \hat{\lambda}^A_{i,j}) \)

First, it is necessary to calculate the inverse of the betting odds in order to obtain a probability measure. It is easy to check that for any match the sum of the odds of the three possible outcomes (home win, draw or away win) is more than 1, indicating that odds are not fair \(^{11}\). Next, we proceed to normalise the odds, so that the sum is equal to 1 (100%).

At this point we can already estimate the parameters. In this part, we make use of: Assumption 1 (Poisson distribution), Assumption 2 (independence) and Assumption 3 (Time-homogeneous Poisson distribution). We will also need the Probability Mass Function (PMF)\(^ {12}\) for a Poisson distribution, defined as:

\[
P(k \text{ events in the interval}) = \frac{e^{-\lambda} \lambda^k}{k!}
\]

where \( \lambda \) is the average number of events per interval and \( k \) takes values 0, 1, 2...

If we apply this formula to the soccer context, at the beginning of a match the probability of observing a particular final score \((k,m)\) is defined by the following joint mass probability function:

\[
P(\text{team } i \text{ scores } k \text{ goals, team } j \text{ scores } m \text{ goals}) = \frac{e^{-\left(X^H_{i,j}\right)}\left(X^H_{i,j}\right)^k}{k!} \times \frac{e^{-\left(X^A_{i,j}\right)}\left(X^A_{i,j}\right)^m}{m!}
\]

where \( X^H_{i,j} \) and \( X^H_{i,j} \) are the goal scoring parameters of team \( i \) and \( j \), respectively, when playing at team \( i \) home. \( k \) takes values 0, 1, 2... and \( m \) takes values 0, 1, 2...

\(^{11}\)In fact, sample mean is about 1.018

\(^{12}\)Discrete equivalent of a pdf.
Notation: In every moment of a match, let $w$ be the current home team score and $q$ the current away team score. Let $t$ be match time (% of total time elapsed). Let $\Theta^t_H$ be the set of possible final home team scores conditioned to the current score (at time $t$). Analogously, let $\Theta^t_A$ be the set of possible final away team scores conditioned to the current score (at time $t$). For example, if the score at time $t$ is $(w=1, q=2)$, $\Theta^t_H = [1, 2, \ldots, \infty]$ and $\Theta^t_A = [2, 3, \ldots, \infty]$. Let $\hat{\lambda}_{i,j,t}^H$ be the estimated goal scoring parameter of team $i$ against team $j$ in time $t$. Let $\hat{\lambda}_{i,j,t}^A$ be the estimated goal scoring parameter of team $j$ against team $i$ at time $t$. Let $a_{i,j}^t$ and $b_{i,j}^t$ be the Betfair Exchange implied probability of draw and local victory at time $t$, respectively.

Using the implicit probabilities of Betfair Exchange and the PMF formula, we can solve the following implicit system of equations\(^{13}\) for every observation\(^{14}\) of every match:

Equation 1:

\[
0 = \sum_{k \leq \Theta^t_H} \sum_{m \leq k \in \Theta^t_H} e^{-\left(\frac{\hat{\lambda}_{i,j,t}^H}{1-t}\right)(k-w)} \frac{e^{-\left(\frac{\hat{\lambda}_{i,j,t}^H}{1-t}\right)(m-q)}}{(k-w)!} - a_{i,j}^t \quad t \subseteq [0,1] \quad i=j=1,\ldots,20 \tag{13}
\]

Equation 2:

\[
0 = \sum_{k \leq \Theta^t_A} \sum_{m \leq k \in \Theta^t_A} e^{-\left(\frac{\hat{\lambda}_{i,j,t}^A}{1-t}\right)(k-w)} \frac{e^{-\left(\frac{\hat{\lambda}_{i,j,t}^A}{1-t}\right)(m-q)}}{(m-q)!} - b_{i,j}^t \quad t \subseteq [0,1] \quad i=j=1,\ldots,20 \tag{14}
\]

The system is perfectly identified, so every pair $(a_{i,j}^t, b_{i,j}^t)$ identify one and only one pair $(\hat{\lambda}_{i,j,t}^H, \hat{\lambda}_{i,j,t}^A)$. Solving all the systems of equations, we find the pair of unknown $(\hat{\lambda}_{i,j,t}^H, \hat{\lambda}_{i,j,t}^A)$ for every second of every match, so we have a entire distribution for every match and every team (ranging from 5600 to 5800 depending of the match duration).

From $\hat{\lambda}$’s to $\hat{\lambda}$’s: Once we have estimated $(\hat{\lambda}_{i,j,t}^H, \hat{\lambda}_{i,j,t}^A)$, $\forall t \subseteq [0,1]$, we have to find the way to summarize all this information (the entire distribution) into a single value. Three possible methods are considered:

- **Value at the beginning of the match.** A simple option is to use match odds at the beginning of the match. But, with this method we are losing information, because we are not considering the entire distribution.

- **Median value along all the match.** A second option is to use the odds of the entire match and calculate the median value of the parameters. Its main advantage is that it collects information from the entire match. In turn, it has the disadvantage of being somewhat conditioned to the circumstances of the match, such as strategic playing, red cards, injuries, refereeing, etc.

\(^{13}\)We solve the implicit systems of equations using fsolve command in Matlab.

\(^{14}\)The frequency is one observation per second.
• The median value between the beginning of the match and 80th minute (truncated median). This is an intermediate option that excludes the final minutes of the match, in which the strategic factor can be very significant.

In principle, there are reasons to consider that the last option is the most appropriate one, since it presents certain advantages over the others, so it will be the measure used throughout the article. However, all other methods will be used as an indicator of robustness in order to verify that the conclusions are maintained.

Finally, we have the input data ($\hat{\lambda}^H_{i,j}$, $\hat{\lambda}^A_{i,j}$, $i=j=1,...,20$) that we were looking for. In the next stage, we will use this information to estimate the models instead of past scores, FIFA rankings or experts’ opinions, the primary data normally used in the soccer modelling literature.

5.2 Step 2: Estimation Methods (Inference)

The goal of this section is to estimate the four quality parameters: home attack, home defense, away attack and away defense. Several estimation methods are introduced. Firstly, we use classical methods: ordinary least squares and weighted ordinary least squares. Secondly, hierarchical bayesian methods are used. Bayesian inference utilizes prior probability distributions and data to estimate the posterior probability distributions, rather than a point estimate, as in classical models. Below, econometric models are described:

5.2.1 (Weighted) Ordinary Least Squares (OLS)

Econometric model:

Home team: $\hat{\lambda}^H_{i,j} = \theta^H_i + \delta^A_j + u^H_{i,j}$ $i = j = 1, 2, ..., 20; \ i \neq j$ (15)

Away team: $\hat{\lambda}^A_{i,j} = \theta^A_j + \delta^H_i + u^A_{i,j}$ $i = j = 1, 2, ..., 20; \ i \neq j$ (16)

Normalization:

First of all, we have to normalize two parameters, one from the home team equation and the other one from the away team equation. Otherwise, we will have multiple solutions, as shown below:

Imagine we have estimated the eighty parameters without applying normalizations. Then ($\hat{\theta}^H, \hat{\delta}^H, \hat{\theta}^A, \hat{\delta}^A, i=j=1,...,20$) forms a solution to the OLS problem. Let $r$ and $s$ be any real constants. Now we can construct the following parameters:

$$\bar{\theta}^H_i = \hat{\theta}^H_i + r; \quad \bar{\delta}^A_j = \hat{\delta}^A_j - r$$ (17)

$$\bar{\theta}^A_j = \hat{\theta}^A_j + s; \quad \bar{\delta}^H_i = \hat{\delta}^H_i - s$$ (18)
Then, for any real values of \( s \) and \( r \), \((\hat{\theta}_i^H, \hat{\delta}_i^H, \hat{\theta}_i^A, \hat{\delta}_i^A, i=j=1,...,20)\) is also a solution for the problem. So, before estimating the model, we normalize two parameters (one from each equation).

