Edward Wheatcroft*

Profiting from overreaction in soccer betting odds

https://doi.org/10.1515/jqas-2019-0009

Abstract: Betting odds are generally considered to represent accurate reflections of the underlying probabilities for the outcomes of sporting events. There are, however, known to be a number of inherent biases such as the favorite-longshot bias in which outsiders are generally priced with poorer value odds than favorites. Using data from European soccer matches, this paper demonstrates the existence of another bias in which the match odds overreact to favorable and unfavorable runs of results. A statistic is defined, called the Combined Odds Distribution (COD) statistic, which measures the performance of a team relative to expectations given their odds over previous matches. Teams that overperform expectations tend to have a high COD statistic and those that underperform tend to have a low COD statistic. Using data from twenty different leagues over twelve seasons, it is shown that teams with a low COD statistic tend to be assigned more generous odds by bookmakers. This can be exploited and a sustained and robust profit can be made. It is suggested that the bias in the odds can be explained in the context of the "hot hand fallacy", in which gamblers overestimate variation in the ability of each team over time.

Keywords: cognitive bias; football betting; football prediction; outcome bias; simulation.

1 Introduction

Betting odds are generally considered to be accurate reflections of the underlying set of probabilities concerning outcomes of sporting events (Spann and Skiera 2009). This is generally explained by the "wisdom of crowds" effect in which the combination of information from a large number of gamblers can yield odds that reflect accurate probabilities (Surowiecki 2005). This is driven by supply and demand for bets on different outcomes of an event. A high volume of bets will cause the bookmaker to reduce their odds on that outcome and increase their odds on other outcomes with the intention of balancing their book and thereby guaranteeing a profit. Combined with the fact that a bookmaker will usually build a profit margin into the odds, giving the bookmaker an inherent advantage, it is difficult for a gambler to make a consistent profit over time.

Despite the fact that betting odds are generally considered to fairly accurately reflect underlying probabilities, the markets have been shown to be biased in a number of ways. The most well known example of this is the so called "favorite longshot bias" in which bookmakers tend to offer better value odds (i.e. a higher expected return) on favorites than on longshots which tend to offer poor value (Cain, Law, and Peel 2000; Shin 2008). Other less well-known examples of biases include the home-underdog bias, in which odds on home teams priced as the underdog tend to offer better value to gamblers than would be expected given the bookmaker's overround (Dare and Dennis 2011), and sentiment bias, in which odds are overly impacted by gamblers' sentiment towards certain teams (Forrest and Simmons 2008).

The aim of this paper is to demonstrate the existence of a bias in soccer betting odds in which odds on teams that have performed above expectations in recent matches tend to offer poorer value than those that have performed below expectations. A statistic is defined called the Combined Odds Distribution (COD) statistic that quantifies the performance of a team over its previous matches, relative to its odds. When teams have tended to perform well relative to the probabilities implied by their odds, they are assigned a high COD statistic whilst those that have performed badly are assigned a low COD statistic. It is shown that the COD statistic can be used to yield a small but robustly profitable betting strategy. An explanation for this bias is proposed in terms of the "hot hand" phenomenon, in which gamblers believe that a player or team is more likely to be successful if they have been successful in previous attempts. The hot hand is the subject of much debate and believed by many to be a fallacy (Gilovich, Vallone, and Tversky 1985). No view is taken in this paper on the *existence* of the "hot hand" in soccer. Instead, it is argued that the effect of the hot hand, if any, is *overestimated* causing gamblers to overestimate the probability of winning a bet.

This paper is organized as follows: Section 2 consists of a discussion of cognitive biases in the context of sports betting markets. In Section 3, background information is given regarding odds formats and implied probabilities

***Corresponding author: Edward Wheatcroft,** London School of Economics and Political Science, Centre for the Analysis of Time Series, Houghton Street, London, United Kingdom of Great Britain and Northern Ireland, e-mail: e.d.wheatcroft@lse.ac.uk. https://orcid.org/0000-0002-7301-0889

Published online May 04, 2020

along with some discussion of cognitive bias related to gambling. In Section 4, the COD statistic is defined and suggested as a measure of the recent performance of a soccer team relative to its odds. In Section 5, the data used to demonstrate the use of the COD statistic are described. Section 6 presents a demonstration of the COD statistic on the 2017/18 Premier League season. In Section 7, the methodology behind the statistical analysis is defined and, in Section 8, the results of the analysis are presented. Section 9 is used for discussion.

2 Biases in sports betting

Humans are known to be subject to a wide variety of biases and these biases often manifest themselves in behavior towards gambling. Perhaps the most well-known example of this is the Gambler's Fallacy in which gamblers wrongly believe that a run of losses in a set of pure chance independent events will be followed by a run of wins or vice-versa (Tversky and Kahneman 1974). In a set of coin tosses, for example, humans have a strong instinct towards believing that a run of heads will be more likely to be followed by a tail, even though those coin tosses are usually independent (Croson and Sundali 2005).

Another cognitive bias known to affect gambling decisions is the hindsight bias in which, given some outcome, a person wrongly believes they are able to explain the events that led to it (Fischhoff and Beyth 1975). Humans are particularly prone to finding non-robust patterns in data and erroneously explaining them using narratives (Winterbottom et al. 2008). It is easy to see how this might happen in the context of betting on sporting events. In football forums and on social media, one can regularly see statements along the lines of "the team started to perform badly when this player stopped playing" or "the team improved when the tactics were changed". Whilst there may be some truth in some of these statements, too much emphasis is often placed on such explanations when randomness often provides a better explanation.

Closely related to the hindsight bias is the outcome bias in which too much weight is placed on the outcome of a decision rather than whether the right decision was made based upon the impact and probability of all possible outcomes. The outcome bias was demonstrated by Baron and Hershey 1988 in a series of experiments in which subjects were asked to rate the quality of thinking of those that made a decision. The subjects tended to rate the quality of the decision making better when the outcome was favourable. Gamblers may be impacted by the outcome bias by rating the performance of a bet purely on the outcome rather than whether the right decision was made to take the bet in the first place.

Biases in sports betting markets have been widely studied. As mentioned in the introduction, the favorite longshot bias describes a phenomenon in which betting odds offer better value (i.e. a lower expected loss) on favorites than on longshots (Cain et al. 2000; Shin 2008). The favorite longshot bias has been demonstrated in a wide variety of sports including soccer (Constantinou and Fenton 2013), handball (Feddersen 2017) and tennis (Abinzano, Muga, and Santamaria 2016). The effect on gamblers of the favorite longshot bias may be enhanced by the counterintuitive finding that bookmakers' odds tend to be better predictors of longshots than favorites (Buhagiar, Cortis, and Newall 2018). A number of explanations have been offered to explain the favorite longshot bias. One suggestion is that humans tend to overestimate small probabilities leading them to be attracted to betting on longshots (Ottaviani and Sørensen 2008) and that bookmakers are forced to reduce their odds to balance their book. Another possible explanation is that bookmakers simply take advantage of the fact that gamblers are willing to bet on longshots on which poor odds are offered (Johnson et al. 2013). Regardless of the explanation, the favorite longshot bias demonstrates that cognitive bias plays an important role in the setting of betting odds.

Another example of bias in betting odds is the so called home-underdog bias. Dare and Dennis found that betting on the home team when they are the underdog yielded better than expected returns given the bookmaker's overround (Dare and Dennis 2011) in the National Football League (NFL). They attribute this to an underestimation of the effects of home advantage, particularly in terms of scoring ability. A later study suggests a different explanation that gamblers prefer to watch, and therefore bet, on the best teams. Those teams will often be favorites when they play away from home and supply and demand will push those prices down and therefore push the home team's odds up (Humphreys, Paul, and Weinbach 2013). Contrary to the home-underdog bias in NFL, evidence of an away-favorite bias, in which home favorites tend to be priced more generously (i.e. are assigned longer odds) than away favorites has been found in European soccer betting markets (Vlastakis, Dotsis, and Markellos 2009). In a later study, on the other hand, Daunhawer, Schoch and Kosub found no evidence for the away-favorite bias in European soccer (Daunhawer, Schoch, and Kosub 2017).

Other evidence of inefficiencies in the sports betting market comes from the existence of arbitrage opportunities. Such opportunities arise when variation in the odds offered by different bookmakers can be exploited by strategically placing bets on different outcomes of an event, thus guaranteeing a profit. Such opportunities have been found to be numerous in European soccer betting (Constantinou and Fenton 2013) with the number of opportunities substantially increasing over time (Gomez-Gonzalez and Del Corral 2018), perhaps due to an increase in the number of bookmakers and a reduction in the average overround. Substantial opportunities for arbitrage have also been demonstrated in a variety of other sports including rugby (Buckle and Huang 2018) and horse racing (Ashiya 2015). Some bookmakers are known to deploy strategies to prevent arbitrage betting such as closing accounts and limiting the stake that can be placed on certain markets (Purdum 2019).

A number of other biases have been identified in sports betting markets. Evidence was found by Forrest and Simmons of a so called sentiment bias. They found that odds on popular Spanish and Scottish soccer teams tended to be more favorable than on less popular teams (Forrest and Simmons 2008). They suggested that, although their results appear to be contrary to what would be expected in a market dominated by supply and demand, it is in fact consistent with bookmakers' profit maximizing behavior in a competitive market. Further evidence of the sentiment bias has been found by Braun and Kvasnicka who showed that, in international soccer, odds were more favorable in the country represented by each team (Braun and Kvasnicka 2013). Evidence of the bias has also been found in American basketball (Feddersen, Humphreys, and Soebbing 2018).

