International Journal of Forecasting 30 (2014) 934–943

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

On determining probability forecasts from betting odds

Erik Štrumbelj *

University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, 1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Sports forecasting
Probability forecasting
Fixed-odds
Betting exchange
Shin's model
Betfair
Calibration

ABSTRACT

We show that the probabilities determined from betting odds using Shin's model are more accurate forecasts than those determined using basic normalization or regression models. We also provide empirical evidence that some bookmakers are significantly different sources of probabilities in terms of forecasting accuracy, and that betting exchange odds are not always the best source, especially in smaller markets. The advantage of using Shin probabilities and the differences between bookmakers decrease with an increasing market size.

© 2014 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

1. Introduction

There has been interest in the scientific literature in the accuracy of betting odds-based probability forecasts both directly, by comparing them to other sources of probability forecasts, and indirectly, through their use in betting strategies and as explanatory variables in statistical models. The probabilities from betting odds are also used in research into issues such as market efficiency and the competitive balance of sports competitions. For reviews, we refer the reader to Humphreys and Watanabe (2012), Stekler, Sendor, and Verlander (2010) and Vaughan Williams (2005).

The widespread use of betting odds is not surprising, as there is substantial empirical evidence that betting odds are the most accurate publicly-available source of probability forecasts for sports. With the growth of online betting, betting odds are also readily available for an increasing number and range of sports competitions. However, we believe that the following two issues with using betting odds as probability forecasts have not yet been addressed sufficiently:

(a) Which method should be used to determine probability forecasts from raw betting odds?
(b) Does it make a difference as to which bookmaker or betting exchange we choose, when two or more are available?

We address these two issues in the context of fixed-odds betting, with an emphasis on evaluating the most commonly used methods for determining probability forecasts from odds. Empirical evaluation is performed using data from several different online bookmakers across 37 competitions and five different team sports (basketball, handball, ice hockey, soccer, and volleyball).

1.1. Related work

As a matter of brevity and convenience, we focus on the most relevant results for fixed-odds betting, which is prevalent in team sports.[1]

The empirical evidence suggests that betting odds are the most accurate source of sports forecasts. Odds-based probability forecasts have been shown to be better than, or

* Tel.: +386 1 4768459.
E-mail address: erik.strumbelj@fri.uni-lj.si.

[1] We omit a substantial subset of the literature on racetrack betting that focuses primarily on parimutuel markets and their efficiency (see Hausch, Lo, & Ziemba, 2008, for a review).

http://dx.doi.org/10.1016/j.ijforecast.2014.02.008
0169-2070/© 2014 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

at least as good as, statistical models using sports-related input variables (Forrest, Goddard, & Simmons, 2005; Song, Boulier, & Stekler, 2007; Štrumbelj & Vračar, 2012), expert tipsters (Song et al., 2007; Spann & Skiera, 2009), and (aggregated) lay predictions (Pachur & Biele, 2007; Scheibehenne & Broder, 2007).

A special subset of betting odds are odds from betting exchanges. Unlike fixed-odds, which are formed by bookmakers, betting exchange odds are formed by bettors. That is, betting exchanges facilitate both backing and laying bets, and can be considered a form of prediction market.

In many different domains, forecasts from prediction markets are more accurate than those produced by traditional forecasting approaches and single forecasters (Arrow et al., 2008; Graefe & Armstrong, 2011; Tziralis & Tatsiopoulos, 2007). In sports forecasting, the term 'betting exchange' in most cases means Betfair, the world's largest betting exchange. There is substantial empirical evidence that the probabilities determined from Betfair odds are more accurate forecasts than those from fixed-odds bookmakers (Franck, Verbeek, & Nuesch, 2010; Smith, Paton, & Williams, 2009; Spann & Skiera, 2009; Štrumbelj & Vračar, 2012). Štrumbelj and Robnik-Šikonja (2010) also showed that there are significant differences between online fixed-odds bookmakers in terms of forecasting accuracies.