Model identification conditions: Two conditions are required:

Condition 1: \((\text{Number of matches } \times 2) > \text{Number of parameters}\) \(\text{(19)}\)

Then, we need at least four complete rounds. Remember that we have to estimate 78 parameters\(^{15}\) and a round has ten matches. So, it is not possible to estimate the model before fifth round\(^{16}\).

Condition 2: Existence of a link between all teams. \(\text{(20)}\)

Our goal, once the model is estimated, is to have a measure to compare the relative quality of any pair of teams. This relative quality can be revealed directly (from the games between both teams) or indirectly (from the games against common rivals). In any case, there must be a link between the two teams. Formally, let \(\Delta_i\) be a set including the following elements: the rivals that have faced the team \(i\) so far (in the present season), the rivals that have faced team \(i\)’s rivals so far and so on ad infinitum. In order to guarantee model identification, we require that, \(\forall i \neq j\), at least one of the following conclusions is true: sets \(\Delta_i\) and \(\Delta_j\) have at least one element in common or team \(i\) faced team \(j\).

Estimation methods:

i) Ordinary Least Square (OLS) regression. In this case, all the matches weigh exactly the same. We are implicitly assuming Assumption 5.i, i.e, parameters are constant over the season.

ii) Weighted Ordinary Least Square regression. We introduce a exponential discount factor \((\beta)\) in order to give more weight to recent data. This is a simplistic approach to allow fluctuations in performance over the season (Assumption 5.ii). Relative weight (RW) between two different rounds, indexed \(x\) and \(z\), is defined as follows:

\[
RW_{x,z} = \frac{\beta^z}{\beta^x} \tag{21}
\]

When choosing the value of \(\beta\), there are two possibilities. On the one hand, we can take the estimated discount factor from previous studies. Dixon & Coles (1997) propose an exponential weighting function \(\phi\) in order to give more value to recent information. On the other hand, an alternative approach is to introduced an endogenous discount rate in the model specification. This way, the value of \(\beta\) is estimated simultaneously with the other model coefficients.

In our study, we consider that the second approach is more convenient, because there are very few studies that apply weighted ordinary least squares in the context of modelling soccer data, so we have few references.

\(^{15}\)We have twenty teams, four parameters per team and we have to exclude the two normalized parameters.

\(^{16}\)Unless previous season games are used.
5.2.2 Hierarchical Poisson-Bayesian Model with Normal random effects priors and Uniform-Gamma hyperpriors (HBM 1).

Below, a two-level hierarchical bayesian model is proposed to model soccer data. From now on, we refer to this model as HBM 1. Subscript $i$ is for home teams ($i=1,...,20$), subscript $j$ is for away teams ($j=1,...,20$) and subscript $k$ is for matches ($k=1,...,38^{17}$).

Level 1 (Structural model). This part contains the central part of the model.

\begin{align}
X_{i,j}^H \mid \theta_i^H, \delta_j^A & \sim \text{Poisson}((\theta_i^H + \delta_j^A) \ast t) \quad i = j = 1, 2, ..., 20; \quad i \neq j \\
X_{i,j}^A \mid \theta_j^A, \delta_i^H & \sim \text{Poisson}((\theta_j^A + \delta_i^H) \ast t) \quad i = j = 1, 2, ..., 20; \quad i \neq j \\
X_{i,j}^H \mid \theta_i^H, \delta_j^A \perp X_{i,j}^A \mid \theta_j^A, \delta_i^H & \quad i = j = 1, 2, ..., 20; \quad i \neq j
\end{align}

Note that the structural model contain the following four assumptions:

- Poisson Assumption (Assumption 1). Given defensive and offensive parameters, goal scoring is modelled using a Poisson distribution.
- Independence Assumption (Assumption 2). Both goal scoring processes are independent.
- Additive Poisson property (Assumption 3). We use this property to decompose the global processes into two parts: $\lambda_{i,j}^H = \theta_i^H + \delta_j^A$ and $\lambda_{i,j}^A = \theta_j^A + \delta_i^H$
- Time inhomogeneous Poisson process (Assumption 4). Match time is adjusted to take into account historical cumulative goal distribution (by means of F function).

Level 2 (Hierarchical Normal Random Effect priors). In this level, the parameters of the structural model ($\theta$’s and $\delta$’s) are modelled. We are assuming Normal Random Effect priors for these parameters. Doing this, we are implicitly using the ”assumption Parameters are constant over the season” (Section 4.5.1). Shocks are normally distributed\(^{18}\) and have zero mean.

\begin{align}
\log(\theta_i^H) & = \alpha_i + \epsilon_{i,k} \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \\
\log(\theta_j^A) & = \tau_j + \epsilon_{j,k} \quad j = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \\
\log(\delta_i^H) & = \rho_i + \epsilon_{i,k} \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \\
\log(\delta_j^A) & = \omega_j + \epsilon_{j,k} \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \\
\epsilon_{i,k}, \epsilon_{j,k} & \sim \text{Normal}(0, \sigma) \quad \forall \quad 1 < i \leq 20; \quad 1 < j \leq 20; \quad 0 < k < 20
\end{align}

\(^{17}\)Every team plays 19 matches at home and 19 away.
\(^{18}\)We have also considered Gamma distributed random shocks, which are mathematically convenient (the gamma distribution is the conjugate prior distribution for poisson), but may be restrictive.
Hyperpriors distributions. Now we are defining the priors (also known as hyperpriors) for the parameters of the level 2. Following the literature (Best et al., 2011), we use non-informative Uniform-Gamma hyperpriors.

\[ \alpha_i \sim \text{Uniform}(\ln -\infty, +\infty) \quad i = 1, 2, \ldots, 20 \] (30)

\[ \tau_j \sim \text{Uniform}(\ln -\infty, +\infty) \quad j = 1, 2, \ldots, 20 \] (31)

\[ \rho_i \sim \text{Uniform}(\ln -\infty, +\infty) \quad i = 1, 2, \ldots, 20 \] (32)

\[ \omega_j \sim \text{Uniform}(\ln -\infty, +\infty) \quad j = 1, 2, \ldots, 20 \] (33)

\[ \sigma \sim \text{Gamma}(0.01, 0.00001) \] (34)

Parameter inference: Once the model is described, the objective is to estimate the parameters (\(\theta\)'s and \(\delta\)'s). The bayesian estimation process is summarize below:

1. Input Data: We use the \(\hat{\lambda}'s\) we have estimated in Section 5.1 as input data. Then, we can use the following expressions in order to estimate the model:

\[ \hat{\lambda}^H_{i,j} = \theta^H_i + \delta^A_j \] (35)

\[ \hat{\lambda}^A_{i,j} = \theta^A_j + \delta^H_i \] (36)

2. Normalization: As in previous section, before estimating the model we normalize two parameters, one from each equation.

3. Estimation algorithm: Closed form bayesian estimation is not possible in this case. A Gibbs Sampler algorithm is proposed\(^\text{19}\). This algorithm is a variant of Monte Carlo Markov Chain and is widely used in Applied Bayesian Modelling. It has also been used in the context of modelling soccer data (Everson & Goldsmith-Pinkham (2008)). Gibbs sampler has the advantage that is applicable when the joint distribution is not known explicitly or when is really hard to sample from directly.