Another bias in sports betting is related to the particular market on which the odds are offered. It was demonstrated by Hassanniakalager and Newall (2018) that overrounds and expected losses are highly dependent on the market with those on, for example, the "correct score" far higher than those on the match outcome market. They used this evidence to argue for better information to be provided to gamblers to judge the risk of a bet.

The phenomenon of interest in this paper is overreaction of sports betting markets to recent information. Overreactions are often attributed to the availability heuristic which conditions humans to overweight information they can more readily remember. This is a well researched phenomenon in stock markets. For example, De Bondt and Thaler 1985 showed that stocks that have been performing badly in recent times tend to perform better in the future than those that have been performing well. This violates the efficient market hypothesis which states that stock prices reflect all available information. Following this, a large body of literature has emerged demonstrating the existence of overreaction in the stock market (Forbes 1996).

The first suggested example of overreaction in sport was presented by Gilovich, Vallone and Tversky (Gilovich et al. 1985) who proposed a "hot hand fallacy" in basketball. They discovered a common belief that a basketball player could have a "hot hand" and were more likely to score if they had been successful with their most recent shot but found no evidence to support this belief. A recent paper by Miller and Sanjurjo (2018) exposed a bias in the reasoning used by Gilovich et al. (1985) and introduced an unbiased version of the methodology, concluding that there is some evidence that the "hot hand" does, in fact, exist.

A number of papers have found evidence of overreaction in betting markets. Camerer found that the betting market for basketball tended to overvalue winning teams and undervalue losing teams (Camerer 1989), showing that gamblers tended to overreact to good and bad runs of form. Further evidence of this was later found in NFL (Badarinathi and Kochman 1994; Tassoni 1996; Vergin 2001). Woodland and Woodland found that gamblers tended to overreact to the performance of a team in the previous season in NBA (Woodland and Woodland 2015a), NFL (Woodland and Woodland 2015b) and Major League Baseball (Woodland and Woodland 2016). Despite the wealth of evidence for overreaction in betting odds for American sports, few studies appear to have investigated this effect in soccer. Choi and Hui found mixed evidence for the overreaction of odds to surprising events in in-play soccer betting markets (Choi and Hui 2014). However, the effect does not appear to have been demonstrated in prematch betting odds and thus, to the author's knowledge, this paper is the first to demonstrate the phenomenon in this context.

3 Background

3.1 Odds-implied probabilities

This paper is concerned with betting odds and the probabilities that can be implied from them. Before proceeding, it is useful to discuss the format of betting odds and how these relate to implied probabilities of match outcomes. Various formats of betting odds are in widespread use in the world with the most popular generally depending on the region. In the context of this paper, it is useful to consider "decimal" (or "European style") odds (to which other formats can easily be converted). Decimal odds simply indicate the multiplier of the stake if that bet turns out to be successful. For example, if a bookmaker offers odds of 3 on an event, a unit stake would generate a return of 3 (a profit of 2 plus return of the stake). Another format, called "fractional" (or "British style") odds, in this case would be stated as "2/1". In this paper, all betting odds are given in decimal format and hereafter will simply be referred to as the "odds". Crucial to this paper is the concept of the "odds-implied" probability. Let the odds for the ith possible event in a book be O_i. The "odds implied" probability is simply the inverse, i.e. $r_i = \frac{1}{O_i}$. For example, if the odds on an event are $O_i = 3$ then $r_i = \frac{1}{3}$. Whilst probabilities over a set of exhaustive events are required to add up to one, this is almost always *not* the case for odds implied probabilities. In fact, due to the bookmaker's overround, the sum of the odds-implied probabilities for an event will usually exceed one. For an event with k possible outcomes, the "overround" is defined as

$$\pi = \left(\sum_{i=1}^{k} \frac{1}{O_i} \right) - 1. \tag{1}$$

As such, to convert the odds-implied probabilities to probabilities that add to one requires some additional methodology. A number of approaches have been proposed with which to do this (Clarke, Kovalchik, and Ingram 2017). In this paper, a simple multiplicative approach is taken in which the probabilities are defined by

$$\tilde{r}_i = \frac{\left(\frac{1}{O_i}\right)}{1 + \pi}. \tag{2}$$

Implied probabilities that have been normalized are hereafter referred to as "normalized odds-implied probabilities" and non-normalized implied probabilities are referred to as "odds-implied probabilities".

4 Defining the combined odds distribution (COD) statistic

In this paper, the COD statistic, which measures the performance relative to expectations of a team in its previous matches, is defined. In soccer leagues, almost universally, three points are awarded for a win, one for a draw and zero for a defeat. The aim of the statistic is to assess how favorably a team's actual point total after a given number of matches compares to the statistical distribution of possible point totals in a world in which the probability of each match outcome is defined by the normalized odds-implied probabilities, i.e. under the assumption that those probabilities are correct.

Whilst each match can only be played in the real world once, if the probability of each outcome in each match is assumed to be that implied by the odds, the matches can be simulated a large number of times using a random number generator (the simulated match ends with a home win, a draw or an away win with the estimated probability of each one of those outcomes). The relative frequencies of each possible number of points in the simulated world can then be used to estimate the underlying distribution of points a team would expect to achieve under the assumption that the normalized odds-implied probabilities are correct. The COD statistic after a given number of matches is then defined as the quantile of the actual point total within the distribution of simulated point totals. Consequently, if a team's actual point total exceeds most of its simulated point totals, it is given a high COD statistic whilst, if its actual point total is lower than most simulated totals, it is assigned a low COD statistic.

Formally, let x_i be the points obtained by a team in their ith match of a league season. The number of points a team playing its *Nth* match of the season has achieved in its last r matches is given by

$$P_{N,r} = \sum_{i=(N-r)}^{N-1} x_i. \tag{3}$$

Let p_i^w, p_i^d and p_i^l be the estimated probability of a win, draw or defeat, respectively for the ith match of the season for that team. The number of points achieved by that team in the ith match from the jth simulation is calculated using the following rule:

$$s_i^j = \begin{cases} 3, & \text{with probability } p_i^w \\ 1, & \text{with probability } p_i^d \\ 0, & \text{with probability } p_i^l \end{cases} \tag{4}$$

The jth simulated point total over the last r matches can therefore be calculated as

$$S_{N,r}^j = \sum_{i=(N-r)}^{N-1} s_i^j. \tag{5}$$

For a window length of r, the COD statistic after N matches is then defined by

$$\phi_{N,r} = \frac{\sum_{j=1}^{m} f(S_{N,r}^j, P_{N,r})}{m} \tag{6}$$

where

$$f(a,b) = \begin{cases} 1, & \text{if } a < b \\ 0.5, & \text{if } a = b \\ 0, & \text{if } a > b \end{cases} \tag{7}$$

and m is the number of simulations. Throughout this paper, a sample size of $m = 512$ is used.

The question of how to interpret the COD statistic is now addressed. Suppose that the normalized odds-implied probabilities represent actual probabilities of each match. If this is the case, the outcome of each match is a random draw from the distribution of probabilities assigned to each match outcome (win, lose or draw). Under this assumption, the COD statistic then purely reflects random chance. Now suppose that the assumption that the normalized odds-implied probabilities represent the true probabilities of each match outcome is dropped. In this case, the COD statistic may be reflective of either random chance, as above, or by mispricing of the odds and therefore inaccuracies of the normalized odds-implied probabilities. These factors, of course, may occur concurrently. The key finding of this paper is that gamblers tend to overreact and attribute good or bad form to mispricing of odds and that this is reflected in future odds. It is found that the odds on teams that have a high COD statistic tend to offer poor value whilst odds on those with a low COD statistic tend to offer relatively good value. Since betting odds are governed by supply and demand in the market, this is likely caused by biases in the gamblers in which the role of random chance is underestimated.

The COD statistic depends on the chosen value of r, that is the number of previous matches considered. Unless otherwise stated, r is set to $N - 1$ such that all previous matches in the current season are taken into consideration. The results are then briefly compared with those obtained from using a variety of different values of r.

5 Data

This paper makes use of the repository of soccer match data available at www.football-data.co.uk. Free-to-access match-by-match data is supplied in comma separated format for a range of European Leagues dating as far back as the 1993/1994 season. From the 2005/06 season onwards, odds from various bookmakers have been provided concerning the outcome of the match (home win, away win or draw) and these are utilized in this paper up to the end of the 2017/18 season. The leagues considered in this paper along with the number of matches available in each are summarized in Table 1. The total number of soccer matches considered is thus 136,011. Cup games, playoffs and other extra matches during the regular season are not recorded and thus are not considered in this paper.

Table 1: Soccer league data used in this paper.

