DiSES Working Papers

Università degli Studi di Salerno
Dipartimento di Scienze Economiche e Statistiche
>>>**www.dises.unisa.it**

ON THE LONGSHOT BIAS IN TENNIS BETTING MARKETS: THE CASCO NORMALIZATION

Vincenzo Candila
Antonio Scognamillo

Working Paper No. 3.236

ISSN: 1971-3029
Publication date: March 20th, 2017

Dipartimento di Scienze Economiche e Statistiche
Università Degli Studi di Salerno
Via Ponte Don Melillo – 84084; Fisciano (SA) – Italy

Tel +39-089-96.21.55
Fax +39-089-96.20.49
E-mail dises@unisa.it
Web www.dises.unisa.it

ON THE LONGSHOT BIAS IN TENNIS BETTING MARKETS: THE CASCO NORMALIZATION

Vincenzo Candila* Antonio Scognamillo†

Abstract. This study focuses on investigating bookmakers' behavior in the tennis gambling market in presence of a clear underdog. The aim of this paper is threefold. First, it investigates the distance (bias) between the true but unobserved probability of a given sport outcome and the published odd by a bookmaker. Second, it tests the predictive skills of the most widespread normalization methods when a player is clearly favourite on another. Third, it proposes a new normalization method (called CaSco normalization), which takes into account the positive relationship between the bias and the distance between the odds. The empirical analysis relies on sample odds provided by Bet365 about over 27,000 matches from 2005 to 2015. Our findings show that. First, when there is a clear underdog, the bookmaker minimises the losses in case of unexpected outcomes by increasing the bias in the public available odds. Second, the normalization methods which take into account the bias generally perform better than the other alternatives. Third, in-sample forecasts based on CaSco normalization always outperform the other methods and more importantly, the proposed technique always guaranties unbiased normalized probabilities.

Keywords: Bookmaker behavior, Betting, Favourite-Longshot Bias, Forecasting

AMS 2010 classifications: 60G25, 9102, 91C99.

JEL classifications: C10, C50, C52.

1. Introduction

Over the past few years, the extensive deregulation, the abolition of national monopolies and the advent of on-line gambling have resulted in the exponential growth of the betting market (Vlastakis et al., 2009). As an example, according to the United Kingdom Gambling Commission, the non-remote betting market has generated a gross gambling yield of £3.25 billion from April 2014 to March 2015. It corresponds to an increase of 12% relative to the size of the market five years before.

Hence, sports betting is increasingly considered a mass participation mainstream leisure activity. Two kinds of actors play the betting markets: traditional bookmakers, who represent the supply side, and who set the odds, and the bettors, who represent the demand side, and who wager money on an event with an uncertain outcome. The bettors can participate in these markets using traditional bookmakers or person-to-person betting exchanges.

The question about as to which of these two (bookmakers or exchanges) has superior skills in predicting outcomes has been the focus of a huge amount of studies. For instance,

*DISES, Via Giovanni Paolo II, 84084, Fisciano (SA), Italy, vcandila@unisa.it

†Food and Agriculture Organization of the United Nations, Rome, Italy, antonio.scognamillo@fao.org

Smith et al. (2009) find a greater efficiency on the part of the exchange markets in reflecting the outcome probabilities with respect to traditional bookmakers.

However, before performing a forecasting evaluation, it is necessary to derive the true probability attributed by the bookmaker to a given outcome. In fact, the published odds are considered a proxy of probability but not the actual probability, due to the presence of the favourite-longshot bias. This bias consists of the difference between the published odds and the probability expectations regarding the outcome set by the market maker, whose behavior depends on their profit strategy.

Generally speaking, the method that is used to derive the implied probabilities starting from the raw odds is called "normalization". Three of the most common normalization methods are: the basic, regression and Shin (Shin, 1991, 1992, 1993) methods. Štrumbelj (2014b) and Štrumbelj (2014a) compare the forecasting performance of these methods and show that Shin normalized probabilities are more accurate forecasts than those determined using basic or regression normalization.

The origin of the bias has been extensively debated in the literature. Early studies explained the bias by reference to the demand-side factors relating to bettor rationality (Weitzman, 1965). These explanations are based on bettors' "risk loving" behavior, or the so called "fans' sentiment". More recently, several explanations based on the behavioral characteristics of market makers (Shin, 1991, 1992, 1993), or structural characteristics of the markets, such as the cost of acquiring information relevant to the outcomes (Hurley and McDonough, 1995; Sobel and Travis Raines, 2003) have become common.

Whatever the nature of the bias is, it is a matter of fact that its existence undermines betting market efficiency and predictability. Thus, while the emphasis of the literature has largely focused on the efficiency of betting markets (Stekler et al., 2010), little attention has been devoted to the normalization procedures and how well they perform. This article aims to shed some light on these issues. More precisely, we are interested in verifying if the previous cited normalization methods are always applicable. This is directly connected with the magnitude of the bias. Our conjecture is that the greater the distance between the probability odds, the larger the bias. From a behavior point of view, we believe that the bookmakers alter the published odds more (allowing the bias to increase) when a clear favourite and a clear underdog are present.