2. Determining outcome probabilities from betting odds

Fixed-odds bookmakers post betting odds, which indicate how much a bet placed with the bookmaker at that time would pay if it were to win. An online bookmaker posted the following betting odds for the 2012 Champions League Final match regular time outcome: Bayern Munich 1.80, Draw 3.75, Chelsea 4.33. Given that regular time ended in a 1:1 draw, we now know that for every unit we had bet on a draw, we would have won 3.75 units (2.75 + the unit we bet). Money bet on either of the other outcomes would have been lost.

The inverse odds are an indication of the bookmaker's underlying probabilistic beliefs. In our case, they suggest that Bayern had at most a $\frac{1}{1.80} = 0.56$ chance of winning, Chelsea 0.23, and that there was at most a 0.27 probability of a draw. However, bookmakers do not offer fair odds, so the sum of the inverse odds (also known as the *booksum*) will always be greater than 1 ($0.56 + 0.27 + 0.23 = 1.06$). In order to use the inverse odds as probability forecasts, we therefore have to account for the excess 6% (also known as the *bookmaker take* or bookmaker margin).

Most studies use basic normalization (dividing the inverse odds by the booksum).[2] In fact, this approach has become almost synonymous with the use of betting odds, although it is not clear whether bookmakers really do add their take proportionately across all possible outcomes. The widespread use of basic normalization can be attributed to its simplicity.

Alternatively, we can view the outcome as a categorical variable and model the probabilities using a historical data set of betting odds and corresponding match outcomes (see for example Forrest et al., 2005; Forrest & Simmons, 2002; Goddard, Beaumont, Simmons, & Forrest, 2005). Due to the categorical nature of the dependant variable, either logistic (probit) regression or multinomial regression is used, depending on the number of outcomes. An ordered model is preferred if there is a natural order to the outcomes.

There are only a few studies that have used an alternative to basic normalization or regression modeling. Smith et al. (2009) used a theoretical model of how bookmakers set their odds that was originally proposed by Shin (1993). Shin's model can be used to reverse-engineer the bookmaker's underlying probabilistic beliefs from the quoted betting odds. For earlier uses of Shin's model, see the works of Cain et al. (for example Cain, Law, & Peel, 2002, 2003; Smith et al., 2009, and references therein). They show that Shin's model-based approach improves on basic normalization. We adopt their term *Shin probabilities* to refer to probabilities determined from betting odds by using Shin's model.

Surprisingly, Shin probabilities have not been adopted widely, and the little use they have seen has focused almost exclusively on racetrack betting. A logical question that follows, therefore, is, can normalization based on Shin's model improve on basic normalization in individual and team sports?

2.1. Basic normalization

Let $\mathbf{o} = (o_1, o_2, \ldots, o_n)$ be the quoted decimal odds for a match with $n \geq 2$ possible outcomes, and let $o_i > 1$ for all $i = 1 \ldots n$. The inverse odds $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$, where $\pi_i = \frac{1}{o_i}$, can be used as latent team strength variables, but do not represent probabilities, because they sum to more than 1.

Let $\beta = \sum_{i=1}^{n} \pi_i$ be the booksum. Dividing by the booksum, $p_i = \frac{\pi_i}{\beta}$, we obtain a set of values that sum to 1 and can be interpreted as outcome probabilities. We refer to this as *basic normalization*.

2.2. Shin's model

Shin (1993) proposed a model which is based on the assumption that bookmakers quote odds which maximize their expected profit in the presence of uninformed bettors and a known proportion of insider traders.

The bookmaker and the uninformed bettors are assumed to share the probabilistic beliefs $p = (p_1, p_2, \ldots, p_n)$, while the insiders are assumed to know the actual outcome before the actual experiment (race, match, etc…). In sports, such superior information can be due either to a better aggregation of publicly available knowledge, or to private information, such as match-fixing.

Without loss of generality, we can assume that the total volume of bets is 1, of which $1 - z$ comes from uninformed bettors and z from insiders. Conditional on outcome i occurring, the expected volume bet on the ith outcome is $p_i(1 - z) + z$. If the bookmaker quotes $o_i = \frac{1}{\pi_i}$

[2] For our example, basic normalization gives Bayern $\frac{0.56}{1.06} = 0.528$, Chelsea 0.217, and Draw 0.255. Using Shin's model, we get 0.535, 0.215, and 0.250, respectively.

for outcome i, then the expected liability for that outcome is $\frac{1}{\pi_i}\,(p_i(1-z)+z)$.