League	Number of matches
English Premier League	8360
English Championship	12,144
English League One	11,567
English League Two	11,567
English National League	5924
Scottish Premier League	5924
Scottish Championship	3154
Scottish League One	3155
Scottish League Two	3148
Spanish Primera Liga	7578
Spanish Segunda Division	8316
Italian Serie A	7284
Italian Serie B	8160
French Ligue One	4524
French Ligue Two	6840
German Bundesliga	7190
German 2. Bundesliga	5364
Belgian First Division A	4850
Dutch Eredivisie	5508
Portugese Primeira Liga	4980

6 Demonstration – 2017/18 premier league season

The COD statistic introduced in this paper is now demonstrated using data from the 2017/18 English Premier League season. As in previous seasons, the league consisted of twenty teams, each playing one another exactly once home and away such that each team played 38 matches in total over the season. The team finishing top of the league would be crowned champions whilst the teams finishing in 18th, 19th and 20th places would be relegated to the league below. The season was characterized by a particularly strong performance from eventual champions Manchester City who reached a record breaking 100 points, finishing 19 points clear of second place. Stoke City, Swansea City and West Bromwich Albion were all relegated having all finished in the bottom three of the table.

Although the bookmakers had installed Manchester City as preseason favorites, few expected them to dominate the league to such an extent. The COD statistic can be used to assess to what extent Manchester City's performance was expected. Their point total over the season is shown in black in the top panel of Figure 1 along with simulated point totals (grey lines). The corresponding COD statistic over time is shown in the lower panel. Here, it is clear that Manchester City tended to outperform their odds over the season and thus their COD statistic

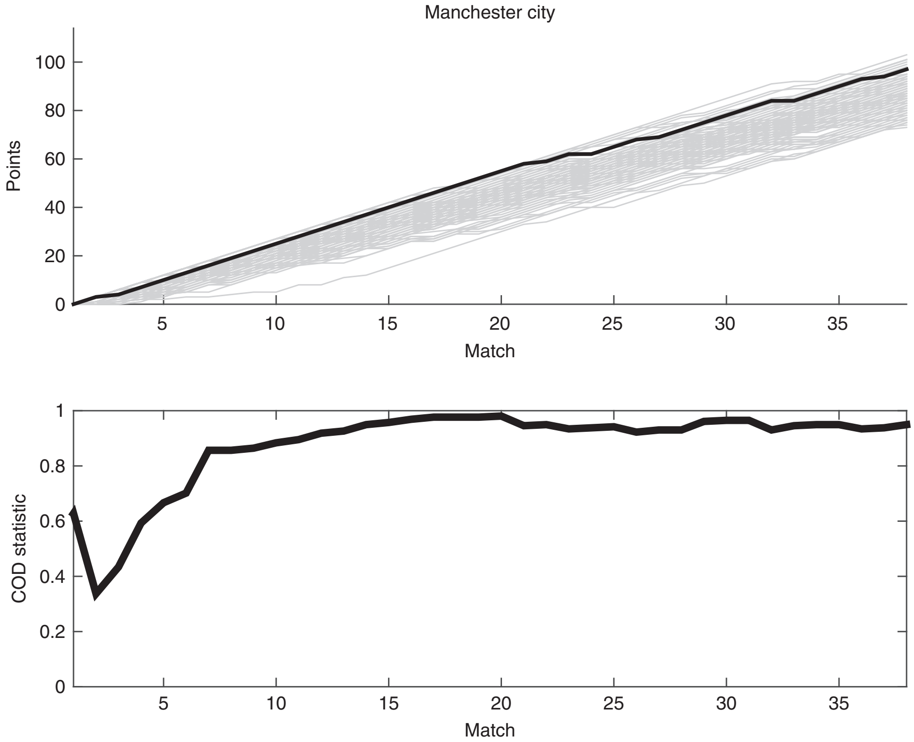

Figure 1: Top: Evolution of 128 simulated point totals (grey) and the actual point total (black) of Manchester City over the 2017/18 Premier League season. Bottom: Manchester City's COD statistic over the season.

is consistently high. If the assumption that the normalized odds-implied probabilities represent actual probabilities of each match holds, some of the explanation behind their extraordinarily successful season can be put down to chance.

To demonstrate the COD statistic further, the performance of a team that performed roughly to expectations over the season is now considered. Everton, whilst far from being one of the title favorites, have consistently been one of the top teams outside of the "big six" teams in the league, regularly finishing in the top half of the table. The 2017/18 season was no exception to this with Everton finishing eighth with 49 points. The actual and simulated point totals of Everton along with their COD statistic are shown in Figure 2. Here, Everton tended to perform roughly as expected, generally falling in the middle of the distribution and generally having a COD statistic not representing a particularly high or low degree of luck (again, under the assumption that the normalized odds-implied probabilities are correct).

A team that had an unexpectedly successful season was Burnley. In preseason, they were made second favorites to be relegated (talkSPORT 2017) and were widely expected to struggle. In the end, they finished in seventh place, qualifying for a place in the Europa League, a Europe-wide competition for high performing clubs who fail to qualify for the more prestigious Champions League. Burnley's actual and simulated point totals and COD statistic are shown in Figure 3. Here, the COD statistic reflects the unexpected nature of Burnley's success showing how their actual point total fell towards the top of the distribution of simulated point totals and therefore how their COD statistic tended to be high throughout the season.

A team that performed somewhat below expectations for the season were Southampton. In preseason, they were priced as 13th favorites to be relegated with decimal odds of 26.0 implying a probability of $\frac{1}{26}$. In practice, the season was one of struggle with an eventual finish of 17th place, just one place and three points above the relegation zone. The actual and simulated point totals of Southampton and their COD statistic are shown in Figure 4. Here, Southampton's poor performance is reflected in their COD statistic which stayed close to zero for most of the season.

The COD statistic of each team in the league is shown as a function of the number of matches played in Figure 5. Here, the top panel shows teams that finished in the top half of the table and the bottom panel teams in the bottom half. Having a low COD statistic is not necessarily reflective of achieving a low number of points or vice versa. For example, both Chelsea and Arsenal ended with low COD

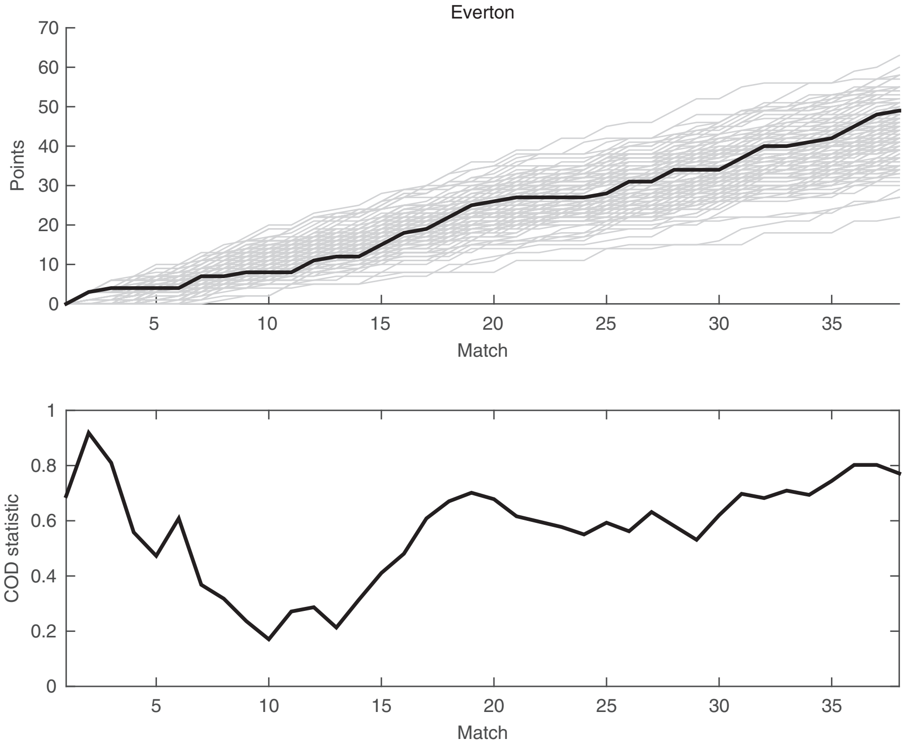

Figure 2: Top: Evolution of 128 simulated point totals (grey) and the actual point total (black) of Everton over the 2017/18 Premier League season. Bottom: Everton’s COD statistic over the season.

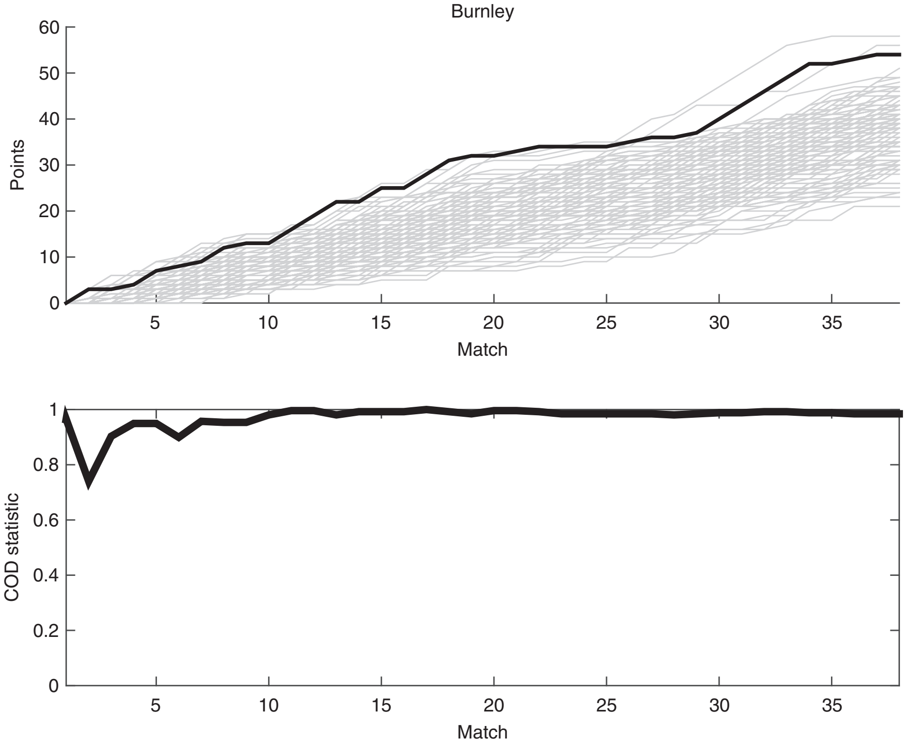

Figure 3: Top: Evolution of 128 simulated point totals (grey) and the actual point total (black) of Burnley over the 2017/18 Premier League season. Bottom: Burnley’s COD statistic over the season.