In this work, we restrict our attention to the tennis betting markets, for two main reasons. First, because of the high amount of data that are publicly available. Second, due to the features of tennis match outcomes, where draws do not exist, which make easier the division of the bias across the two players.

In the literature, different contributions highlight the presence of such a bias in tennis betting markets. For instance, Lahvička (2014) argues that, considering a dataset of over 44,000 single tennis matches, not only is the favourite-longshot bias present, but it appears more heavily in matches between lower-ranked players, in later-rounds and in high-level tournaments. Very similar results have been found by Abinzano et al. (2016), in tennis betting exchange markets. Given the existence of such a bias, it is fundamental to adopt methods that are able to determine as much as possible the true probability of winning for each player. No less important is the robustness of these methods to all type of matches.

Hence, the aim of this work is threefold. First, we investigate if the bias increases with the distance between the probability odds. Second, we test how the most widespread nor-

malization methods behave, when one player is the clear favourite. Third, we propose a new normalization method, called Candila-Scognamillo (CaSco), that is robust to all the distances between the odds.

Overall, we find that the bias increases as long as the distance between the odds increases. Moreover, the analysed normalization approaches are generally not robust to matches where the distance between the odds is large. Rather, the probabilities obtained from the CaSco approach are not biased, independently of the distance between the odds. From a forecasting point of view, the CaSco probabilities have a superior forecasting ability with respect to the other approaches, in terms of in-sample estimation.

The rest of the paper proceeds as follows: Section 2 presents the three normalization methods cited above. Section 3 is devoted to the illustration of data and the investigation of the relationship bias - distance. Section 4 introduces the CaSco normalization technique. Section 5 evaluates the CaSco normalization method against the other normalization approaches, in terms of forecasting ability and bias robustness. Section 6 concludes.

2. Normalization methods

In individual sports with two possible outcomes, there are only two results: the victory of a player and the defeat of the other. Thus, let $\boldsymbol{o_j} = (o_{1,j}, o_{2,j})$ be the observed odds for a match j and players 1 and 2. The odds $o_{1,j}$ and $o_{2,j}$ denote how much a bettor receives investing one dollar, for instance, if player 1 or 2 wins, respectively. The probability odds $\boldsymbol{\pi_j} = (\pi_{1,j}, \pi_{2,j})$ are:

$$\pi_{1,j} = \frac{1}{o_{1,j}} \qquad \text{and} \qquad \pi_{2,j} = \frac{1}{o_{2,j}}. \tag{1}$$

Each probability can be considered as a proxy of the player strength. In fact, the smaller the quote for player i is, the larger π_i is, more likely the winning of that player is. Within this framework, we refer to the booksum Π as the sum between the probability odds associated to the two possible outcomes calculated as:

$$\Pi_j = \sum_{i=1}^{2} \pi_{i,j}. \tag{2}$$

The empirical evidence generally shows that, since the booksum is normally greater than one, probability odds do not represent the true probabilities of winning for each player attributed by the bookmakers. This means that there is a difference between the real but unobserved probability of winning and the bookmakers' published odds. The literature on this topic has highlighted two main reasons to explain this evidence. The first reason is that the booksum incorporates the margin of the bookmaker, say $m_j = \Pi_j - 1$. The second reason concerns the fact that the observed quotes are adjusted by the bookmakers in order to maximize their profits given some possible external factors. This second element is usually referred as "longshot bias". According to Forrest et al. (005a), "a (positive) longshot bias implies that financially superior returns (i.e. smaller losses) accrue to a strategy of wagering on short-odds rather than long-odds players". Lahvička (2014) reviews the huge amount of literature on this topic, identifying three possible explanations for the phenomenon. The first explanation concerns the fact that the bookmakers know that the bettors risk function

is locally convex and they take advantage of that by lowering the longshot odds (Friedman and Savage, 1948). The second explanation relies on the assumption that bookmakers increase their profit by setting lower longshot odds because of the bettors' bounded cognitive ability, which lead them to overestimate the longshot winning probability (Kahneman and Tversky, 1979). The third explanation assumes that both bettors and available information are not homogeneous. As a result, the existence of a number of insider traders, who preventively know the match outcome or react faster to the new available information, exposes the bookmakers to huge potential losses. As a coping strategy against this type of loss, bookmakers tend to offer lower longshot odds (Shin, 1991).

As said above, the methods to determine the true (but unobserved) probabilities, also called implied probabilities, from the published odd are generally defined with the acronym of normalization. In what follows, we briefly present three of the most common normalization methods.

Basic normalization

The basic normalization consists of dividing the probability odds by their sum. This is the most simple normalization technique. It has largely and commonly applied in the majority of the studies (for example: Franck et al. (2010)) so that this approach has become almost a synonymous of betting odds. Formally, the probability of winning of player i, for the match j, denoted with $p_{i,j}$, is obtained as:

$$p_{i,j} = \frac{\pi_{i,j}}{\Pi_j}. \tag{3}$$

The basic normalization does not rely on any assumptions and divides proportionally the part of the booksum exceeding one. In doing so, this method does not take into account the part of the margin related to the longshot bias. In other words, the ratio between the probability odds is the same as the ratio between the normalized ones.