By assuming that the bookmaker has probabilistic beliefs $\mathbf{p}$, we get the bookmaker's unconditional expected liabilities

$$\sum_{i=1}^{n} \frac{p_i}{\pi_i}\,(p_i(1-z)+z)\,,$$

and the total expected profit

$$T(\boldsymbol{\pi}) = 1 - \sum_{i=1}^{n} \frac{p_i}{\pi_i}\,(p_i(1-z)+z)\,.$$

The bookmaker sets $\boldsymbol{\pi}$ to maximize the expected profit, subject to the constraints $0 \leq \pi_i \leq 1$ and $\beta < \beta_{\max}$. This leads to

$$\pi_i = \frac{\beta_{\max}\sqrt{zp_i + (1-z)p_i^2}}{\sum_{j=1}^{n}\sqrt{zp_j + (1-z)p_j^2}}, \qquad (1)$$

and, taking into account the fact that, in economic equilibrium, the expected payoff of a bookmaker will be 0, we can determine $\beta_{\max}$ (see Shin, 1993), and get the following solution:

$$\pi_i = \sqrt{zp_i + (1-z)p_i^2}\sum_{j=1}^{n}\sqrt{zp_j + (1-z)p_j^2}. \qquad (2)$$

We have the reverse task of determining the probabilistic beliefs $\mathbf{p}$ given the quoted $\boldsymbol{\pi}$. Jullien and Salanié (1994) showed that Eq. (2) can be inverted to give the *Shin probabilities*

$$p_i = \frac{\sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\beta}} - z}{2(1-z)}. \qquad (3)$$

Then, all that remains is to compute the proportion of insider trading z. We can use the condition $\sum_{i=1}^{n} p_i = 1$ to obtain

$$z = \frac{\sum_{i=1}^{n}\sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\beta}} - 2}{n-2}, \qquad (4)$$

which can be solved using fixed-point iteration starting at $z_0 = 0$.

Note that as the booksum β approaches 1, the proportion of insider traders z goes to 0, reducing Shin's approach to basic normalization (see Eq. (3)).

In the special case of two possible outcomes ($n = 2$), Eq. (4) has a tractable analytical solution

$$z = \frac{(\pi_+ - 1)(\pi_-^2 - \pi_+)}{\pi_+(\pi_-^2 - 1)},$$

where $\pi_+ = \pi_1 + \pi_2$ and $\pi_- = \pi_1 - \pi_2$.

2.3. Regression analysis

An alternative is to use a statistical model to predict the outcome probabilities from odds. Unlike basic normalization and Shin probabilities, this approach requires a historical set of betting odds and match outcomes, which we can use to estimate the parameters of the model.

We use standard discrete choice models, with (inverse) betting odds as input variables. For sports with three outcomes (home [H], draw [D], away [A] in handball, ice hockey, and soccer), we use an ordered logistic regression model (Train, 2009, Chapter 7). For two outcomes (home [H] and away [A] win in basketball and volleyball), it simplifies to a logistic regression model (Train, 2009, Chapter 3).

3. Data set

We compiled a data set of sports matches with outcomes and betting odds. These include the Betfair betting exchange and the following major online fixed-odds bookmakers: bet.at.home, Bet365, Betclic, Betsafe, bwin, Betway, DOXXbet, Expekt, Interwetten, Jetbull, Ladbrokes, NordicBet, and Unibet. The betting odds in our data are final odds, recorded just before the start of the event.

A summary of the sports and competitions is found in Table 1. Note that not every league and/or game is covered by every bookmaker. Unless otherwise noted, all leagues are the top tier competitions in their respective countries.

4. Methodology

We compare four different methods for determining probabilities from betting odds: basic normalization (*Basic*), Shin probabilities (*Shin*), and two variants of the regression-based approach. The first variant of the regression-based approach (*Logit*) uses the first third of the data to estimate a model. The second variant (*LogitR*) updates the model in a rolling manner. That is, for every match, the probabilities are determined using a model estimated on the first third of the data and all matches from the evaluation part of the data that precede the current match. Note that we omit the π_A predictor from all regression models in order to avoid having linearly dependent predictors.