4. Sofware: JAGS (Just another Gibbs sampler) is a statistical software specifically designed for analysis of Bayesian hierarchical models using Markov Chain Monte Carlo (and its variants). So it is the perfect programming environment to estimate this kind of models. The algorithm was implemented calling JAGS from Matlab using MATJAGS interface.

5. Results: Finally, we obtain the posterior probability distribution of the parameters (\(\hat{\theta}\)'s and \(\hat{\delta}\)'s)). Using this methodology, we are able to estimate the four parameters for each team: home attack, home defense, away attack and away defense.

\(^\text{19}\) Comprehensive information about Gibbs Sampler algorithm is provided in: Casella & George (1992) and Gelman et al. (2014).
5.2.3 Hierarchical Poisson-Bayesian Model with autoregressive priors and Uniform-Gamma hyperpriors distribution (HBM 2).

Below, another two-level hierarchical bayesian model is proposed to model soccer data. From now on, we refer to this model as HBM 2. Unlike the previous model, in this case it is allowed that the parameters change over the season. Now we focus on level 2 and the hyperpriors, given the structural model (level 1) is exactly the same as in Section 5.2.2:

Level 2 (Autoregressive prior distributions)
In this case, temporal dependence of the data is specifically taken into account. Shocks are modeled as first-order autoregressive processes, following the methodology described in (Best et al., 2011). This scheme allows for fluctuations in the parameters over the season (Assumption 5.i).

\[
\theta_{i,k}^H = \alpha_i + \rho \ast (\theta_{i,k}^H - \alpha_{t-1,i}) + \epsilon_{t,i} \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \tag{37}
\]

\[
\theta_{j,k}^A = \nu_j + \rho \ast (\theta_{j,k}^A - \nu_{t-1,j}) + \epsilon_{j,k} \quad j = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \tag{38}
\]

\[
\delta_{i,k}^H = \epsilon_i + \rho \ast (\delta_{i,k}^H - \epsilon_{t-1,i}) + \epsilon_{i,k} \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \tag{39}
\]

\[
\delta_{j,k}^A = \pi_j + \rho \ast (\delta_{j,k}^A - \pi_{t-1,j}) + \epsilon_{j,k} \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \tag{40}
\]

\[
\epsilon_{i,k}, \epsilon_{j,k} \sim Normal(0, \sigma) \quad \forall \quad 1 < i \leq 20; \quad 1 < j \leq 20; \quad 0 < k < 20 \tag{41}
\]

Hyperpriors distributions
Hyperpriors for \(\alpha, \nu, \epsilon\) and \(\pi\) are non-informative. At the same time, we restrict \(\kappa\) values to be between 0 and 1 and we propose a Gamma hyperprior for the parameter \(\sigma\).

\[
\alpha_i \sim Uniform(-inf, +inf) \quad i = 1, 2, ..., 20 \tag{42}
\]

\[
\nu_j \sim Uniform(-inf, +inf) \quad j = 1, 2, ..., 20 \tag{43}
\]

\[
\epsilon_i \sim Uniform(-inf, +inf) \quad i = 1, 2, ..., 20 \tag{44}
\]

\[
\pi_j \sim Uniform(-inf, +inf) \quad j = 1, 2, ..., 20 \tag{45}
\]

\[
\kappa \sim Uniform(0,1) \tag{46}
\]

\[
\sigma \sim Gamma(0.001, 0.001) \tag{47}
\]

Parameter inference: The estimation process is analogous to previous section. Nonetheless, now the output is totally different because we obtain a different posterior probability distribution for each week.
6 Application: Liga 2013-2014

Once we have designed the methodology to model soccer data using prediction markets information, we just apply this approach to Spanish Football League data (Liga 2013-2014) in order to check its validity and its prediction properties.

6.1 Parameter estimates:

Using Betfair Betting Exchange data from the first \(k\) weeks of the season, we generated estimates for each teams parameters. We repeat this process for \(k=5,\ldots, 38\). Remember that we need at least 40 matches (4 weeks) to estimate the models, that is the reason we start at week 5.

As an illustrative example, Table 2 shows OLS estimations using information from the first half of the season (until week 19). Note that OLS method provides point estimates instead of predictive distributions (Bayesian case), computing the "best guess" of an unknown population parameter.

<table>
<thead>
<tr>
<th>Team</th>
<th>OLS estimation</th>
<th>Weighted OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quality index (week by week)</td>
<td>Quality index (week by week)</td>
</tr>
<tr>
<td>Barcelona</td>
<td>2.70</td>
<td>2.30</td>
</tr>
<tr>
<td>Real Madrid</td>
<td>2.63</td>
<td>2.23</td>
</tr>
<tr>
<td>Athletic</td>
<td>1.82</td>
<td>1.58</td>
</tr>
<tr>
<td>Athletic</td>
<td>1.12</td>
<td>1.11</td>
</tr>
<tr>
<td>Villarreal</td>
<td>1.24</td>
<td>1.13</td>
</tr>
<tr>
<td>Sevilla</td>
<td>1.31</td>
<td>1.17</td>
</tr>
<tr>
<td>Real Sociedad</td>
<td>1.10</td>
<td>1.14</td>
</tr>
<tr>
<td>Valencia</td>
<td>1.33</td>
<td>1.20</td>
</tr>
<tr>
<td>Espanyol</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>Getafe</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Granada</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>Mlaga</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>Levante</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>Osasuna</td>
<td>0.75</td>
<td>0.86</td>
</tr>
<tr>
<td>Celta</td>
<td>0.84</td>
<td>0.95</td>
</tr>
<tr>
<td>Almeria</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>Elche</td>
<td>0.77</td>
<td>0.78</td>
</tr>
<tr>
<td>Rayo Vallecano</td>
<td>0.84</td>
<td>0.94</td>
</tr>
<tr>
<td>Valladolid</td>
<td>0.76</td>
<td>0.82</td>
</tr>
<tr>
<td>Betis</td>
<td>0.93</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Figure 4 shows Bayesian posterior distributions for some selected teams’s attack parameters. (Real Madrid, Barcelona, Sevilla and Atletico). Parameters have been estimated using model BHM 1 and Gibbs sampling (100,000 draws). As in the previous case, algorithms only include information from the first half of the season, so it is updated up to week 19.
Figure 5 shows Bayesian posterior distributions for Real Madrid’s attack parameter for the whole season. Parameters have been estimated using model BHM 2 and Gibbs sampling. In this case, temporal dependence of the data is modelled and posterior distributions evolve over the season.

Figure 4: BHM 1: Posterior Probability Distributions for the home attack parameters of Real Madrid, Barcelona, Atletico, Sevilla y Valencia (Updated at Week 19).