Normalization by regression

The normalization by regression consists of regressing the observed outcomes on historical betting odds. Forrest et al. 005a and Goddard and Asimakopoulos 2004, among others, have reently used this approach to derive implied probabilities. When the outcome of the even under consideration is is binary, the probit or logit estimators are used. Formally, let $Y_{1,j}$ be the observed outcome of the j-th match for player 1. If the probit estimator is employed, $p_{1,j}$ is obtained after running the following regression:

$$Pr(Y_{1,j} = 1|(\pi_{1,j}, \pi_{2,j})) = \Phi\left(\beta_0 + \beta_1\pi_{1,j} + \beta_2\pi_{2,j}\right), \quad \text{with} \quad j = 1, \cdots, J.$$

Once run the regression for a given dataset, β_0, β_1 and β_2 are estimated such that $p_{1,j}$ can be easily calculated. Obviously, the other player's probability of winning for the same match is obtained by difference, i.e. $(1 - p_{1,j})$.

This methodology has some shortcomings. First, it requires an historical set of betting odds and match outcomes and, unfortunately, for some sports the betting quotes are not so publicly available. The smaller this set is, the less reliable and precise the estimated probabilities are. Second, and more importantly, it does not take into account the problem of the longshot bias.

Shin normalization

The model relies on the assumption that bookmakers formulate the betting odds maximizing their expected profit in a market where both uninformed bettors and a small portion of insider traders are present. In particular, bookmakers set a spread in the published odds in a bid to minimize the losses arising from the existence of a group of gamblers who have insider information on the outcome of the event. The bookmaker cannot distinguish gamblers who have insider information from those who do not, but has some idea on the proportion of one group. Assuming that the spread is increasing with the incidence of insider trading, the size of the observed spread provides some indication of the severity of market distortion due to insider trading that results in a (positive) longshot bias. Jullien and Salanie (1994), starting from the solution of the game proposed by Shin (1993), reverse the problem and derive the probabilistic beliefs (i.e. the Shin probabilities) given the odds. In a nutshell, Shin probabilities are the normalized odds obtained by taking into account the longshot bias due to the existence of insider traders.

Formally, let the "distance" be the observed span between the probability odds for the two players, that is: $d_j = \pi_{1,j} - \pi_{2,j}$, for a match j. Note that d_j can be positive or negative. The probability of winning for player i is:

$$p_{i,j} = \frac{\sqrt{z_j^2 + 4(1 - z_j)\frac{\pi_{i,j}^2}{\Pi_j}} - z_j}{2(1 - z_j)}, \tag{4}$$

where z_j represents the proportion of insider traders. Moreover, the larger z_j is, the greater the bias is (Smith et al., 2006). In case of sports with only two outcomes, Jullien and Salanie (1994) and Cain et al. (2001) demonstrate that that z_j depends only on the margin and the distance between the probability odds:

$$z_j = \frac{m_j(d_j^2 - \Pi_j)}{\Pi_j(d_j^2 - 1)}, \tag{5}$$

where, as seen above, m_j represents the margin for the match j. The Shin method has the great advantage to take into account the longshot bias, incorporating the behavior of bookmakers in order to face the problem of insider traders. Nevertheless, this method also has a drawback. In fact, the Shin method relies on a set of *a priori* assumptions about the existence and proportion of insider traders that, in turn, are based on the observed spread between the probability odds of each possible outcome.

Table 1: Summary statistics

	All rounds			1st rounds			Semifinals		
	J	Margin	Distance	J	Margin	Distance	J	Margin	Distance
ATP 250	12249	0.076	0.356	5692	0.076	0.351	836	0.069	0.331
ATP 500	4105	0.073	0.402	1807	0.076	0.376	229	0.067	0.428
ATP 1000	5806	0.070	0.402	2278	0.076	0.351	181	0.057	0.484
ATP Finals	159	0.058	0.458	-	-	-	22	0.057	0.484
Grand Slam	5234	0.068	0.533	2613	0.069	0.533	85	0.056	0.550
Tot./Median	27553	0.070	0.402	12390	0.076	0.376	1353	0.067	0.376

Notes: Columns "J" show the number of matches, per type of tournament. Columns "Margin" show the median of the margins obtained from the summation of the probability odds offered by Bet365 minus one, per type of tournament. Columns "Distance" show the median of the distance between the probability odds, per type of tournament.

3. Bias in the tennis betting markets

Once presented the most common normalization techniques, let us verify the bias in tennis betting market and how these techniques deal with it. In particular, we use the betting quotes offered by the professional bookmaker Bet365 in the male tennis market, collecting over 27,000 matches from the Tennis Data provider.[1] The matches cover the period 2005-2015 and consider all the four Grand Slams (Australian Open, Roland Garros, Wimbledon, U.S. Open) in each year, as well as all the Association of Tennis Professionals (ATP) world tour tournaments, namely ATP 250, ATP 500 and ATP 1000, plus the ATP Finals.