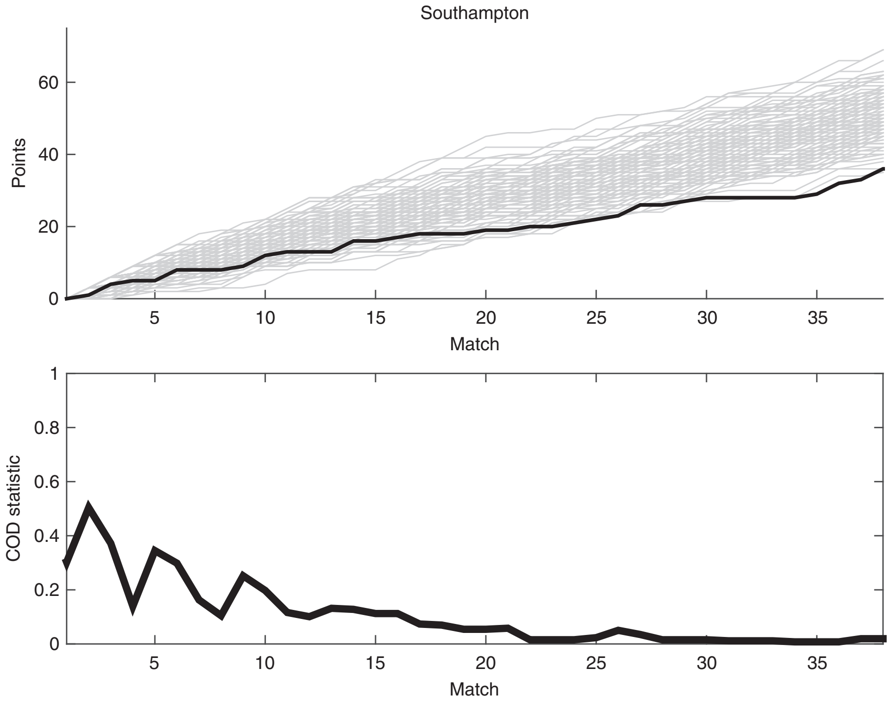

Figure 4: Top: Evolution of 128 simulated point totals (grey) and the actual point total (black) of Southampton over the 2017/18 Premier League season. Bottom: Southampton's COD statistic over the season.

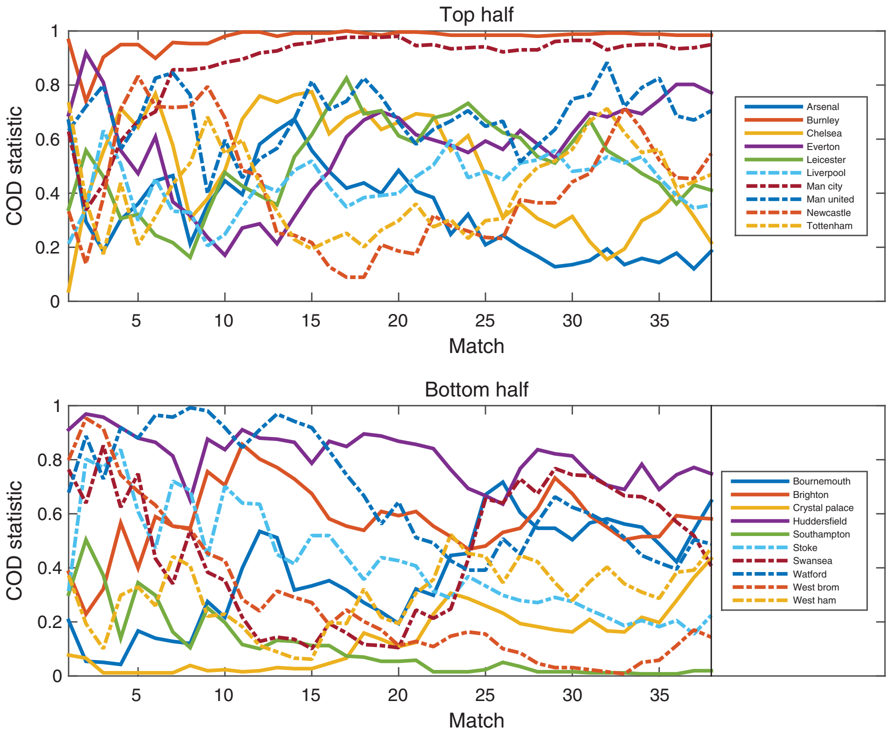

Figure 5: Evolution of the COD statistic for Premier league teams finishing in the top (top panel) and bottom (bottom panel) half of the table in the 2017/18 season.

statistics but finished fifth and sixth in the league, respectively. The low COD statistic was reflective, however, of a somewhat below par performance for the season. Burnley, on the other hand, achieved a high COD statistic yet finished below both Chelsea and Arsenal. The COD statistic, however, clearly does have some impact on league position and this can be seen in the fact that teams in the top half of the table tended to have higher COD statistics than those in the lower half.

7 Methodology

In the analysis presented in this paper, the effect of the COD statistic on the odds of each match outcome is assessed. This is done using data from the twenty leagues listed in Section 5. For each team, when applicable, the COD statistic is calculated over all of its previous matches in that season (that is, using $\phi_{N,N-1}$ for the Nth match of the season) using the maximum betting odds over all available bookmakers.

When fewer than six previous matches have been played by a team in a particular season, due to the small sample size, the effect of the COD statistic is left out of any further analysis. Similarly, occasionally, for certain matches, no odds are available. When calculation of a COD statistic requires this information, these matches are left out of further analysis. The calculations allow a COD statistic to be assigned both to the home and away teams for each match. The effect of the COD statistic on match odds is investigated by assessing the profit and loss that would have been made by betting on teams with different values of the COD statistic as well as different combinations of both the home and away COD statistics.

8 Results

The results of the analysis are presented in terms of the mean profit/loss that would have been made by only betting on teams with COD statistics within some range. In addition, the effect of different combinations of the home and away COD statistics on profit/loss is assessed. A moving average approach is taken in which the returns over the entire data set are ordered by the COD statistic of the home team. A long vector is created from the returns that would have been achieved by betting on the home team in each available and relevant match. This vector is ordered by the COD statistic of the home team such that the first elements of the vector correspond to returns in matches in which the home team has a low COD statistic and the last elements of the vector correspond to returns in matches in which the home team has a high COD statistic. A moving average of the ordered vector of returns is calculated using a window length of ten thousand matches. Therefore, the first element of the moving average corresponds to the average return over the ten thousand matches with the lowest home COD statistic and the last element the average return over the ten thousand matches with the highest COD statistic. A large window is chosen to ensure that the results are robust. Ninety percent bootstrap resampling intervals of the mean profit/loss are formed using 256 random resamples of the ten thousand matches in each window. The whole process described above is also repeated for away teams such that the profit achieved by betting on the away team is ordered by the away team's COD statistic.

The results are shown in Figure 6 in which the mean profit and bootstrapped intervals are shown with the black line and grey area, respectively for home (top) and away (bottom) teams. The dashed line shows the moving average of the overround corresponding to the same window. These results show that better value odds, on average, are available for teams with a low COD statistic than those with a high COD statistic. In fact, the difference is so large that, for the teams with the very lowest COD statistics, since the entire resampling range falls above the zero line, there is significant evidence that a profit can be made both for home and away teams. This is despite the fact that the overround is positive, on average. The fact that the moving average of the overround stays relatively constant rules out the possibility that the difference in profits is due to differences in the overrounds of the bookmakers when teams with different COD statistics are involved.