Figure 5: BHM 2: Posterior Probability Distributions for the home attack parameter of Real Madrid (from week 1 to week 38).
6.2 Predictions for single matches (short-term forecast):

Once we have estimated the parameters, we are able to predict any single match. Firstly, let’s introduce some notation: let \( P_{i,j}^H \) be the probability of local win (team i), let \( P_{i,j}^D \) be the probability of draw and let \( P_{i,j}^A \) be the probability of away win (team j). Now we can define the set of possible forecasts \( (\Gamma_{i,j}) \) as:

\[
\Gamma_{i,j} = \{(P_{i,j}^H, P_{i,j}^D, P_{i,j}^A) \subseteq [0,1]^3 : (P_{i,j}^H + P_{i,j}^D + P_{i,j}^A) = 1\}
\] (48)

In the OLS framework, we use the point estimates (\( \theta \)’s and \( \delta \)’s) in order to get the predictions. For a given match played by team i and team j, the OLS prediction is computed as follows:

Goal scoring processes:

Home Goal Scoring (team i): \( X_{i,j}^H \sim \text{Poisson}(\hat{\theta}_i^H + \hat{\delta}_j^A) \) (49)

Home Goal Scoring (team j): \( X_{i,j}^A \sim \text{Poisson}(\hat{\theta}_j^A + \hat{\delta}_i^H) \) (50)

Once we know both goal scoring processes, we use the joint probability distribution in order to compute the probability of all possible outcomes for the match. Then we compute match probabilities as follows:

\[
P_{i,j}^H = \sum_{x=1}^{\infty} \sum_{y=0}^{x-1} P(X_{i,j}^H = x) \times P(X_{i,j}^A = y)
\] (51)

\[
P_{i,j}^D = \sum_{x=0}^{\infty} P(X_{i,j}^H = x) \times P(X_{i,j}^A = x)
\] (52)

\[
P_{i,j}^A = \sum_{y=1}^{\infty} \sum_{x=0}^{y-1} P(X_{i,j}^H = x) \times P(X_{i,j}^A = y)
\] (53)

In the Bayesian context, the procedure is a little more complex since we have posterior probability distributions instead of point estimates. Then the posterior sampling procedure is described as follows:

1. We randomly choose one value for each parameter from the posterior probability distributions (\( \hat{\theta}_{i,h}^H, \hat{\delta}_{j,h}^A, \hat{\delta}_{j,h}^A, \hat{\theta}_{i,h}^H \)).
2. We compute: \( \hat{\theta}_{i,h}^H + \hat{\delta}_{j,h}^A \) and \( \hat{\theta}_{j,h}^A + \hat{\delta}_{i,h}^H \).
3. Obtain \( P_{i,j,h}^H, P_{i,j,h}^D \) and \( P_{i,j,h}^A \), as in the OLS case.
4. Repeat steps n times (for h=1,...,n^{20}).
5. Compute the median value:

\(^{20}\)In our study, n=100,000.
Once we have developed a systematic method to compute single match probabilities from predictive markets, we check its performance. In the literature (Everson & Goldsmith-Pinkham, 2008), a normal method used to evaluate the goodness of a prediction is to calculate the DeFinnetti distance (DeFinetti, 1972) which is the Euclidean distance between the point correspondent to the outcome and that one correspondent to the forecast. For instance, consider predicted match probabilities (0.2, 0.3, 0.5). If the actual outcome is (0,1,0), De Finetti measure for that prediction equals 0.78 (0.2^2+0.7^2+0.5^2). Note that if we assume complete uncertainty about the final outcome (1/3, 1/3, 1/3), i.e., equiprobable predictor, De Finetti distance always equals 2/3. The historical-mean predictor is also considered. Now, we can use this values as reference to evaluate our prediction model performance. If our model is not able to win equiprobable and historical-mean predictors, our model is not useful at all.

Below, we carry out an exercise to compare Betfair Exchange predictions and the predictions offered by our models. Firstly, we use our models to compute match probabilities week by week, starting on week 5 and finishing on week 38. So, we have to forecast ten matches per week. When calculating match probabilities for some week, we use information from the beginning of the season until the previous week. We do not use information of the current week in order to be on equal terms with markets. Secondly, we compute implied match probabilities from Betfair Exchange for all matches of the season. The data is collected at the beginning of the game. Finally, we compute the sum of all De Finetti distances week by week for both markets prediction and our models’ predictions. In order to have references we also include De Finetti’s distances for the equiprobable predictor and for the historical-mean predictor.

Results are shown in Table 3. If we consider the whole season, market predictions are a bit better than models’ predictions according to ”De Finetti distance”. Remember that a smaller De Finetti distance indicates better fit. At the same time, all the models perform much better than equiprobable and historical-mean predictors.

However, if we evaluate predictions only from week 5 to week 32, excluding last 6 weeks of the season, models’ predictions are more accurate than market ones, because all of them present a smaller De Finetti measure in the subtotal (J5-J32). In conclusion, our forecasting models perform really well throughout the season and are able to beat betting markets, except in the last weeks.