4.1. Evaluating probability forecasts

Let $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ be our probability estimates and $\mathbf{a}$ the vector indicating the actual outcome. We evaluate the forecasting accuracy of probabilities using two standard scores for categorical forecasts, the Ranked Probability Score (RPS) (Epstein, 1969) and the Brier score (Brier, 1950).

The Brier score of a single forecast is defined as

$$\mathrm{BRIER}(\pi, \mathbf{a}) = \frac{1}{n}\|(\pi - \mathbf{a})\|^2$$

and the RPS as

$$\mathrm{RPS}(\pi, \mathbf{a}) = \frac{1}{n}\|(C(\boldsymbol{\pi}) - C(\mathbf{a}))\|^2,$$

where $C(\mathbf{x}) = (c_1, c_2, \ldots, c_n), c_i = \sum_{j=1}^{i} x_i$ is the cumulative distribution.

RPS is preferred due to the ordinal outcomes of some sports (Home, Draw, Away), where it makes sense to look at the difference between the cumulative probabilities

Table 1
Sports competitions and games in our data set (for the period 2008/09–2011/12).

Sport	Competition	N_{matches}	Market size[a]
Basketball	Greek A1	694	27 402
	Italian Lega A	944	14 577
	Russian Superleague A	105	–
	Spanish ACB	884	24 011
	Turkish TBL	928	6 952
	NBA (USA)	4 657	150 684
Handball	Danish Jack Jones Ligaen	704	690
	French Division 1	621	48
	German Bundesliga	1 191	3 390
	Polish Ekstraklasa	505	533
	Portugese LPA[b]	366	–
	Spanish Liga Asobal	950	689
Hockey	Czech Extraliga	1 454	903
	Finish SM Liga	1 648	7 024
	German DEL	1 561	1 314
	Norvegian Eliteserien[b]	665	1 256
	Russian KHL	2 556	5 325
	Swedish Eliteserien	1 320	5 812
	Swiss NLA	1 200	271
	NHL (USA)	4 899	9 765
Soccer	Brazilian Campeonato[b]	1 140	150 221
	English Championship (2nd tier)	2 206	165 952
	English League 1 (3rd tier)	2 206	41 124
	English League 2 (4th tier)	2 208	29 732
	English Premier League	1 517	2 998 480
	French Ligue 1	1 510	222 938
	German Bundesliga	1 220	752 293
	Italian Serie A	1 519	902 854
	Dutch Erdedivisie	1 220	241 368
	Scottish Premier League	792	171 174
	Spanish Primera Division	1 518	1 299 918
	MLS (USA)	1 094	83 676
Volleyball	Belgian League[b]	251	–
	French Pro A	631	–
	German Bundesliga[b]	396	659
	Italian Seria A1[b]	572	3 397
	Polish Plusliga[b]	274	1 704
	Total	48 126	

[a] Median volume matched (in GBP) on Betfair on an event from this competition (match outcome bets only).
[b] Data only available from 2009/10 to 2011/12.

and the cumulative actual outcome (see Constantinou & Fenton, 2012, for a more detailed argument[3]). Note that the RPS simplifies to the Brier score for binary forecasts (basketball, volleyball).

The Brier score is utilized because the mean Brier score for a set of forecasts can be decomposed conveniently into calibration (bias) and resolution components, to provide further insights. We use a generalized decomposition of the Brier score for situations where the observations are stratified into bins of probabilities rather than directly on the issued probabilities (Stephenson, Coelho, & Jolliffe, 2008):

$$
\begin{aligned}
BS &= \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj}-o_{kj})^2 \\
&= \frac{1}{n}\sum_{k=1}^{m} n_k(\bar{f}_k-\bar{o}_k)^2 \\
&\quad - \frac{1}{n}\sum_{k=1}^{m} n_k(\bar{o}_k-\bar{o})^2 + \bar{o}(1-\bar{o}) \\
&\quad + \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj}-\bar{f}_k)^2 \\
&\quad - \frac{2}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj}-\bar{f}_k)(o_{kj}-\bar{o}_k), \qquad (5)
\end{aligned}
$$

where m is the number of bins, n_k the number of forecasts in the kth bin, n the total number of forecasts, f_{kj} the jth forecast in the kth bin and o_{kj} the observed outcome, $\bar{f}_k$ and $\bar{o}_k$ the within-bin averages, and $\bar{o}$ the overall relative frequency.