Another factor that is now ruled out in terms of explaining the difference in profit resulting from betting on teams with different COD statistics is that of the favorite longshot bias. The favorite long shot bias, as described in the introduction, is the tendency for odds on favorites to offer better value than odds on long shots. A moving average of the profit/loss ordered by the odds-implied probabilities rather than the COD statistic is shown in Figure 7 for home and away teams along with 90 percent bootstrapped intervals. Here, there appears to be some limited evidence that betting on teams with higher implied probabilities (i.e. favorites) results in a better return than betting on teams with lower implied probabilities (i.e. long shots). The effect, however, is much weaker than that of the COD statistic. To rule this effect out in terms of explaining the relationship between the gambling return and the COD statistic, however, the relationship between the COD statistic and the odds-implied probabilities needs to be assessed. The moving average of the odds-implied probability as a function of the moving average of the COD statistic when ordered according to the COD statistic is shown in Figure 8 both for home and away teams. Here, there is clearly a relationship between the COD statistic and the odds-implied probability with teams that have high COD

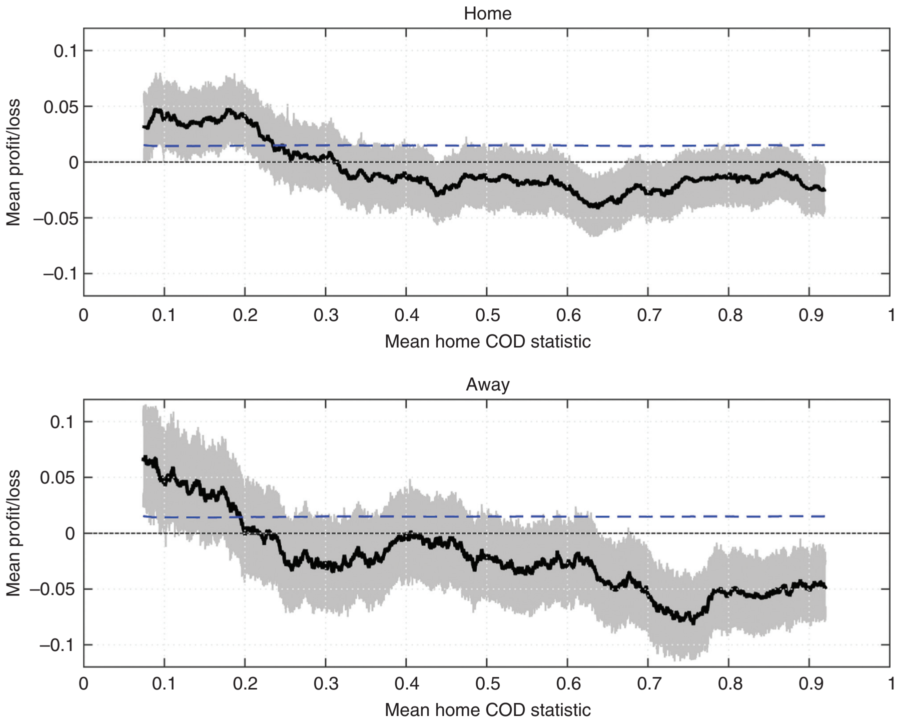

Figure 6: Top: Moving average of the profit made by betting on the home team when ordered from the lowest home COD statistic to the highest. The grey area represents 90 percent bootstrap resampling intervals of the mean profit and the dashed line shows the corresponding moving average of the overround. Bottom: the same but for betting on the away team and ordered by the away COD statistic.

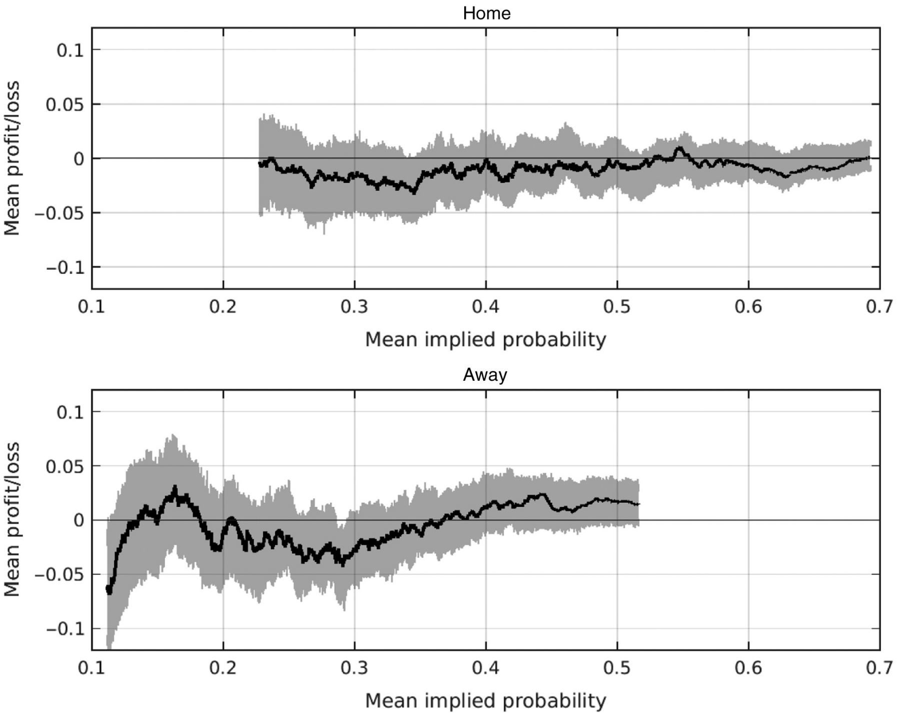

Figure 7: Top: Moving average of the profit made by betting on the home team when ordered from the lowest home odds-implied probability to the highest. The grey area represents 90 percent bootstrap resampling intervals of the mean profit and the dashed line shows the corresponding moving average of the overround. Bottom: the same but for betting on the away team and ordered by the away odds-implied probability.

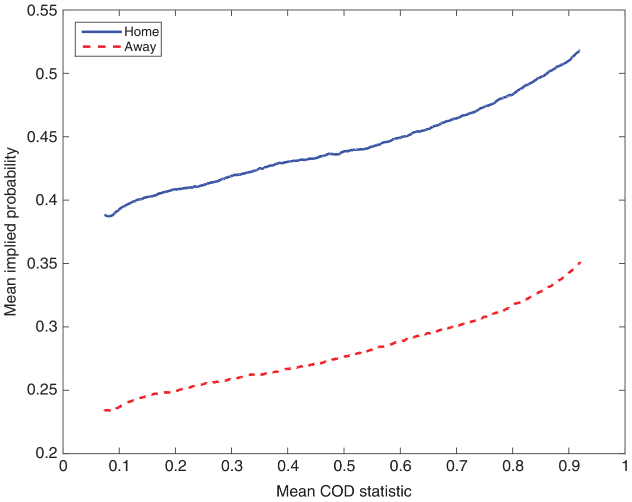

Figure 8: Moving average of the odds-implied probability as a function of the moving average of the COD statistic when ordered by the COD statistic for home (solid line) and away (dashed line) teams.

statistics more likely to have high odds-implied probabilities. This is not surprising since bookmakers' odds will tend to react to good runs of form by reducing the odds on those teams. However, if anything, this relationship is likely to dampen the effect seen in Figure 6 since teams with a high odds-implied probability would be expected to have better value odds than those with a low odds-implied probability due to the favorite long shot bias. Therefore, teams with a low COD statistic tend to have longer odds, on average, which tend to offer poorer value. It is therefore notable that, despite the impact of the favorite long shot bias, there is still a clear relationship in that low COD statistics tend to result in high gambling returns.

In order to demonstrate that the profit/loss made by betting on teams with different COD statistics is fairly consistent over time, the cumulative profit as a function of time from placing bets on teams with COD statistics in each decile is shown in Figure 9 for home (upper panel) and away (lower panel) teams. Here, the thickness and the color of the lines indicates the decile of the COD statistic with thicker lines and colder colors indicating a lower decile of COD statistic and thinner lines and warmer colors indicating a higher decile. Consistently with Figure 6, the lower the COD statistic, the higher the return from betting on those teams. Betting on the lowest three deciles for home teams and the lowest two for away teams would have resulted in a profit over the duration of the data set. Here, there appears to be little evidence that the bias in the odds has increased or decreased over time.

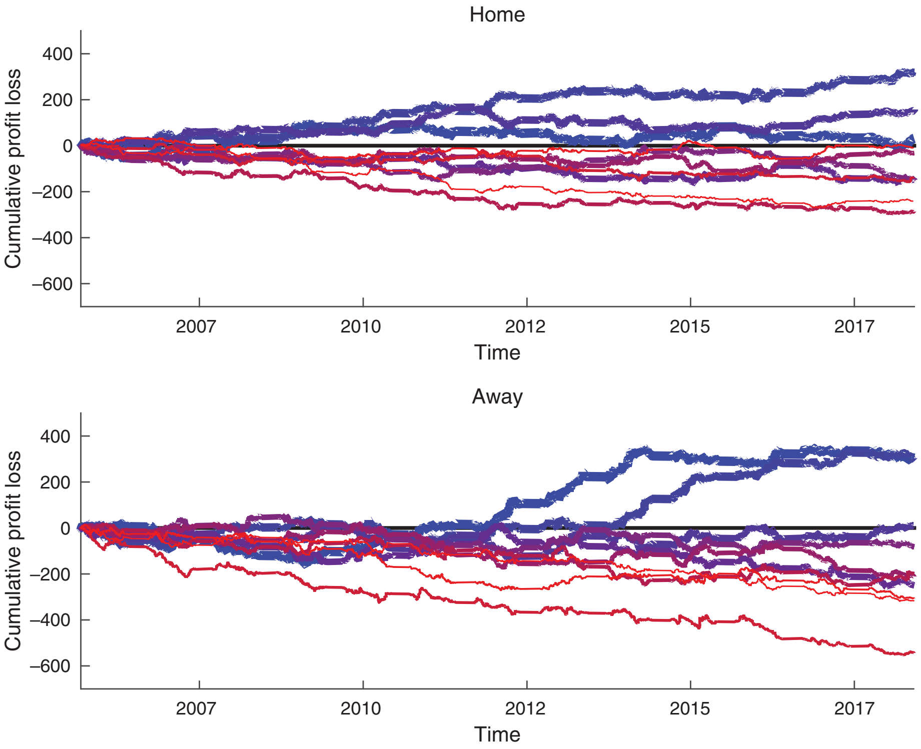

Figure 9: Profit over time from betting on home teams within each decile of the home COD statistic (upper panel) and on away teams within each decile of the away COD statistic (lower panel). Thicker blue lines correspond to higher deciles of the COD statistic whilst thinner red lines correspond to lower deciles.