Table 1 presents the number of matches as well as the median of the margin m and distance d per tournament typologies and rounds.

First of all, we note that as long as the tournament has a greater importance, the median of the margin decreases while the median of the distance increases. Looking at their correlation, we find a strong negative linear relationship (-0.57). That is, the greater the distance is, the clearer a favourite player is, the smaller the margin is and vice versa. This holds independently of the round of the tournament considered.

Similarly, regardless of the type of tournament, the (median of the) margin is greater in the first rounds of a competition (0.076) than those in the later rounds (0.067).

Our intuition is that the relationship between the margin and the distance leads the bookmaker to alter more the odds when there is a clear favourite and a clear longshot player, reducing the margin and increasing the bias.

Generally speaking, the existence and the nature of the bias in the tennis market have been largely debated in the literature. A first study by (Cain et al., 2003) did not find any clear pattern regarding the existence of such a bias. Unfortunately, the study considers only a very small sample of matches played at Wimbledon 1996. On the other hand, exploiting a larger dataset of 5892 matches from 2001 to 2004, Forrest and McHale (2005) find that the tennis betting market is characterized by a positive longshot bias similar to that found in horse betting markets. The already cited works of Lahvička (2014) and Abinzano et al. (2016) confirm the existence of longshot bias in the market under consideration. Our analysis corroborates the previous results, as summarized in Table 2. In fact, betting on the favorite yields larger returns (or smaller losses). In particular, a statistically significant bias has been found in all the considered years (except 2013) in the whole period. However,

[1] www.tennis-data.co.uk

Table 2: Returns on betting on favorites and longshots

	All matches			Distance > 0.8		
	Favourite	Longshot	J	Favourite	Longshot	J
2005	-0.038^{***}	-0.184^{***}	2600	-0.002	-0.662^{***}	194
2006	-0.056^{***}	-0.151^{***}	2516	0.018	-0.838^{***}	194
2007	-0.038^{***}	-0.164^{***}	2587	-0.017	-0.487^{***}	229
2008	-0.051^{***}	-0.141^{***}	2496	-0.012	-0.540^{***}	226
2009	-0.062^{***}	-0.145^{***}	2511	-0.012	-0.556^{***}	305
2010	-0.052^{***}	-0.147^{***}	2508	-0.011	-0.571^{***}	239
2011	-0.034^{***}	-0.179^{***}	2514	-0.014	-0.506^{***}	264
2012	-0.044^{***}	-0.139^{***}	2531	-0.022	-0.234	354
2013	-0.075^{***}	-0.073^{*}	2448	-0.025^{*}	-0.156	301
2014	-0.048^{***}	-0.150^{***}	2373	-0.014	-0.463^{***}	242
2015	-0.040^{***}	-0.175^{***}	2469	-0.012	-0.385^{**}	273
2005-2015	-0.049^{***}	-0.150^{***}	27553	-0.013^{***}	-0.465^{***}	2821

Notes: The first three columns show the returns on betting on favourites, longshots and the number of matches. The last three columns show the returns on betting on favourites, longshots and the number of matches whose distance between the probability odds is greater than 0.8. *, ** and *** denote significance of the t-test at the 10%, 5% and 1% levels, respectively.

differently from (Forrest and McHale, 2005), we find that both the strategies yield negative returns and no-profitable strategy exists. In accordance with Cain et al. (2003), we have also considered the sub-sample of matches in which there is a clear favourite, i.e. the odd probabilities for the favourite are greater than 0.80. In this case, betting on the favourite allows the bettor to break even in several cases (the losses are not statistically different from zero) or experience very small losses. On the other hand, betting on the clear underdog yields very high losses, greater than in the full-sample scenario.

Overall, these results only confirm that a positive longshot bias in the tennis betting market exists. Now, let us verify how the most common normalization methods face the bias problem and how they behave when the distance between the odds increases. In particular, according to Lahvička (2014), we regress the result of the matches on the implied probabilities, by using a simple OLS estimator and heteroskedasticity robust standard errors. More precisely, the following regression is carried out:

$$Y_{i,j} = \beta_0 + \beta_1 p_{i,j}, \quad \text{with } j = 1, \cdots, J, \tag{6}$$

where, $Y_{i,j} = 1$ if player i wins and 0 otherwise and $p_{i,j}$ is the implied probability for that player under a given normalization method. In eq. (6), for each match j, a player and consequently his implied probability is randomly chosen. According to the literature, the null of no bias occurs when $\beta_0 = 0$ and $\beta_1 = 1$, while the presence of bias occurs when $\beta_0 < 0$ and $\beta_1 > 1$. This is because under the efficient market hypothesis, prices (odds) should reflect all the available information related to the outcome of the event (Coombes et al., 1998) such that they can predict perfectly the probability associated to the outcome (Forrest and McHale, 2005). The results of the estimations are in Table 3. In the table, the significance of the previous null hypotheses is reported. Many points can be underlined. First, when all the matches are considered, the implied probabilities provided by the Basic and normalization by regression are always biased. Instead, the Shin probabilities are not biased. Second, when all the matches whose distance between the odds is greater than 0.7 are analysed (center panel of the table), not only the bias signalled by the Basic and normalization by regression increases, but also that of the Shin. The situation becomes clearer when the distance between the odds overcomes 0.8. In this regard, all the considered normalization methods provide biased probabilities. In fact, even for the Shin method, the null of no bias is rejected at 10% significance for β_0 and 1% for β_1.