The five components in the decomposition of Eq. (5) are reliability (REL), resolution (RES), uncertainty (UNC), within-bin variance (WBV), and within-bin correlation (WBC), respectively. The generalized decomposition can be written as BS $=$ REL $-$ GRES $+$ UNC, where GRES $=$ REL $-$ WBV $+$ WBC is the generalized resolution.

The two within-bin terms help to compensate for the decrease in the resolution component when the bin size is increased, which makes the generalized decomposition less sensitive to the choice of bin size (Stephenson et al., 2008). In the special case when all forecasts within a bin are the same (that is, when the observations are stratified directly on the issued probabilities), the two within-bin terms vanish and we are left with the standard decomposition of the Brier score into Reliability, Resolution, and Uncertainty components (Murphy, 1973).

5. Results

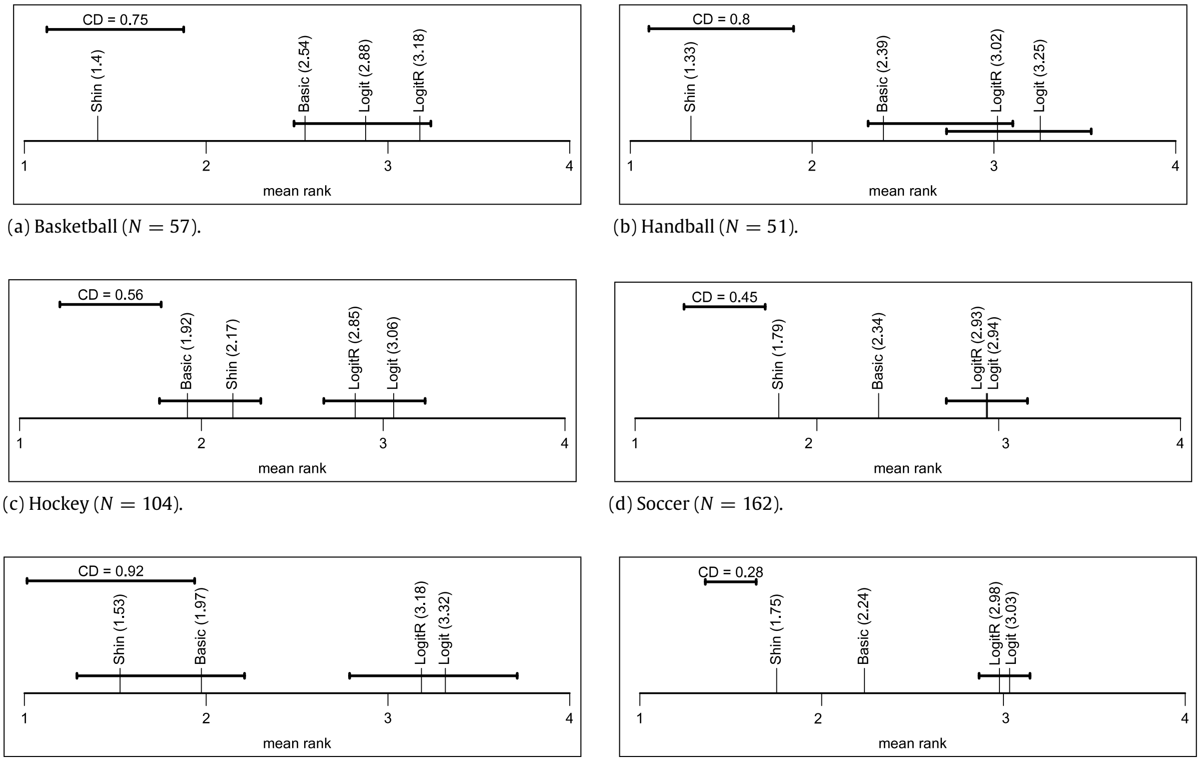
First, we computed the average RPS scores of the four methods for deriving probabilities from bookmaker odds. The scores were computed for each sport (and all sports combined) and each bookmaker/competition pair separately. We then used these RPS scores to rank the methods from 1 (best) to 4 (worst), for each bookmaker/competition pair. Such a non-parametric comparison is preferred because forecasting scores violate the assumption of normality.