8.1 Regression analysis

The analysis above demonstrates the effect of the home team's COD statistic on the home team's odds and of the away team's COD statistic on the away team's odds. However, it seems likely that the COD statistic of the away team should also have an effect on the home team's odds and vice-versa. In this section, the effect of both teams' COD statistics on each team's odds are investigated. In order to determine the effect of the home and away COD statistics on the probability of a home win, the results of a logistic regression analysis are now presented in which combinations of the odds-implied probability and the home and away COD statistics are used as explanatory variables for predictions of the binary outcome of whether the home team wins or not. This is then repeated for predictions of away wins. The results are shown in Table 2 for the prediction of home wins and in Table 3 for the prediction of away wins. The performance of each of the combinations of variables is compared using Akaike's Information Criterion (AIC). For clarity, both statistics are given relative to that of the best performing model in each case such that the best performing model has a relative AIC of zero. All coefficients for all combinations of variables are found to be significant at the 0.1 percent level.

Table 2: AIC relative to the best model for each combination of variables as predictors of a home win, along with the values of the logistic regression coefficients.

Implied prob.	COD statistic home	COD statistic away	Δ AIC
+4.721	−0.166	+0.136	0
+4.621	−0.156		+18.9
+4.612		+0.123	+29.3
+4.538			+44.5
	+0.476	−0.501	+5744.5
		+0.516	+6048.6
	+0.483		+6093.8

All variables are found to be strongly significant ($p < 0.001$) for all combinations of variables.

Table 3: AIC relative to the best model for each combination of variables as predictors of an away win, along with the values of the logistic regression coefficients.

Implied prob.	COD statistic away	COD statistic home	Δ AIC
+5.233	−0.171	+0.086	0
+5.179	−0.165		+4.9
+5.113		+0.074	+25.7
+5.071			+28.8
	+0.502	−0.527	+5209.9
		−0.534	+5489.9
	+0.510		+5516.9

All variables are found to be strongly significant ($p < 0.001$) for all combinations of variables.

For prediction of both home and away win probabilities, the combination of variables that provides the best predictor of the outcome includes all three of the candidate predictor variables, suggesting that the COD statistic of both teams has an effect on the probability of a home or away win. Since the odds-implied probabilities and the COD statistics are on the same range, that is on (0, 1), the coefficients in the logistic regression are roughly comparable. Unsurprisingly, the odds-implied probabilities have the highest magnitude coefficients and therefore make the biggest contribution to the final forecast. Consistently with the results shown in the previous section, when predicting the probability of a home win, a negative coefficient is found on the home COD statistic and a positive coefficient is found for the COD statistic of the away team. Similar results are found for the prediction of away wins. The results presented here shows how the COD statistic of both the home and away team impacts the bias in the odds.

The combined effect of the home and away COD statistics is now demonstrated in terms of the profit made from betting on matches with different combinations of this statistic. In Figure 10, a scatter plot is shown of the home COD statistic against the away COD statistic for each match, colored according to the mean profit over that match and matches with similar combinations of these statistics. In order to determine the similarity of the statistics, for a given point, the Euclidean distance is calculated to each of the other points and the profit/loss from the nearest $n - 1$ points and the point itself are averaged over, resulting in an average over a total of n points. Each panel in the figure corresponds to a different value of n. The darkest blue points correspond to the biggest mean profits whilst the darkest red correspond to the biggest losses. Consistent with the results of the logistic regression above, there is a tendency for the biggest profits to come from betting on points in the top left of the figures, that is matches in which the home COD statistic is low and the away COD statistic is high. The impact of the home team's COD statistic appears to be larger than that of the away team on the profit/loss from placing bets on the home team. Note how the colors of the points are impacted by the value of n. For the top left panel case, in which $n = 5000$, the colors are very pronounced because the points averaged over are relatively close. For the bottom right case, in which $n = 40{,}000$, the colors are less pronounced than in the other cases. This is because the effect of the COD statistic is reduced because a wider range of points are considered. There is, however, a clear pattern in this panel,

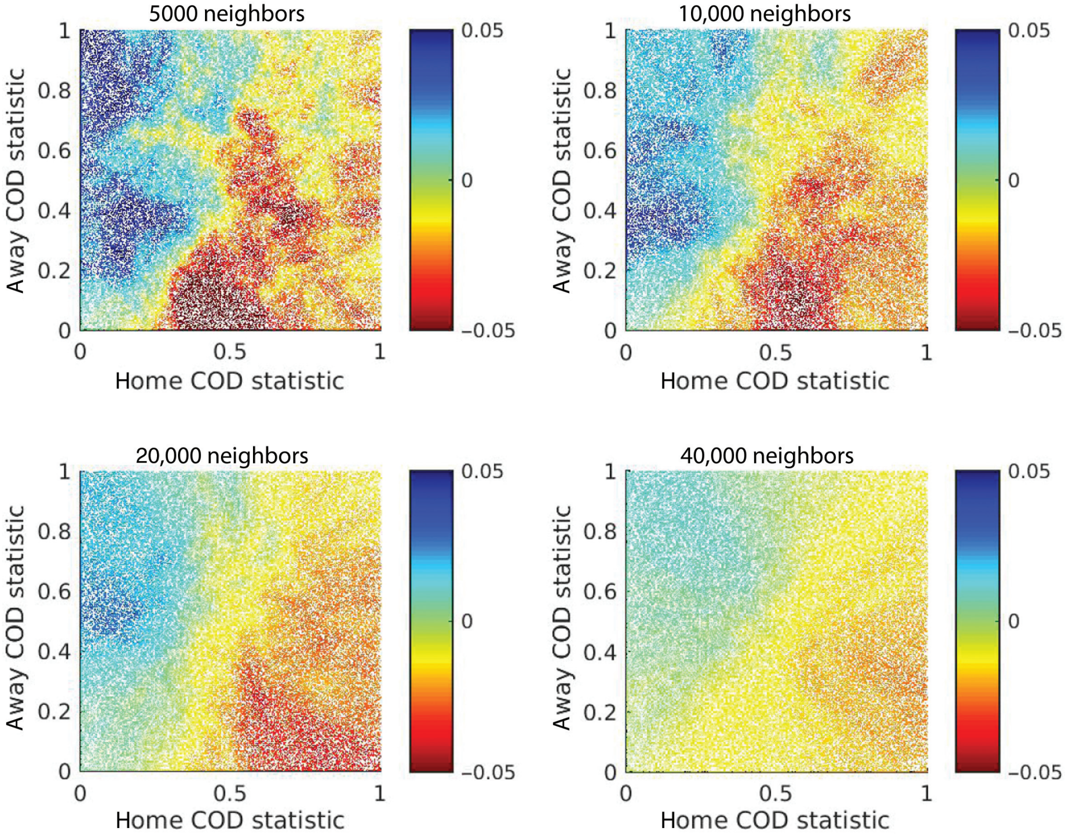

Figure 10: Scatter plot of the home and away COD statistics for each match colored according to the mean profit made by betting on the home team over the number of neighbors stated.

and in all panels, with points in the top left corresponding to a higher mean profit than those in the bottom right. The same scatter plot but for profit/loss from away wins is shown in Figure 11. Here, the results are very similar and consistent with the results of Figure 10. These results clearly demonstrate how the COD statistics of both teams have an impact on the match odds offered on both the home and away teams.

8.2 Effect of match history length

So far in this paper, the COD statistic has been calculated on the basis of all matches played by a team so far in the season, that is the COD statistic used is $\phi_{N,N-1}$ for each value of N. The reasoning behind this is that gamblers may be affected by discrepancies between the state of the league table and their expectations. This means that the number of previous matches considered increases as the season progresses. In this section, this approach is contrasted with the case in which r, the number of previous matches considered, is fixed. Here, the profit/loss achieved when setting r to different values from 4 to 20 is compared with the case in which $r = N - 1$. In order to make a fair comparison, only matches in which the COD statistic can be calculated for all values of r are considered, that is all matches after the 20th match of the season for each team.

In Figure 12, the mean profit/loss achieved from betting on home teams with a COD statistic below α (blue) or above $1 - \alpha$ (red) is shown for $\alpha = \{0.125, 0.25, 0.375, 0.5\}$ and different values of r along with that achieved from setting $r = N - 1$ (represented with dashed horizontal lines). The same but for away teams is shown in Figure 13. In both cases, there is no robust evidence of a difference in the mean profit/loss achieved from using different values of r.