Table 3: Longshot bias with respect to the normalization methods

	Basic	Regression	Shin
		Whole sample	
β_0	-0.0362^{***}	-0.0368^{***}	-0.0040
β_1	1.0760^{***}	1.0709^{***}	1.0117
		Distance > 0.7	
β_0	-0.0344^{***}	0.0263^{***}	-0.0055
β_1	1.0757^{***}	0.9573^{***}	1.0180^{***}
		Distance > 0.8	
β_0	-0.0388^{***}	0.0329^{***}	-0.0119^{*}
β_1	1.0855^{***}	0.9451^{***}	1.0318^{***}

Notes: The table reports the estimated coefficients of the OLS regression "Results of the matches on the implied probabilities", where the variable "Result" can assume value 1 or 0 and the implied probabilities are the normalized probabilities according to the methods in column. The number of matches included in the top, center and bottom panel are 27,553, 4,954 and 2,821, respectively. *, ** and *** denote significance of the t-test at the 10%, 5% and 1% levels, respectively.

So far, the first two aims of this work can be answered. The bias between the true but unobserved probabilities attributed by the bookmaker and the implied probability increases when the matches have a clear favourite. This means that the bookmakers behave differently under these cases. In our opinion, this is because, when a player is clearly a favourite, the bookmaker losses, in case of a longshot victory, are potentially very high. In this regard, the bookmaker "fears" the underdog so much to diminish the longshot odd (meaning that his probability odd increases). Such bookmaker behavior prevents huge losses arising from unexpected outcomes. Conversely, when the distance between the observed probability odds is small, the bookmaker can reduce the longshot bias, offering more attractive odds with no high loss risk. This means that, under these circumstances, there is less difference between the probabilities odds and the true but unobserved probabilities.

With reference to our second aim, we can argue that the normalization methods analysed in this work are not robust to all the type of matches. Really, two of the three methods produce biased implied probability independently of the distance between odds. Instead, the Shin method fails to solve the bias problem only when the distance between the published odds is high. Thus, a proper method, robust to all the types of matches, is needed. The rest of the paper is devoted to the presentation and verification of the proposed approach.

4. The CaSco normalization

The CaSco normalization technique is designed to work under all types of matches. It assumes that the probability of winning for a player is obtained subtracting one half of the margin to the probability odd of that player, when the distance is low. Instead, when the distance is greater than a given threshold ψ, the probability of winning for a player is obtained by subtracting a percentage of the margin δ from the probability odd of that player. By subtracting one half of the margin, under low-distance matches, we assume that the bookmaker alters less the published odds. Instead, under high-distance matches, the behavior of bookmaker becomes more substantial.

Formally, the CaSco normalization calculates the probability of winning of player i for the match j, as follows:

$$p_{1,j} = \begin{cases} \pi_{1,j} - 0.5 \cdot m_j & \text{if } |d_j| \leq \psi \\ \pi_{1,j} - \delta \cdot m_j & \text{if } |d_j| > \psi \text{ and } d_j > 0 \\ \pi_{1,j} - (1-\delta) \cdot m_j & \text{if } |d_j| > \psi \text{ and } d_j < 0 \end{cases} . \tag{7}$$

In (7), the key quantities are ψ and δ, respectively the threshold and "amount" of the margin to subtract from the probability odds. If d_j is positive, then player 1 for that particular match is the favourite and vice versa. The sense of (7) is that if player 1 is the favourite, then its probability odd is diminished of a quantity varying from $\delta \cdot m_j$ to, at most, $0.5 \cdot m_j$. The greater the distance d_j is, more favourite player 1 is, the less the CaSco normalization subtracts the margin from the probability odds. If, instead, player 1 is the longshot, meaning that d_j is negative, the probability odds are diminished of a quantity greater than $0.5 \cdot m_j$.

Note that once calculated the probability for player 1, the other is obtained as its complementary. Moreover, simple algebra revels that

$$p_{1,j} + p_{2,j} = 1, \tag{8}$$

independently of which player enters (7).

Suppose that player 1 is the favourite and that, for a given match j, $d_j > \psi$. Under this situation, we would have:

$$p_{1,j} = \pi_{1,j} - \delta \cdot m_j \quad \text{and} \quad p_{2,j} = \pi_{2,j} - (1-\delta) \cdot m_j.$$

Replacing the last expressions in eq. (8), we obtain:

$$\begin{aligned} \pi_{1,j} - \delta \cdot m_j + \pi_{2,j} - (1-\delta) \cdot m_j &= 1; \\ \pi_{1,j} + \pi_{2,j} - \delta \cdot m_j - m_j + \delta \cdot m_j &= 1; \\ 1 + m_j - m_j &= 1. \end{aligned}$$

Finally, this kind of normalization is asymmetric. Even if the distance is smaller than the threshold ψ, the operation $\pi_i - 0.5 \cdot m$ alters the ratio between the original probability odds and the resulting implied probabilities while the ratio is more heavily altered when the distance is above the threshold ψ. Thus, this method takes into account the longshot bias, as it depends on the margin and distance.