The Shin probabilities were best for 217 of the 412 bookmaker/competition pairs, basic normalization for 103, Logit for 67, and LogitR for 25. On average, the Shin probabilities ranked 1.75, basic normalization 2.24, LogitR 2.98, and Logit 3.03. The average ranks for individual sports are given in parentheses in Fig. 1.

We used the Friedman test, a non-parametric equivalent of the one-way ANOVA (see for example Demšar, 2006), to test whether any of the methods rank significantly higher or lower than others. The null hypothesis of all methods having the same average rank was rejected for every sport and for all sports combined (all p-values $<$ 0.001).

To identify specific pairs of methods where one is better than the other, we performed a pair-wise comparison using the Nemenyi post-hoc test, thus adjusting for multiple

[3] Prior to Constantinou and Fenton, the RPS had also been used in the context of evaluating sports forecasts by Štrumbelj and Robnik-Šikonja (2010).



(a) Basketball ($N = 57$).

(b) Handball ($N = 51$).

(c) Hockey ($N = 104$).

(d) Soccer ($N = 162$).

(e) Volleyball ($N = 38$).

(f) All sports.

Fig. 1. Mean ranks of the methods' scores, broken down by sport and across all sports. Differences beyond the Nemenyi post-hoc test critical distance (CD) are statistically significant at the 0.01 level.

Table 2
Mean and median RPS scores and average improvement of probabilities when using Shin probabilities.

Bookmaker	Basic		Shin		Δ[a] (in %)
	Mean	Median	Mean	Median	
bet.at.home	0.2114	0.1857	0.2109	0.1818	0.59
bet365	0.2110	0.1837	0.2106	0.1820	0.35
Betclic	0.2116	0.1882	0.2110	0.1841	0.58
Betsafe	0.2112	0.1861	0.2107	0.1828	0.40
bwin	0.2113	0.1833	0.2109	0.1803	0.44
DOXXbet	0.2113	0.1895	0.2107	0.1858	0.57
Expekt	0.2115	0.1870	0.2110	0.1837	0.39
Interwetten	0.2123	0.1963	0.2113	0.1895	1.05
NordicBet	0.2112	0.1871	0.2107	0.1842	0.47
Betfair	0.2101	0.1833	0.2100	0.1810	0.38

[a] An indication of how much the probabilities obtained using basic normalization would have to be moved towards the actual outcomes in order to match the forecasting accuracy of the Shin probabilities. The improvement is small but consistent across all bookmakers.

comparisons. Under the null hypothesis of equal ranks, any pair of groups (methods) that differ in rank by more than the *critical distance*

$$CD = q_{k,\alpha}\sqrt{\frac{k(k-1)}{6N}}$$

are statistically significantly different. The critical value q depends on the number of groups k and the chosen significance level. For $k = 4$ and our chosen 0.01 significance level, $q_{4,0.01} = 3.11$ (see Demšar, 2006, for details).

The results for all pair-wise comparisons in a single Nemenyi test can be represented on a single axis, by plotting the average rank of each method and using the critical distance (CD) to indicate where the difference (distance) between two methods is not large enough to be statistically significant.

The results of our post-hoc comparison are summarized in Fig. 1. Shin probabilities are significantly better than basic normalization overall and for three sports (basketball, handball, and soccer). They are also better than basic normalization for volleyball, while basic normalization produces more accurate forecasts for hockey, but these two differences are not found to be significant. Furthermore, Shin probabilities are significantly better than both regression-based methods in all cases, and these in turn are not significantly different from each other for any of the sports.