21There are three possible outcomes: (1,0,0), (0,1,0) and (0,0,1).
22This predictor uses the historical means (from 1998 to 2016) for the match probabilities (.48, .25, .27).
23We calculate the inverse of the odds and later we normalize values to sum 1.
<table>
<thead>
<tr>
<th>Week</th>
<th>OLS</th>
<th>Weighted OLS</th>
<th>BHM 1</th>
<th>BHM 2</th>
<th>Betfair Exchange</th>
<th>Equiprob. Predictor</th>
<th>Hist. Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 5</td>
<td>4.94</td>
<td>5.13</td>
<td>5.15</td>
<td>5.41</td>
<td>6.14</td>
<td>6.67</td>
<td>6.34</td>
</tr>
<tr>
<td>Week 6</td>
<td>2.99</td>
<td>3.05</td>
<td>3.29</td>
<td>3.06</td>
<td>2.88</td>
<td>6.67</td>
<td>5.38</td>
</tr>
<tr>
<td>Week 7</td>
<td>6.62</td>
<td>6.43</td>
<td>6.84</td>
<td>6.62</td>
<td>6.91</td>
<td>6.67</td>
<td>7.06</td>
</tr>
<tr>
<td>Week 8</td>
<td>4.07</td>
<td>4.22</td>
<td>4.13</td>
<td>4.16</td>
<td>4.24</td>
<td>6.67</td>
<td>5.38</td>
</tr>
<tr>
<td>Week 9</td>
<td>8.91</td>
<td>8.29</td>
<td>8.24</td>
<td>8.34</td>
<td>8.21</td>
<td>6.67</td>
<td>7.06</td>
</tr>
<tr>
<td>Week 10</td>
<td>4.15</td>
<td>4.52</td>
<td>5.02</td>
<td>4.87</td>
<td>5.07</td>
<td>6.67</td>
<td>5.15</td>
</tr>
<tr>
<td>Week 12</td>
<td>4.49</td>
<td>4.39</td>
<td>4.48</td>
<td>4.49</td>
<td>4.55</td>
<td>6.67</td>
<td>5.38</td>
</tr>
<tr>
<td>Week 13</td>
<td>3.94</td>
<td>6.66</td>
<td>5.99</td>
<td>5.97</td>
<td>6.02</td>
<td>6.67</td>
<td>7.10</td>
</tr>
<tr>
<td>Week 14</td>
<td>4.19</td>
<td>4.22</td>
<td>4.11</td>
<td>4.20</td>
<td>4.37</td>
<td>6.67</td>
<td>6.16</td>
</tr>
<tr>
<td>Week 15</td>
<td>5.57</td>
<td>5.54</td>
<td>5.65</td>
<td>5.58</td>
<td>5.48</td>
<td>6.67</td>
<td>6.24</td>
</tr>
<tr>
<td>Week 16</td>
<td>5.54</td>
<td>5.60</td>
<td>5.48</td>
<td>5.48</td>
<td>5.72</td>
<td>6.67</td>
<td>5.87</td>
</tr>
<tr>
<td>Week 17</td>
<td>5.09</td>
<td>5.06</td>
<td>5.07</td>
<td>5.17</td>
<td>5.00</td>
<td>6.67</td>
<td>7.02</td>
</tr>
<tr>
<td>Week 18</td>
<td>3.33</td>
<td>3.30</td>
<td>3.48</td>
<td>3.44</td>
<td>3.30</td>
<td>6.67</td>
<td>5.38</td>
</tr>
<tr>
<td>Week 19</td>
<td>5.92</td>
<td>5.82</td>
<td>5.99</td>
<td>5.97</td>
<td>5.75</td>
<td>6.67</td>
<td>6.24</td>
</tr>
<tr>
<td>Week 21</td>
<td>5.05</td>
<td>5.04</td>
<td>5.01</td>
<td>4.94</td>
<td>5.08</td>
<td>6.67</td>
<td>5.79</td>
</tr>
<tr>
<td>Week 22</td>
<td>6.83</td>
<td>6.70</td>
<td>6.83</td>
<td>6.76</td>
<td>6.86</td>
<td>6.67</td>
<td>6.28</td>
</tr>
<tr>
<td>Week 23</td>
<td>5.71</td>
<td>5.73</td>
<td>5.85</td>
<td>5.64</td>
<td>5.67</td>
<td>6.67</td>
<td>5.97</td>
</tr>
<tr>
<td>Week 24</td>
<td>5.79</td>
<td>5.64</td>
<td>5.63</td>
<td>5.66</td>
<td>5.59</td>
<td>6.67</td>
<td>6.05</td>
</tr>
<tr>
<td>Week 25</td>
<td>6.98</td>
<td>6.88</td>
<td>6.90</td>
<td>6.92</td>
<td>6.74</td>
<td>6.67</td>
<td>6.69</td>
</tr>
<tr>
<td>Week 26</td>
<td>5.14</td>
<td>5.19</td>
<td>5.11</td>
<td>5.83</td>
<td>5.84</td>
<td>6.67</td>
<td>5.91</td>
</tr>
<tr>
<td>Week 27</td>
<td>6.80</td>
<td>6.31</td>
<td>6.38</td>
<td>5.36</td>
<td>6.58</td>
<td>6.67</td>
<td>6.23</td>
</tr>
<tr>
<td>Week 28</td>
<td>4.33</td>
<td>4.21</td>
<td>4.43</td>
<td>4.39</td>
<td>4.04</td>
<td>6.67</td>
<td>6.28</td>
</tr>
<tr>
<td>Week 29</td>
<td>5.78</td>
<td>5.80</td>
<td>5.88</td>
<td>5.84</td>
<td>5.90</td>
<td>6.67</td>
<td>6.65</td>
</tr>
<tr>
<td>Week 30</td>
<td>5.34</td>
<td>5.34</td>
<td>5.30</td>
<td>5.31</td>
<td>5.19</td>
<td>6.67</td>
<td>6.24</td>
</tr>
<tr>
<td>Week 31</td>
<td>6.80</td>
<td>6.51</td>
<td>6.12</td>
<td>6.34</td>
<td>6.30</td>
<td>6.67</td>
<td>7.06</td>
</tr>
<tr>
<td>Week 32</td>
<td>3.87</td>
<td>3.89</td>
<td>3.99</td>
<td>3.63</td>
<td>3.89</td>
<td>6.67</td>
<td>5.79</td>
</tr>
<tr>
<td>Subtotal (J5-32)</td>
<td>153.11</td>
<td>152.67</td>
<td>153.82</td>
<td>153.47</td>
<td>153.98</td>
<td>186.67</td>
<td>177.25</td>
</tr>
<tr>
<td>Week 33</td>
<td>5.34</td>
<td>5.27</td>
<td>5.35</td>
<td>5.43</td>
<td>5.11</td>
<td>6.67</td>
<td>6.28</td>
</tr>
<tr>
<td>Week 34</td>
<td>5.45</td>
<td>5.36</td>
<td>5.34</td>
<td>5.35</td>
<td>5.12</td>
<td>6.67</td>
<td>5.87</td>
</tr>
<tr>
<td>Week 35</td>
<td>5.42</td>
<td>5.29</td>
<td>5.45</td>
<td>5.44</td>
<td>4.69</td>
<td>6.67</td>
<td>6.01</td>
</tr>
<tr>
<td>Week 37</td>
<td>9.11</td>
<td>9.07</td>
<td>8.77</td>
<td>8.82</td>
<td>8.85</td>
<td>6.67</td>
<td>6.32</td>
</tr>
<tr>
<td>Week 38</td>
<td>5.51</td>
<td>5.42</td>
<td>5.38</td>
<td>5.62</td>
<td>4.80</td>
<td>6.67</td>
<td>6.24</td>
</tr>
<tr>
<td>Subtotal (J33-J38)</td>
<td>40.67</td>
<td>40.25</td>
<td>39.97</td>
<td>40.14</td>
<td>48.01</td>
<td>40.00</td>
<td>38.46</td>
</tr>
<tr>
<td>Total</td>
<td>193.78</td>
<td>192.92</td>
<td>193.79</td>
<td>193.60</td>
<td>192.01</td>
<td>226.67</td>
<td>215.71</td>
</tr>
</tbody>
</table>
Finally, if we observe De Finetti distances in the last part of the season, it can be concluded that forecasting models perform really bad, since they are even beaten by predictors. This result should not be a matter of concern since last weeks are a little bit weird and are really hard to model. Quality is not as relevant as before and other factors become important, such as motivation, incentives, position in the table or even illegal bonuses for victory or defeat. At the end of the season, some teams do not play anything while other teams have a lot at stake.

An example can be used to illustrate this fact. In the last week of the 2013-2014 Liga, Rayo Vallecano faced Getafe. The OLS model predictions for this match are\(^{24}\): Rayo Vallecano wins the match (50%), draw probability (35%) and Getafe wins the match (15%). However, the implied probabilities of the Betfair Exchange Market were quite different: Rayo Vallecano wins the match (17%), draw probability (22%) and Getafe wins the match (61%). These market probabilities are not consistent with the relative strength of Rayo Vallecano (position 13 in the table) and Getafe (position 18 in the table). But markets have more information and they took into account the different level of motivation between both teams. Rayo Vallecano was in a safe place and it had no incentives to win and Getafe was fighting to not be relegated. Then markets incorporate all this information and update match probabilities according it.

Table 4 shows the mean distance per match. It can be observed that the mean distance deeply increase in the last part of the season, not only for the models considered but also for markets’ predictions. There is therefore evidence that the last matches of the championship are more difficult to forecast in general. However, the increase is lower in Betfair predictions, because of the amount of information available to markets.

<table>
<thead>
<tr>
<th>De Finetti Measure (mean per match)</th>
<th>OLS</th>
<th>Weighted OLS</th>
<th>BHM 1</th>
<th>BHM 2</th>
<th>Betfair Exchange</th>
<th>Equiprob. Predictor</th>
<th>Hist. Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtotal (5-32)</td>
<td>0.547</td>
<td>0.545</td>
<td>0.550</td>
<td>0.548</td>
<td>0.550</td>
<td>0.667</td>
<td>0.633</td>
</tr>
<tr>
<td>Subtotal (33-38)</td>
<td>0.678</td>
<td>0.671</td>
<td>0.666</td>
<td>0.669</td>
<td>0.634</td>
<td>0.667</td>
<td>0.641</td>
</tr>
<tr>
<td>Total</td>
<td>0.570</td>
<td>0.567</td>
<td>0.570</td>
<td>0.569</td>
<td>0.565</td>
<td>0.667</td>
<td>0.634</td>
</tr>
</tbody>
</table>

Table 5 shows the percentage of matches in which the historical-mean predictor loses. These percentages are higher than those obtained in Bueno et al. (2010), i.e., 62.5% of matches, although both results are not directly comparable since their study uses World Cup data instead of Spanish League data.