At this point, to make (7) feasible, we need to estimate ψ and δ, subject to the constraints $0 \leq \delta \leq 0.5$ and $\min(d) \leq \psi \leq \max(d)$. Note that the positivity of δ excludes implied probabilities greater than one, while ψ has to be included between the smallest and the greatest distance, for all the considered matches. For ease of notation, from now on, we will consider only player 1 such that $p_{1,j}$ becomes p_j.

Estimation of threshold ψ and amount δ

The estimation of ψ and δ as resulting by eq. (7) is carried out by the linear minimum mean square error (LMMSE) estimator, belonging to the wider class of M-estimators with

Table 4: $\hat{\delta}$, $\hat{\psi}$ and 95% boostrap CI

	$\hat{\theta}$	CI_{LB}	CI_{UB}
δ	0.165	0.022	0.300
ψ	0.800	0.730	0.820

Notes: The table reports the estimated coefficients of the threshold ψ and amount δ. The estimator used is the LMMSE (eq. (7)). The number of matches is 20,190. CI_{LB} and CI_{UB} denote the lower and upper bound of the 95% bootstrap confidence interval, respectively.

constraints. The asymptotic properties of constrained M-estimation are largely discussed in Shapiro (2000) and Geyer (1994), for instance. With our notation, the LMMSE estimator is:

$$\begin{aligned} \underset{\psi,\delta}{\arg\min} \quad & \sum_{j=1}^{J} \left(Y_j - p_{j,(\psi,\delta)}\right)^2 \\ \text{subject to} \quad & 0 \leq \delta \leq 0.5, \\ & \min(d) \leq \psi \leq \max(d). \end{aligned} \tag{9}$$

As above, Y_j in (9) denotes the observed outcome of the match j for player 1. If $Y_j = 0$, then player 1 has been defeated. Otherwise, if $Y_j = 1$, then player 1 has defeated player 2 in the match j, with $j = 1, \cdots, J$. Moreover, the dependence of the normalized probability for player 1 and match j from ψ and δ is expressed with $p_{j,(\psi,\delta)}$. Practically speaking, the estimation of ψ and δ is carried out by numerically solving eq. (9).

In order to perform an in-sample analysis, we consider only the period 2005-2012, consisting of $J = 20,190$ matches for the estimation of ψ and δ. The remaining period is left for the out-of-sample evaluation, to be realized in the future.

The estimated threshold and amount are in Table 4. Because of unknown distribution of δ and ψ, we adopt a bootstrap procedure in order to find the confidence intervals (CI). Repeating the estimation of (9) 200.000 times, we obtain the 95% CI, as highlighted in column 2 and 3 of the same table.

As regards the inference issues, being the zeros not included in the 95% CIs, we can state that $\hat{\delta}$ and $\hat{\psi}$ are statistically different from zero. Thus, if the distance between two probability odds is smaller or equal to $\hat{\psi} = 0.80$, then the CaSco normalization obtains the probability of victory of player 1 by subtracting one half of the margin from π_1. If, instead, the distance is greater than 0.80 and player 1 is the favourite, then his probability of victory is obtained by subtracting $\hat{\delta} = 16.5\%$ of the margin from π_1. Finally, if again the distance is greater than 0.80 but player 1 is the longshot, then his probability of victory is obtained by subtracting 83.5% of the margin from π_1. In doing so, the probability for the longshot player is changed more than that for the favourite.

5. Forecasting and bias evaluation with CaSco probabilities

This section presents the evaluation of the CaSco normalization in terms of forecasting ability and bias robustness.

Table 5: Forecasting evaluation scores

Loss	Functional form
Brier	$1/J \sum_{j=1}^{J} (Y_{1,j} - \hat{p}_{1,j})^2$
K-L	$1/J \sum_{j=1}^{J} [Y_{1,j} \cdot \log(Y_{1,j}/\hat{p}_{1,j}) + (1 - Y_{1,j}) \cdot \log((1 - Y_{1,j})/(1 - \hat{p}_{1,j}))]$
Skill-score	$1 - \sum_{j=1}^{J} \left[(Y_{1,j} - \hat{p}_{1,j})^2 / (Y_{1,j} - E(Y_1))^2 \right]$

Notes: $\hat{p}_{1,j}$ represents the predicted probability of winning for player 1 and match j, under a given normalization method. $E(Y_1)$ represents the average of outcomes for player 1.