5.1. *The size of the effect of using Shin probabilities*

The advantage of using Shin probabilities instead of basic probabilities does not depend on the choice of bookmaker (see Table 2 for all sports combined). Ladbrokes, Jetbull, Unibet, and Betway are excluded from the comparison, due to their limited coverage. The results are for matches covered by all listed bookmakers ($N = 33\,981$). Shin probabilities improved on basic normalization not

Table 3
Decomposition of the mean Brier scores of binned forecasts into reliability, generalized resolution, and uncertainty components.

Sport	Method	REL	GRES	UNC	BRIER
Basketball	Shin	3.985E−04	4.831E−02	2.387E−01	0.1908
	Basic	7.429E−04	4.827E−02		0.1912
	LogitR	1.703E−03	4.785E−02		0.1926
Handball	Shin	7.865E−04	5.451E−02	1.858E−01	0.1320
	Basic	1.179E−03	5.448E−02		0.1325
	LogitR	6.576E−04	5.324E−02		0.1332
Hockey	Shin	1.790E−04	7.675E−03	2.132E−01	0.2057
	Basic	1.041E−04	7.662E−03		0.2057
	LogitR	1.095E−04	7.275E−03		0.2061
Soccer	Shin	8.279E−05	1.434E−02	2.140E−01	0.1998
	Basic	1.776E−04	1.432E−02		0.1999
	LogitR	7.497E−05	1.377E−02		0.2003
Volleyball	Shin	8.128E−04	6.934E−02	2.462E−01	0.1776
	Basic	1.507E−03	6.931E−02		0.1784
	LogitR	4.181E−03	6.894E−02		0.1814

just for every bookmaker across all sports, but for every bookmaker/sport pair.[4]

5.2. Decomposition of forecasts and tests for calibration in the small

For each sport and outcome, we divided the forecasts across all matches and bookmakers into eight bins of equal size. The bins are based on the quantiles of the inverse betting odds for match and bookmaker. Therefore, for a given sport, the bins are same for all four methods.

The results of the decomposition are shown in Table 3. Shin probabilities have the best Brier scores for each sport, due to having the best (highest) generalized resolution (GRES). The reliability (calibration) components of the Shin probabilities' Brier scores are best for basketball and volleyball only. Again, this implies that Shin's model is not entirely accurate in its modelling of the way in which bookmakers set odds for matches where draws are possible.

In a search for additional insights, we check for calibration-in-the-small; that is, for potential bias of subsets of odds, and in particular, of individual odds (probability) subintervals. The results are summarized in the calibration plots in Fig. 2.

We can test the significance of these departures from calibration by observing that, under the null-hypothesis of perfect calibration, the standardized within-bin difference between the predicted and observed proportions will approximate a standard normal distribution (see for example Lahiri & Wang, 2013, for details). Here, the within-bin biases appear to be relatively small, though none of the methods are perfectly calibrated-in-the-small (all p values < 0.001).

The high uncertainty components (see Table 3) show that the actual outcome of a sports match is relatively difficult to forecast outright. The values are averaged across all three outcomes for handball, hockey, and soccer. Note that the uncertainty component $\overline{o}(1-\overline{o})$ in Eq. (5) depends only on the sport, and is therefore the same for all methods for a given sport.

5.3. Comparing bookmakers

The data used to obtain the results reported in Table 2 also facilitate a comparison of the forecasting accuracies of probabilities determined from different fixed-odds bookmakers and Betfair, while controlling for variance between sports and leagues.

Similarly to testing the significance of differences between methods, we first apply the Friedman test and then undertake post-hoc analysis using the Nemenyi test. Again, there are significant differences for all sports ($p < 0.001$ for all sports). The post-hoc pairwise comparison results are shown in Fig. 3. The procedure was performed for Shin probabilities and basic normalization separately.

When Shin probabilities are used to determine probabilities from odds, Betfair is the best source (see Fig. 3(f)), but it is not found to be significantly better than bet365. With the exception of soccer, the Betfair betting exchange is not the obvious best choice for any of the sports.

If basic normalization is used instead of Shin probabilities (Fig. 3, in gray below the axis), Betfair stands out as the significantly best source overall and for all five sports (tied with bwin for handball). When interpreting this result, we have to take into account the previous result that Shin probabilities are uniformly better across all bookmakers. That is, this discrepancy is a result, not of Shin probabilities being worse than basic normalization for determining probabilities from Betfair odds, but of basic normalization being less appropriate for determining odds from fixed-odds bookmakers.