<table>
<thead>
<tr>
<th>% of matches in which the historical-mean predictor loses</th>
<th>OLS</th>
<th>Weighted OLS</th>
<th>BHM 1</th>
<th>BHM 2</th>
<th>Betfair Exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtotal (5-32)</td>
<td>66.8%</td>
<td>65.0%</td>
<td>65.0%</td>
<td>65.3%</td>
<td>63.6%</td>
</tr>
<tr>
<td>Subtotal (33-38)</td>
<td>56.7%</td>
<td>55.0%</td>
<td>58.3%</td>
<td>55.0%</td>
<td>63.3%</td>
</tr>
<tr>
<td>Total</td>
<td>65.0%</td>
<td>63.53%</td>
<td>63.8%</td>
<td>63.5%</td>
<td>63.5%</td>
</tr>
</tbody>
</table>

\(^{24}\)Weighted OLS and Bayesian predictions are quite similar.
Figure 1 shows correlation between Betfair Exchange and OLS predictions week by week. It can be easily observed that the correlations are really high (close to one) all season except last four weeks. From week 34, correlations deeply decrease, evidencing that the last part of the season is anomalous. Figure 2 repeats the analysis for the Bayesian predictions, observing exactly the same pattern.

Figure 6: Correlation between Betfair Exchange and OLS predictions (week by week).

Figure 7: Correlation between Betfair Exchange and Bayesian predictions (week by week).
6.3 Model Ranking.

Now, we can rank the models according to their short-term prediction performance. Weighted OLS provides the most accurate predictions according to De Finetti measure (0.567 per match on average). BHM 2 ranks in second place (0.569 per match on average), followed by OLS and BHM 1 both with 0.57.

An alternative approach, only valid for bayesian model comparisons, consist on computing the Deviance Information criterion (DIC). DIC is a widely used hierarchical modeling generalization of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The idea is that models with smaller DIC should be preferred to models with larger DIC. DIC is easily calculated from the samples generated by a MCMC simulation. In our case, according to this criterion, the BHM 2 model (DIC=6112) is better than the BHM 1 model (DIC=6576), in the sense that it fits the data better. This results are also consistent with what was observed before, i.e, BHM 2 presents better short-term prediction capabilities.

In view of the results, the differences between models are small. There is slight evidence that Assumption 5.ii (parameters change over the season) seems more reasonable than Assumption 5.i. (parameters are constant), because Weighted OLS is better than OLS and BHM 2 is also better than BHM 1.

6.4 Tournament simulation (long-term forecast):

The forecasting algorithms proposed in this paper can also be applied to tournament simulation. At any date, we can simulate the rest of the season, obtaining predicted points at the end of the season for each team together with a set of simulated variables, such as position, games won, goals scored, probability of winning the league, chances of being relegated, etc.

The simulation procedure is described as follows. Firstly, we collect information of the first half of the season\(^{25}\) (Betfair Exchange data) in order to estimate model parameters for each team (home attack, home defense, away attack and away defense). To do this, we focus on OLS Model. Secondly, we use these estimates to calculate the intensity\(^{26}\) of the goal scoring processes for the remaining matches until the end of the season. To do this, we use the methodology described in Section 6.2. Finally, we run Monte Carlo simulations\(^{27}\), sampling from these distributions. In each simulated league, we have to sample from 760 different Poisson distributions, given there exist two goal scoring processes per match. During a simulation, we apply the Liga official rules for tie breaking\(^{28}\). We repeat this process \(n\) times\(^{29}\) and we get a simulated distribution for each variable.

\(^{25}\)This exercise can be implemented at any point of the season.

\(^{26}\)(\(\hat{\lambda}^\text{H}_{ij}, \hat{\lambda}^\text{A}_{ij}\)), \(i=j=1,...,20\)

\(^{27}\)Simulations are implemented in Matlab.

\(^{28}\)In case of tie between six or more teams, we just break the tie randomly, because they are unlikely and it is really hard to code such situations.

\(^{29}\)In our case, we consider that \(n=100,000\) is enough.
Table 6 presents a summary of the simulation results.\(^{30}\) Note that we are simulating from week 19, approximately five months before the end of the season. Firstly, the medians\(^{31}\) of the simulated distributions of several variables are presented. The list of variables includes: matches won, matches drawn, matches lost, goals scored (GF), goals conceded (GC) and goal difference (GD). In the Appendix (figures A13, A14 and A15), we compare the actual and predicted values of GF, GC and GD. Secondly, we pay special attention to points, so that in addition to the median, it is also included the 5th percentile and the 95th percentile of the simulated distribution. Finally, some probabilities are presented: probability of winning La Liga, probability of qualifying for Champions League (UCL)\(^{32}\), probability of qualifying for Europa League (UEL) and probability of being relegated to second division\(^{33}\).

Table 6: Simulated Season (Liga 2013-2014)