Table 6: In-sample normalization methods evaluation

	Brier	K-L	Skill score	R^2
Basic	0.1871***	0.5531***	0.2505	0.3246
Regression	0.1884***	0.5587***	0.2456	0.3133
Shin	0.1868***	0.5516***	0.2518	0.3277
CaSco	**0.1868****	**0.5512****	**0.2519**	**0.3284**

Notes: The table reports in the first three columns the averages of loss functions mapping the distance between the each method in row and the observed outcome. In the last column the Nagelkerke's R^2 is reported, as it results from the regression of the observed outcome on the implied probability obtained from each of the methods in row. The number of matches is 20,190. *** denote significance at the 1% levels.

In order to compare the CaSco forecasting ability with respect to the other analysed approaches, we consider the Brier (Brier, 1950), Kullback− Leibler (K-L) (Lai et al. (2011), eq. (2.1)) and Skill score (Lahiri and Yang (2013), eq. (40)) as well as the Nagelkerke's R^2 (Nagelkerke, 1991). The smaller the first two scores are, the better the relative method is while the opposite holds for the last two scores. The Brier, K-L and the Skill scores are expressed in Table 5.

In-sample analysis

The results of the forecasting evaluation are in Table 6. It results that the CaSco normalization has the best forecasting accuracy, regardless of the score function employed. In fact, it has the smallest distance from the true outcome, on average, and the greatest R^2. Furthermore, accordingly to the previous literature (Štrumbelj, 2014b), we find that the normalization by regression has the worst performance.

The improvement of the forecasting performances guaranteed by the CaSco normalization is also confirmed by the Diebold-Mariano (DM) test, in the version proposed for binary outcomes by Gneiting and Katzfuss (2014) (eq. (11)). In particular, we consider the two-tailed test, comparing the performance of the CaSco with respect to the other three methods, taking into account both the Brier and K-L loss functions. Table 7 shows that the null hypothesis of equal predictive ability between the CaSco and each of the other alternative method is always rejected. Moreover, being the DM statistics negative, the CaSco procedure is always preferred to the Basic, regression and Shin normalization methods.

With reference to the bias issue in terms of distance robustness, we repeat the OLS regression as in (6), this time only considering the period 2005-2012, and adding the performance of the CaSco technique. The results are in Table 8. Again, the null of no bias is always rejected under the Basic and normalization by regression. The Shin technique works properly only when the whole sample is considered while it presents some bias when the matches

Table 7: In-sample differences evaluation

Loss	Basic	Regression	Shin
Brier	-4.0487***	-6.5408***	-2.0767**
K-L	-4.6480***	-5.6388***	-2.0735**

Notes: The table reports the DM statistics of the two tailed test of equal predictive accuracy between the CaSco and each model in column, using the loss function in row. The number of matches is 20,190. ** and *** denote significance at the 5% and 1% levels, respectively.

Table 8: In-sample performance of normalization methods in terms of bias

	Basic	Regression	Shin	CaSco
		Whole sample		
β_0	-0.0379^{***}	-0.0300^{***}	-0.0042	0.0024
β_1	1.0788^{***}	1.0576^{***}	1.0114	0.9980
		Distance > 0.7		
β_0	-0.0344^{***}	0.0230^{***}	-0.0135^{*}	0.0024
β_1	1.0757^{***}	0.9554^{***}	1.0217^{**}	0.9980
		Distance > 0.8		
β_0	0.0471^{***}	0.0307^{***}	-0.0174^{**}	0.0024
β_1	1.1011^{***}	0.9505^{***}	1.0415^{***}	0.9980

Notes: The table reports the estimated coefficients of the OLS regression "Results of the matches on the implied probabilities", where the result variable can assume value 1 or 0 and the implied probabilities are the normalized probabilities according to the methods in column. The number of matches included in the top, center and bottom panel are 20,190, 3,646 and 1,997, respectively. *, ** and *** denote significance of the t-test at the 10%, 5% and 1% levels, respectively.

involve a clear favourite. Instead, the null hypothesis of no bias is always not rejected, under the CaSco method, independently of the type of matches considered.

6. Conclusions

Over the past few years, the extensive deregulation, the abolition of national monopolies and the advent of on-line betting have resulted in the exponential growth of the betting markets (Vlastakis et al., 2009). This growth makes theoretical and empirical analysis of multiple aspects of betting behavior more and more relevant for economic and social research. This study has focused on investigating bookmaker behavior in the tennis betting market. As has been widely discussed in the literature, in these markets there is a bias between the odds published by the bookmakers and the true but unobserved probabilities of the outcome of interest.

The focus of this paper has been on this bias and how the most frequently used methods for deriving the implied probabilities from the published odds deal with it. More specifically, the aim of this paper has been threefold. First, we have investigated the relationship between the bias and the distance between the probability odds. Second, we have tested how the normalization methods considered behave, with respect to matches where one player is clearly the favourite. Third, we have proposed and tested a new normalization method, named CaSco, the goal of which is to produce unbiased probabilities, robustly to all types of matches.

As regards the first aim, we have found that the bias increases with the distance between the odds. In this regard, our results are in line with those of Hurley and McDonough (1995) and Smith et al. (2006), that suggest that the bias is positively related to the bookmakers' profit margin, since the market maker deductions are inversely related to the distance between the two alternative outcome odds. However, our intuition is that the bias is due to the bookmakers' decision process in a risky environment, rather than due to the information search cost. Fearing the underdog's victory, bookmakers alter the published odds more substantially in order to minimise the losses in case of an unexpected outcome, when the match has a clear favourite.