9 Discussion

The results presented in this paper show that match odds in soccer tend to overreact to favorable and unfavorable runs of form where form is defined relative to expectations given a team's odds. Betting odds are partly driven by supply and demand and, therefore the bias is likely driven by bias in the gamblers, although bias in the bookmakers

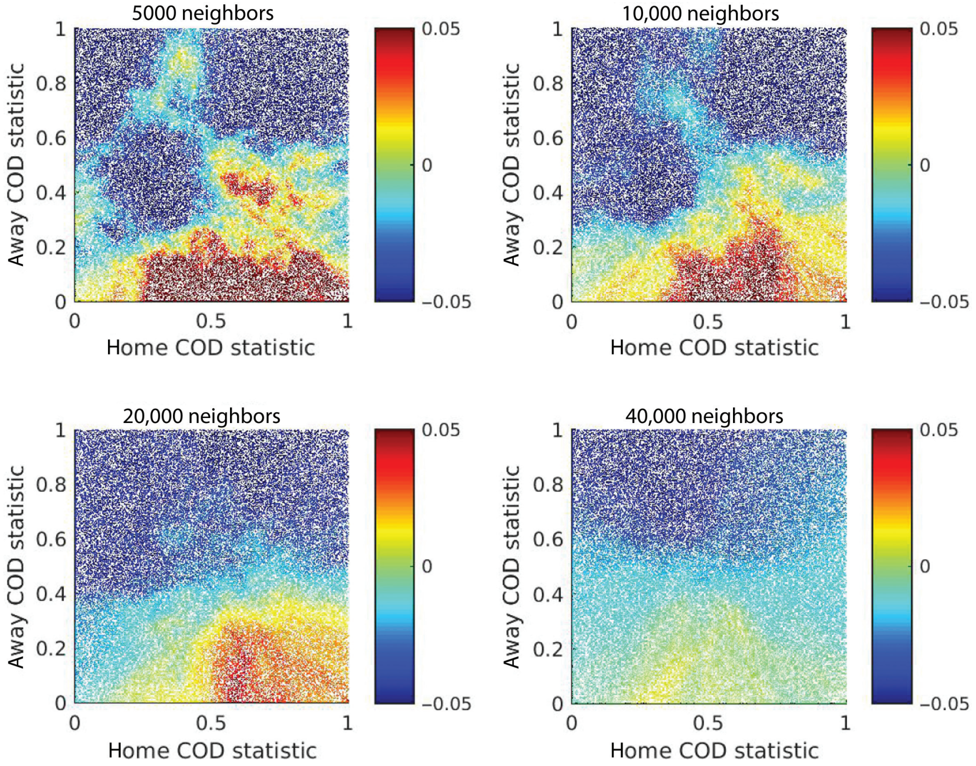

Figure 11: Scatter plot of the home and away COD statistics for each match colored according to the mean profit made by betting on the away team over the number of neighbors stated.

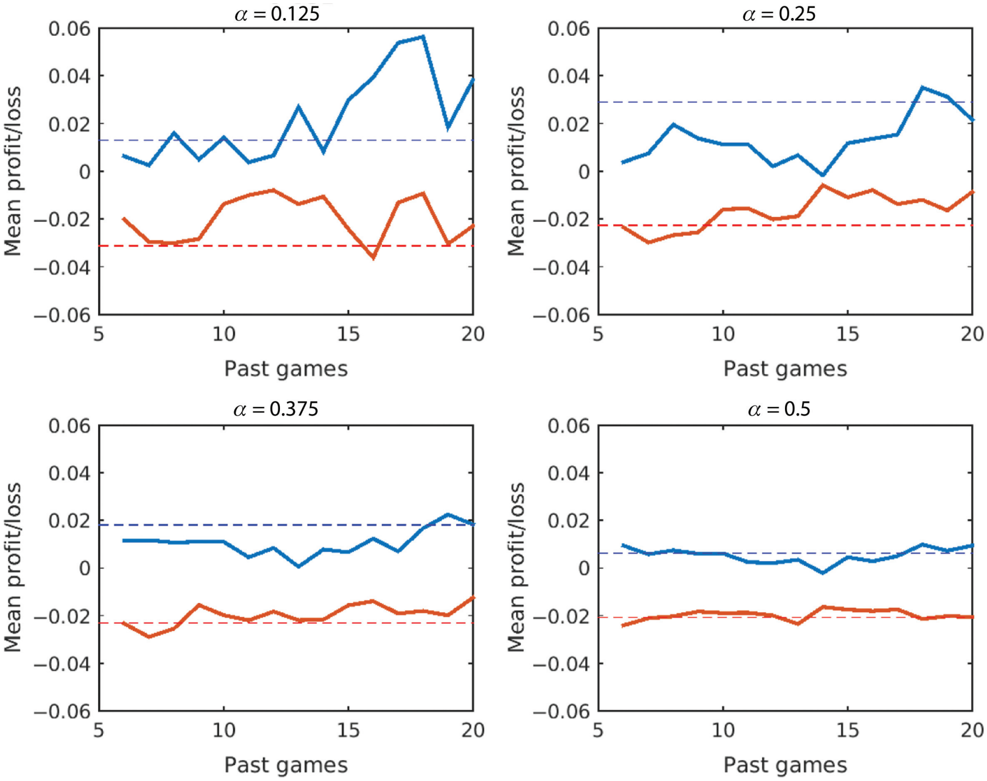

Figure 12: Mean profit/loss achieved by betting on home teams with a COD statistic lower than α (blue) and greater than $1 - \alpha$ (red) for different values of N. The horizontal lines represent the mean profit/loss achieved when using the COD statistic with $r = N - 1$ in each case.

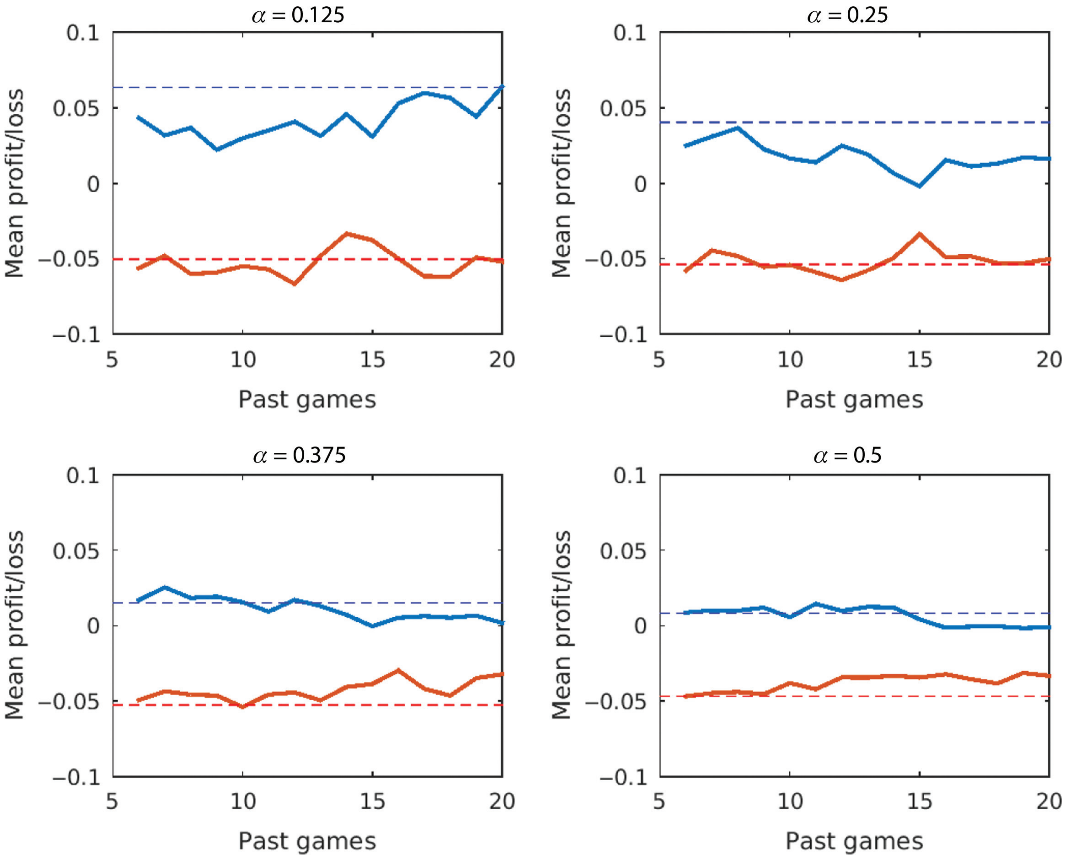

Figure 13: Mean profit/loss achieved by betting on away teams with a COD statistic lower than α (blue) and greater than $1 - \alpha$ (red) for different values of N. The horizontal lines represent the mean profit/loss achieved when using the COD statistic with $r = N - 1$ in each case.

may also play a role (Kaunitz, Zhong, and Kreiner 2017). The hot hand phenomenon is a well known concept that is often argued to cause a mistaken belief that a player or a team is likely to perform better if they have achieved recent favorable results (Gilovich et al. 1985), although this is now disputed by some.

The COD Statistic provides a means with which to assess the performance of a team relative to expectations in previous games. A high COD statistic implies that a team has performed favorably relative to expectations, and the poor value of odds offered on these teams implies that gamblers tend to bet disproportionately on them, overreacting to recent performances. Whilst there is some disagreement as to the existence of the "hot-hand", this paper does not advocate for its existence or otherwise, but merely suggests that its effect, if any, is overestimated.

Whilst the results in this paper have largely been presented in the context of unconscious biases, it is possible that bookmakers are aware of the inefficiency in the odds but are willing to leave such opportunities available for commercial reasons. It is believed that bookmakers are often willing to offer inefficient odds in order to maximise profit (Kuypers 2000) and may close the accounts of those who persistently take advantage of such opportunities (Purdum 2019).