<table>
<thead>
<tr>
<th>Pos.</th>
<th>Team</th>
<th>Played</th>
<th>Won</th>
<th>Drawn</th>
<th>Lost</th>
<th>GF</th>
<th>GC</th>
<th>GD</th>
<th>Median</th>
<th>Pctl 5</th>
<th>Pctl 95</th>
<th>Win Liga</th>
<th>UCL</th>
<th>UEL</th>
<th>Relegated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Barcelona</td>
<td>38</td>
<td>30.9</td>
<td>4.46</td>
<td>2.7</td>
<td>108.8</td>
<td>24.7</td>
<td>79.1</td>
<td>97.0</td>
<td>90</td>
<td>103</td>
<td>66.6%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>2</td>
<td>Real Madrid</td>
<td>38</td>
<td>29.4</td>
<td>4.67</td>
<td>3.9</td>
<td>102.6</td>
<td>35.1</td>
<td>67.5</td>
<td>92.9</td>
<td>85</td>
<td>100</td>
<td>23.2%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>3</td>
<td>Atletico</td>
<td>38</td>
<td>26.2</td>
<td>5.89</td>
<td>3.9</td>
<td>81.5</td>
<td>24.3</td>
<td>57.1</td>
<td>90.4</td>
<td>82</td>
<td>98</td>
<td>10.2%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>4</td>
<td>Athletic</td>
<td>38</td>
<td>18.5</td>
<td>7.92</td>
<td>11.6</td>
<td>55.0</td>
<td>46.1</td>
<td>9.9</td>
<td>63.4</td>
<td>55</td>
<td>72</td>
<td>36.3%</td>
<td>47%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>5</td>
<td>Villarreal</td>
<td>38</td>
<td>18.2</td>
<td>8.59</td>
<td>11.2</td>
<td>61.9</td>
<td>42.3</td>
<td>19.6</td>
<td>63.1</td>
<td>55</td>
<td>72</td>
<td>44.3%</td>
<td>44.3%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>6</td>
<td>Sevilla</td>
<td>38</td>
<td>16.4</td>
<td>10.7</td>
<td>10.9</td>
<td>62.0</td>
<td>51.1</td>
<td>10.9</td>
<td>59.8</td>
<td>51</td>
<td>69</td>
<td>16.0%</td>
<td>46.7%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>7</td>
<td>Real Sociedad</td>
<td>38</td>
<td>16.4</td>
<td>9.76</td>
<td>11.9</td>
<td>59.6</td>
<td>51.1</td>
<td>8.6</td>
<td>58.9</td>
<td>50</td>
<td>68</td>
<td>12.0%</td>
<td>42.2%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>8</td>
<td>Valencia</td>
<td>38</td>
<td>15.4</td>
<td>6.55</td>
<td>16.1</td>
<td>52.2</td>
<td>53.2</td>
<td>-1.0</td>
<td>52.7</td>
<td>44</td>
<td>61</td>
<td>14.3%</td>
<td>14.3%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>9</td>
<td>Espanyol</td>
<td>38</td>
<td>12.2</td>
<td>8.66</td>
<td>17.1</td>
<td>42.0</td>
<td>56.2</td>
<td>-4.2</td>
<td>45.4</td>
<td>37</td>
<td>54</td>
<td>0%</td>
<td>9.9%</td>
<td>3.6%</td>
<td>0%</td>
</tr>
<tr>
<td>10</td>
<td>Granada</td>
<td>38</td>
<td>12.2</td>
<td>6.65</td>
<td>19.1</td>
<td>36.9</td>
<td>52.6</td>
<td>-15.7</td>
<td>43.3</td>
<td>35</td>
<td>52</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6.8%</td>
</tr>
<tr>
<td>11</td>
<td>Malaga</td>
<td>38</td>
<td>11.3</td>
<td>9.83</td>
<td>16.9</td>
<td>39.5</td>
<td>49.4</td>
<td>-9.9</td>
<td>43.6</td>
<td>35</td>
<td>52</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6.9%</td>
</tr>
<tr>
<td>12</td>
<td>Getafe</td>
<td>38</td>
<td>12.2</td>
<td>6.93</td>
<td>18.9</td>
<td>37.6</td>
<td>57.8</td>
<td>-20.3</td>
<td>43.4</td>
<td>35</td>
<td>52</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>7.5%</td>
</tr>
<tr>
<td>13</td>
<td>Levante</td>
<td>38</td>
<td>10.7</td>
<td>9.65</td>
<td>17.6</td>
<td>34.9</td>
<td>55.8</td>
<td>-20.9</td>
<td>41.8</td>
<td>34</td>
<td>50</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>11.8%</td>
</tr>
<tr>
<td>14</td>
<td>Osasuna</td>
<td>38</td>
<td>11.1</td>
<td>7.80</td>
<td>19.1</td>
<td>34.7</td>
<td>57.1</td>
<td>-22.4</td>
<td>41.0</td>
<td>33</td>
<td>50</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>15.4%</td>
</tr>
<tr>
<td>15</td>
<td>Celta</td>
<td>38</td>
<td>10.2</td>
<td>8.76</td>
<td>19.0</td>
<td>41.7</td>
<td>59.2</td>
<td>-17.5</td>
<td>39.4</td>
<td>31</td>
<td>48</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>24.7%</td>
</tr>
<tr>
<td>16</td>
<td>Almeria</td>
<td>38</td>
<td>9.01</td>
<td>8.60</td>
<td>19.5</td>
<td>39.0</td>
<td>67.8</td>
<td>-28.8</td>
<td>38.3</td>
<td>30</td>
<td>47</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>31.5%</td>
</tr>
<tr>
<td>17</td>
<td>Elche</td>
<td>38</td>
<td>8.46</td>
<td>10.6</td>
<td>18.6</td>
<td>33.9</td>
<td>56.4</td>
<td>-22.5</td>
<td>37.2</td>
<td>29</td>
<td>45</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>39.0%</td>
</tr>
<tr>
<td>18</td>
<td>Valladolid</td>
<td>38</td>
<td>8.18</td>
<td>11.9</td>
<td>17.9</td>
<td>38.1</td>
<td>58.9</td>
<td>-20.8</td>
<td>36.4</td>
<td>28</td>
<td>45</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>44.2%</td>
</tr>
<tr>
<td>19</td>
<td>Rayo Vallecano</td>
<td>38</td>
<td>10.2</td>
<td>5.58</td>
<td>22.2</td>
<td>38.2</td>
<td>74.8</td>
<td>-38.6</td>
<td>36.2</td>
<td>28</td>
<td>45</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>47.4%</td>
</tr>
<tr>
<td>20</td>
<td>Betis</td>
<td>38</td>
<td>8.06</td>
<td>9.90</td>
<td>20.0</td>
<td>36.4</td>
<td>63.7</td>
<td>-27.3</td>
<td>34.1</td>
<td>26</td>
<td>41</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>61.9%</td>
</tr>
</tbody>
</table>

---

\(^{30}\) Actual table of LIGA 2013-2014 is presented in the appendix (Table 7)

\(^{31}\) The mean gives similar results.

\(^{32}\) The top four teams are qualified to Champions League

\(^{33}\) The bottom four teams are relegated.
Figure 8 shows the 90% interval for the points at the end of the season. The actual points fall within the bands for 18 out of the 20 teams. Only Barcelona (models overestimate its performance) and Celta (models infraestimate its performance) are outside the interval. At the same time, figure 9 compares the actual points and the median predicted points, observing a good fit in general terms.

Figure 8: 90% confident interval for points.

Figure 9: Points: Actual vs Simulation.
As an example, figure 10 shows a probability distribution of points for Atletico de Madrid. The simulation was made considering only information from the first half of the season. At that point, 91 points was the most likely outcome with a probability of 8.5%, followed by 90 and 92 points, with a probability of 8% and 7.8%, respectively. Atletico de Madrid finally got 90 points at the end of the season.

![Distribution of Points for Atlético de Madrid](image)

Figure 10: Simulated distribution of points for Atletico de Madrid at the end of the season.

7 Conclusions

In this article a market-based algorithm for predicting the outcome of events is proposed and illustrated for the 2013-2014 Liga. The main methodological contribution is the use of high-frequency data from betting exchanges as primary information for the estimation of predictive models. Until now, the models presented in the literature have been based solely on historical results or rankings, but have never used information from competitive markets. When modelling the stochastic processes that determine the goals scored by the different teams, we believe that markets can offer us better quality information that is more up-to-date than the final scores that, after all, are a simple realisation of each stochastic process.

First, in order to facilitate the process of inference, several assumptions are introduced. On the one hand, it is assumed that the goals scored by a certain team follow a time-homogeneous Poisson process. On the other hand, it is considered that the processes that determine the goals of the local and visiting team are statistically independent. These assumptions are widely used in the literature and the tests performed to ascertain their validity find no evidence against them.

Next, predictive market information is used to infer the parameters of the stochastic
processes that determine the goals scored by each team. For each second of each match of the season, a system of equations is proposed that uses as inputs the probabilities provided by the markets and outputs the estimated intensity of the Poisson processes for each of the teams. In this way, for each match we have a whole probability distribution for each of the possible final scores instead of a single result.

Subsequently, the additive property of the Poisson distributions is used to estimate a total of four parameters per team: home attack, home defense, away attack and away defense. The intensities of the Poisson processes previously estimated are used as primary information. Several estimation methods are proposed: (1) (weighted) ordinary least squares, (2) hierarchical bayesian model with normal random shocks and (3) hierarchical bayesian model with autorregresive shocks. Now there is a mechanism by which any two teams can be paired and the outcome of their future matches predicted.

The methodology described above is applied to predict the probability of each Liga 2013-2014 match, calculated week by week, starting on Matchday 5 and ending on Matchday 38. The prediction capability of the proposed models is tested using "De Finetti" measure, which is defined as the Euclidean distance between the prediction and the realisation. The results obtained indicate that the models have good short-term predictive properties, being able to improve market predictions, except in the last weeks of the season.

Finally, we verify the predictive ability of proposed algorithms when simulating a league. In this exercise, information from the betting markets (first half of the season) is used for estimating the structural parameters and then the rest of the season is simulated. Our models have also good long-term predictive properties, since the predictions derived from the simulation closely resemble what is observed in reality.

These models can be easily extended to other sports just changing the structural model and the distributions. This methodology can also be applied to other prediction markets, such as stock markets, political markets...

As a final remark, it can be concluded that the use of information from betting markets can lead to more accurate estimates of the probability distribution of future events.
8 References


Huang, X., Knottenbelt, W., & Bradley, J. (2011). Inferring tennis match progress from in-play betting odds. Final year project), Imperial College London, South Kensington Campus, London, SW7 2AZ.