With respect to the second aim, we have verified that the implied probabilities based on the theoretical model developed by Shin (1993) perform better than the other alternatives. This is because the Shin probabilities take into account the presence of a bias in the published odds. However, the Shin probabilities still produce biased probabilities in the case of matches with clear longshots.

Regarding the third aim, the forecasting performance of the proposed approach, expressed as the average distance between the implied probability and the observed outcome, is quite satisfactory. More specifically, the CaSco normalization, from an in-sample perspective, has consistently outperformed the other normalization methods, regardless of the loss functions employed. More importantly, the CaSco probabilities are unbiased, independent of the distance between the published odds (in contrast to the other approaches).

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*(8), 605–608.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review 78*(1), 1–3.

Cain, M., D. Law, and D. Peel (2003). The favourite-longshot bias, bookmaker margins and insider trading in a variety of betting markets. *Bulletin of Economic Research 55*(3), 263–273.

Cain, M., D. Law, and D. A. Peel (2001). The incidence of insider trading in betting markets and the Gabriel and Marsden anomaly. *The Manchester School 69*(2), 197–207.

Coombes, R., L. Frazer, R. Johnson, J. Hockaday, and C. Otto (1998). A case study on the informational efficiency of markets: The market for horse racing in australia. *Journal of Gambling Studies 14*(4), 401–411.

Forrest, D., J. Goddard, and R. Simmons (2005a). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting 21*(3), 551–564.

Forrest, D. and I. McHale (2005). Longshot bias: insights from the betting market on men's professional tennis. In L. Vaughan Williams (Ed.), *Information Efficiency in Financial and Betting Markets*, pp. 215–230. Cambridge: Cambridge University Press.

Franck, E., E. Verbeek, and S. Nüesch (2010). Prediction accuracy of different market structures - bookmakers versus a betting exchange. *International Journal of Forecasting 26*(3), 448–459.

Friedman, M. and L. J. Savage (1948). The Utility Analysis of Choices Involving Risk. *Journal of Political Economy 56*(4), 279–304.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics 22*(4), 1993–2010.

Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application 1*(1), 125–151.

Goddard, J. and I. Asimakopoulos (2004). Forecasting Football Results and the Efficiency of Fixed-odds Betting. *Journal of Forecasting 23*(1), 51–66.

Hurley, W. and L. McDonough (1995). A note on the Hayek hypothesis and the favorite-longshot bias in parimutuel betting. *The American Economic Review 85*(4), 949–955.

Jullien, B. and B. Salanie (1994). Measuring the Incidence of Insider Trading: A Comment on Shin. *The Economic Journal 104*(427), 1418–1419.

Kahneman, D. and A. Tversky (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica 47*(2), 263–292.

Lahiri, K. and L. Yang (2013). Forecasting Binary Outcomes. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Chapter 19, pp. 1025–1106. Elsevier.

Lahvička, J. (2014). What causes the favourite-longshot bias? Further evidence from tennis. *Applied Economics Letters 21*(2), 90–92.

Lai, T. L., S. T. Gross, and D. B. Shen (2011). Evaluating probability forecasts. *Annals of Statistics 39*(5), 2356–2382.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika 78*(3), 691–692.

Shapiro, A. (2000). On the asymptotics of constrained local M-estimators. *Annals of Statistics*, 948–960.

Shin, H. S. (1991). Optimal Betting Odds Against Insider Traders. *The Economic Journal 101*(408), 1179–1185.

Shin, H. S. (1992). Prices of State Contingent Claims with Insider Traders, and the Favourite-Longshot Bias. *The Economic Journal 102*(411), 426–435.

Shin, H. S. (1993). Measuring the Incidence of Insider Trading in a Market for State-Contingent Claims. *The Economic Journal 103*(420), 1141–1153.

Smith, M. A., D. Paton, and L. V. Williams (2006). Market efficiency in person-to-person betting. *Economica 73*(292), 673–689.

Smith, M. A., D. Paton, and L. V. Williams (2009). Do bookmakers possess superior skills to bettors in predicting outcomes? *Journal of Economic Behavior & Organization 71*(2), 539–549.

Sobel, R. S. and S. Travis Raines (2003). An examination of the empirical derivatives of the favourite-longshot bias in racetrack betting. *Applied Economics 35*(4), 371–385.

Stekler, H. O., D. Sendor, and R. Verlander (2010). Issues in sports forecasting. *International Journal of Forecasting 26*(3), 606–621.

Štrumbelj, E. (2014a). A comment on the bias of probabilities derived from betting odds and their use in measuring outcome uncertainty. *Journal of Sports Economics 17*(1), 12–26.

Štrumbelj, E. (2014b). On determining probability forecasts from betting odds. *International Journal of Forecasting 30*(4), 934–943.

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*(5), 426–444.

Weitzman, M. (1965). Utility analysis and group behavior: An empirical study. *Journal of Political Economy February*.