The results demonstrated in this paper show an inherent bias present within bookmakers' odds on European soccer matches. To attempt to assess how widespread this bias is in betting markets, an interesting next step would be to study bookmakers' odds in other sports such as rugby, cricket, tennis and American Football, as well as non-sporting markets such as those offered in politics. It may, however, be somewhat more difficult to do this since no sport is as widely played as soccer globally and hence the number of data points with which to detect this bias would be lower. Another interesting thing to consider would be whether, over time, the bias demonstrated will reduce as gamblers become more sophisticated (it should be noted, however, that there is no sign of this happening in the data that have been considered). Computer algorithms, which are becoming more and more widespread in the prediction of sporting outcomes for betting purposes, may be argued to be less prone to cognitive biases and thus less likely to be impacted by the biases described in this paper. Having said that, since algorithms are designed by humans, these kinds of biases may find their way into such algorithms anyway. Regardless, the results in this paper clearly show that, over the last 12 years, an inherent bias has been present in betting odds of European soccer matches from which a demonstrable profit could have been made. Understanding this bias through the COD statistic therefore represents an opportunity for serious gamblers to make a long term profit and for a better understanding of the behavior of gamblers to be obtained.

Acknowledgment: I am grateful to Leonard A. Smith who provided intellectually stimulating discussion regarding the concept and naming of the Combined Odds Distribution (COD) statistic. I am also grateful to several anonymous reviewers who provided invaluable feedback on earlier versions of this paper.

References

Abinzano, I., L. Muga, and R. Santamaria. 2016. “Game, Set and Match: the Favourite-Long Shot Bias in Tennis Betting Exchanges.” *Applied Economics Letters* 23:605–608.

Ashiya, M. 2015. “Lock! Risk-Free Arbitrage in the Japanese Racetrack Betting Market,” *Journal of Sports Economics* 16:322–330.

Badarinathi, R. and L. Kochman. 1994. “Does the Football Market Believe in the" Hot Hand"?” *Atlantic Economic Journal* 22:76–76.

Baron, J. and J. C. Hershey. 1988. “Outcome Bias in Decision Evaluation.” *Journal of Personality and Social Psychology* 54:569–579.

Braun, S. and M. Kvasnicka. 2013. “National Sentiment and Economic Behavior: Evidence from Online Betting on European Football.” *Journal of Sports Economics* 14:45–64.

Buckle, M. and C.-S. Huang. 2018. “The Efficiency of Sport Betting Markets: An Analysis Using Arbitrage Trading within Super Rugby. *International Journal of Sport Finance* 13:279–294.

Buhagiar, R., D. Cortis, and P. W. Newall. 2018. “Why do Some Soccer Bettors Lose more Money than Others?” *Journal of Behavioral and Experimental Finance* 18:85–93.

Cain, M., D. Law, and D. Peel. 2000. “The Favourite-Longshot Bias and Market Efficiency in UK Football Betting.” *Scottish Journal of Political Economy* 47:25–36.

Camerer, C. F. 1989. “Does the Basketball Market Believe in Thehot Hand?” *The American Economic Review* 79:1257–1261.

Choi, D. and S. K. Hui. 2014. “The Role of Surprise: Understanding Overreaction and Underreaction to Unanticipated Events Using In-Play Soccer Betting Market.” *Journal of Economic Behavior & Organization* 107:614–629.

Clarke, S., S. Kovalchik, and M. Ingram. 2017. “Adjusting Bookmaker’s Odds to Allow for Overround.” *American Journal of Sports Science* 5:45–49.

Constantinou, A. and N. Fenton. 2013. “Profiting from Arbitrage and Odds Biases of the European Football Gambling Market.” *Journal of Gambling Business and Economics* 7:41–70.

Croson, R. and J. Sundali. 2005. “The Gambler’s Fallacy and the Hot Hand: Empirical data from Casinos.” *Journal of risk and uncertainty* 30:195–209.

Dare, W. H. and S. A. Dennis. 2011. “A Test for Bias of Inherent Characteristics in Betting Markets.” *Journal of Sports Economics* 12:660–665.

Daunhawer, I., D. Schoch, and S. Kosub. 2017. “Biases in the Football Betting Market. Available at SSRN: https://ssrn.com/abstract=2977118 or http://dx.doi.org/10.2139/ssrn.2977118.

De Bondt, W. F. and R. Thaler. 1985. “Does the Stock Market Overreact?” *The Journal of finance* 40:793–805.

Feddersen, A., B. R. Humphreys, and B. P. Soebbing. 2018. “Sentiment Bias in National Basketball Association Betting.” *Journal of Sports Economics* 19:455–472.

Feddersen, A. 2017. “Market Efficiency and the Favorite–longshot Bias: Evidence from Handball Betting Markets. in *Economics of Sports Betting*, Edward Elgar Publishing, Incorporated, 105–117.

Fischhoff, B. and R. Beyth. 1975. “I knew it would Happen: Remembered Probabilities of Once—Future Things.” *Organizational Behavior and Human Performance* 13:1–16.

Forbes, W. P. 1996. “Picking Winners? A Survey of the Mean Reversion and Overreaction of Stock Prices Literature.” *Journal of Economic Surveys* 10:123–158.

Forrest, D. and R. Simmons. 2008. “Sentiment in the Betting Market on Spanish Football.” *Applied Economics* 40:119–126.

Gilovich, T., R. Vallone, and A. Tversky. 1985. “The Hot Hand in Basketball: On the Misperception of Random Sequences.” *Cognitive Psychology* 17:295–314.

Gomez-Gonzalez, C. and J. Del Corral. 2018. “The Betting Market Over Time: Overround and Surebets in European Football.” *Economics and Business Letters* 7:129–136.

Hassanniakalager, A. and P. W. Newall. 2018. “A Machine Learning Perspective on Responsible Gambling.” *Behavioural Public Policy* 1–24.

Humphreys, B. R., R. J. Paul, and A. P. Weinbach. 2013. “Bettor Biases and the Home-Underdog Bias in the NFL.” *International Journal of Sport Finance* 8:294–311.

Johnson, J., M. Sung, D. McDonald, and C. Tai. 2013. “Forecasting the Presence of Favourite-Longshot Bias in Alternative Betting Markets.

Kaunitz, L., S. Zhong, and J. Kreiner. 2017. “Beating the Bookies with their Own Numbers-and how the Online Sports Betting Market is Rigged. *arXiv preprint arXiv:1710.02824*.

Kuypers, T. 2000. “Information and Efficiency: An Empirical Study of a Fixed Odds Betting Market.” *Applied Economics* 32:1353–1363.

Miller, J. B. and A. Sanjurjo. 2018. “Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers.” *Econometrica* 86:2019–2047.

Ottaviani, M. and P. N. Sørensen. 2008. “The Favorite-Longshot Bias: An Overview of the Main Explanations.” *Handbook of Sports and Lottery Markets* 83–101.

Purdum, D. 2019. “Won and Done? Sportsbooks Banning the Smart Money. http://www.espn.com/chalk/story/_/id/24425026/gambling-bookmakers-growing-us-legal-betting-market-allowed-ban-bettors, accessed: 30/01/2019.

Shin, H. S. 2008. “Prices of State Contingent Claims with Insider Traders, and the Favourite-Longshot Bias. Pp. 343–352 in *Efficiency of Racetrack Betting Markets*. World Scientific.

Spann, M. and B. Skiera. 2009. “Sports Forecasting: A Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters. *Journal of Forecasting* 28:55–72.

Surowiecki, J. 2005. *The Wisdom of Crowds*. Anchor.

talkSPORT. 2017. “Odds to be Relegated from the Premier League in 2017/18: Every Club Ranked from Most to Least Likely. https://talksport.com/football/249068/odds-be-relegated-premier-league-201718-every-club-ranked-most-least-likely-170624244236/, accessed: 11/11/2018.

Tassoni, C. J. 1996. “Representativeness in the Market for Bets on National Football League Games.” *Journal of Behavioral Decision Making* 9:115–124.

Tversky, A. and D. Kahneman. 1974. “Judgment Under Uncertainty: Heuristics and Biases.” *science* 185:1124–1131.

Vergin, R. C. 2001. “Overreaction in the NFL Point Spread Market. *Applied Financial Economics* 11:497–509.

Vlastakis, N., G. Dotsis, and R. N. Markellos. 2009. “How Efficient is the European Football Betting Market? Evidence from Arbitrage and Trading Strategies.” *Journal of Forecasting* 28:426–444.

Winterbottom, A., H. L. Bekker, M. Conner, and A. Mooney. 2008. “Does Narrative Information Bias Individual’s Decision Making? A Systematic Review.” *Social science & medicine* 67:2079–2088.

Woodland, B. M. and L. M. Woodland. 2015a. “Testing Profitability in the NBA Season Wins Total Betting Market. *International Journal of Sport Finance* 10:160–174.

Woodland, L. M. and B. M. Woodland. 2015b. “The National Football League Season Wins Total Betting Market: The Impact of Heuristics on Behavior.” *Southern Economic Journal* 82:38–54.

Woodland, B. M. and L. M. Woodland. 2016. “Additional Evidence of Heuristic-Based Inefficiency in Season Wins Total Betting Markets: Major League Baseball.” *Journal of Economics and Finance* 40:538–548.