Keller, J. B. (1994). A characterization of the Poisson distribution and the probability
of winning a game. The American Statistician, 48(4), 294-298.


9 Statistical Appendix

9.1 Appendix A: Tables and figures

Figure A11: Rootograms: Home goals

Figure A12: Rootograms: Away goals
### Table 7: Liga 2013-2014 (Table)

<table>
<thead>
<tr>
<th>Pos.</th>
<th>Team</th>
<th>Played</th>
<th>Won</th>
<th>Drawn</th>
<th>Lost</th>
<th>GF</th>
<th>GC</th>
<th>GD</th>
<th>Points</th>
<th>Win Liga</th>
<th>UEFA CL</th>
<th>UEFA EL</th>
<th>Relegated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Atletico</td>
<td>38</td>
<td>28</td>
<td>6</td>
<td>4</td>
<td>77</td>
<td>26</td>
<td>51</td>
<td>90</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Barcelona</td>
<td>38</td>
<td>27</td>
<td>6</td>
<td>5</td>
<td>100</td>
<td>33</td>
<td>67</td>
<td>87</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Real Madrid</td>
<td>38</td>
<td>27</td>
<td>6</td>
<td>5</td>
<td>104</td>
<td>38</td>
<td>66</td>
<td>87</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Athletic</td>
<td>38</td>
<td>20</td>
<td>10</td>
<td>8</td>
<td>66</td>
<td>49</td>
<td>17</td>
<td>70</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Sevilla</td>
<td>38</td>
<td>18</td>
<td>9</td>
<td>11</td>
<td>69</td>
<td>52</td>
<td>17</td>
<td>63</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>Villarreal</td>
<td>38</td>
<td>17</td>
<td>8</td>
<td>13</td>
<td>60</td>
<td>44</td>
<td>16</td>
<td>59</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Real Sociedad</td>
<td>38</td>
<td>16</td>
<td>11</td>
<td>11</td>
<td>62</td>
<td>55</td>
<td>7</td>
<td>59</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>Valencia</td>
<td>38</td>
<td>13</td>
<td>10</td>
<td>5</td>
<td>51</td>
<td>53</td>
<td>-2</td>
<td>49</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>Celta</td>
<td>38</td>
<td>14</td>
<td>7</td>
<td>17</td>
<td>54</td>
<td>14</td>
<td>40</td>
<td>70</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>Levante</td>
<td>38</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>35</td>
<td>43</td>
<td>8</td>
<td>48</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>Malaga</td>
<td>38</td>
<td>12</td>
<td>9</td>
<td>17</td>
<td>39</td>
<td>46</td>
<td>-7</td>
<td>45</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>12</td>
<td>Rayo Vallecano</td>
<td>38</td>
<td>13</td>
<td>4</td>
<td>21</td>
<td>46</td>
<td>80</td>
<td>-34</td>
<td>43</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>13</td>
<td>Getafe</td>
<td>38</td>
<td>11</td>
<td>9</td>
<td>18</td>
<td>35</td>
<td>54</td>
<td>-19</td>
<td>42</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>14</td>
<td>Espanyol</td>
<td>38</td>
<td>11</td>
<td>9</td>
<td>18</td>
<td>43</td>
<td>51</td>
<td>-10</td>
<td>42</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>15</td>
<td>Granada</td>
<td>38</td>
<td>12</td>
<td>5</td>
<td>21</td>
<td>52</td>
<td>56</td>
<td>-24</td>
<td>41</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>16</td>
<td>Elche</td>
<td>38</td>
<td>9</td>
<td>13</td>
<td>16</td>
<td>31</td>
<td>50</td>
<td>-20</td>
<td>40</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>17</td>
<td>Almeria</td>
<td>38</td>
<td>11</td>
<td>7</td>
<td>20</td>
<td>43</td>
<td>71</td>
<td>-28</td>
<td>40</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>18</td>
<td>Osasuna</td>
<td>38</td>
<td>10</td>
<td>9</td>
<td>19</td>
<td>32</td>
<td>82</td>
<td>-50</td>
<td>39</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>19</td>
<td>Valladolid</td>
<td>38</td>
<td>7</td>
<td>15</td>
<td>16</td>
<td>38</td>
<td>60</td>
<td>-22</td>
<td>36</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>20</td>
<td>Betis</td>
<td>38</td>
<td>6</td>
<td>7</td>
<td>25</td>
<td>36</td>
<td>78</td>
<td>-42</td>
<td>25</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

![Figure A13: Goals Scored (Actual vs Simulation).](image-url)
Figure A14: Goals Conceded (Actual vs Simulation).

Figure A15: Goals Difference (Actual vs Simulation).
9.2 Appendix B: Alternative models

Hierarchical Poisson-Bayesian Model with Gamma random effects priors and Normal-Gamma hyperpriors distribution.

Level 1 (Structural model). It contains the central part of the model.

\[
X_{i,j}^H \mid \theta_i^H, \delta_j^A \sim \text{Poisson}((\theta_i^H + \delta_j^A) \ast F(t)) \quad i = j = 1, 2, ..., 20; \quad \forall i \neq j \tag{57}
\]

\[
X_{i,j}^A \mid \theta_j^A, \delta_i^H \sim \text{Poisson}((\theta_j^A + \delta_i^H) \ast F(t)) \quad i = j = 1, 2, ..., 20; \quad \forall i \neq j \tag{58}
\]

\[
X_{i,j}^{H,A} \mid \theta_i^H, \delta_j^A \perp X_{i,j}^{A,H} \mid \theta_j^A, \delta_i^H \quad i = j = 1, 2, ..., 20; \quad \forall i \neq j \tag{59}
\]

Level 2 (Hierarchical Normal Random Effect priors). In this level, the parameters of the structural model (\(\theta\)'s and \(\delta\)'s) are modelled. We are assuming Gamma Random Effect priors for these parameters. Doing this, we are using the "assumption Parameters are constant over the season" (Section 4.5.1).

\[
\theta_i^H \sim \text{Gamma}(\mu_1, \sigma_1) \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \tag{60}
\]

\[
\theta_j^A \sim \text{Gamma}(\mu_2, \sigma_2) \quad j = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \tag{61}
\]

\[
\delta_i^H \sim \text{Gamma}(\mu_3, \sigma_3) \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \tag{62}
\]

\[
\delta_j^A \sim \text{Gamma}(\mu_4, \sigma_4) \quad i = 1, 2, ..., 20 \quad k = 1, 2, ..., 38 \tag{63}
\]

Hyperpriors distributions. Now we are defining the priors (also known as hyperpriors) for the parameters of the level 2. Following the literature (Best et al., 2011), we use non-informative Normal-Gamma hyperpriors.

\[
\mu_1 \sim \text{Normal}(0, 0.0001) \tag{64}
\]

\[
\mu_2 \sim \text{Normal}(0, 0.0001) \tag{65}
\]

\[
\mu_3 \sim \text{Normal}(0, 0.0001) \tag{66}
\]

\[
\mu_4 \sim \text{Normal}(0, 0.0001) \tag{67}
\]

\[
\sigma_1 \sim \text{Gamma}(0.001, 0.001) \tag{68}
\]

\[
\sigma_2 \sim \text{Gamma}(0.001, 0.001) \tag{69}
\]

\[
\sigma_3 \sim \text{Gamma}(0.001, 0.001) \tag{70}
\]

\[
\sigma_4 \sim \text{Gamma}(0.001, 0.001) \tag{71}
\]