5.4. Effects of market size

Some sports are more popular and attract a larger volume of bets. Larger markets should result in a higher level

[4] The mean (min) improvements are 6.54×10^{-4} (7.1×10^{-5}) and 5.44×10^{-3} (5.18×10^{-4}) for the mean and median RPS scores, respectively.

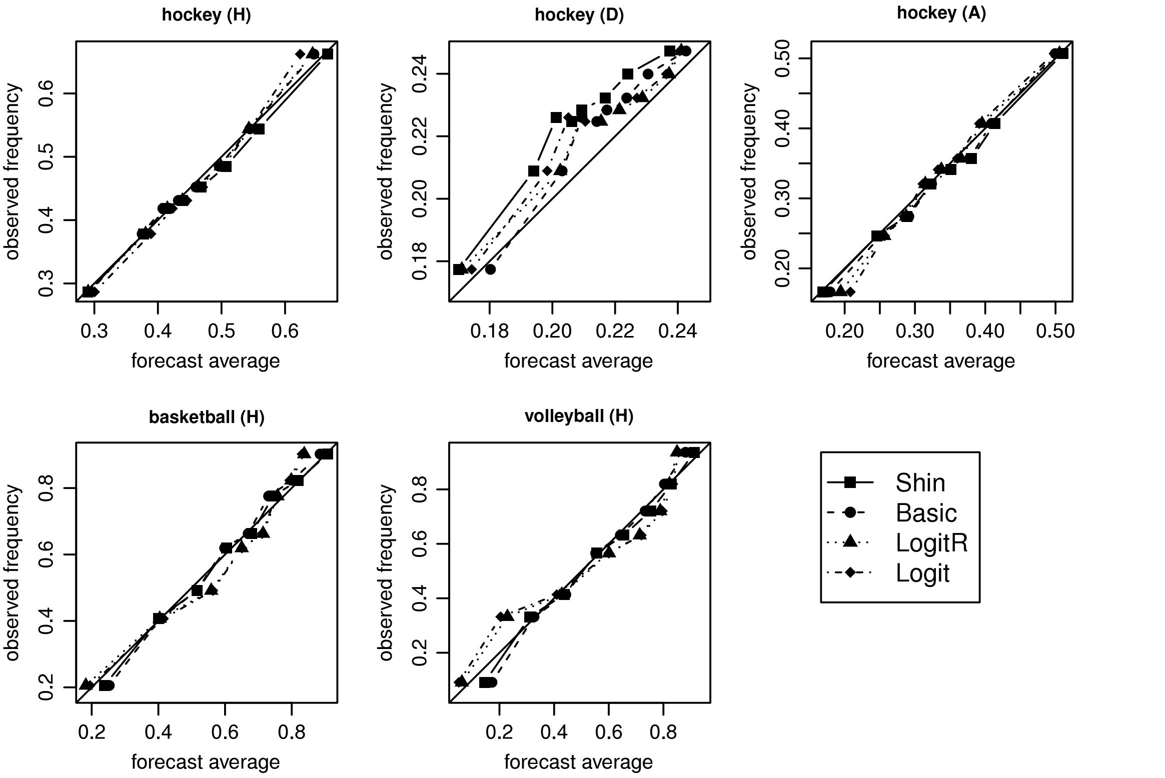


Fig. 2. Plots of observed relative frequencies against the average forecast for three sports. The results for soccer and handball are similar to those for hockey.

Table 4
Pearson correlation coefficients between log-market size (V), bookmaker take (T), difference (D) and absolute difference $|D|$ in median RPS between bookmaker and Betfair, and absolute difference (improvement) in median RPS between Shin probabilities and basic normalization (I).

Bookmaker	$R_{V,T}$	$R_{V,D}$	$R_{V,\|D\|}$	$R_{V,\|I\|}$
bet.at.home	−0.622**	−0.044	−0.332	−0.503*
bet365	−0.788**	0.294	−0.253	−0.511*
Betclic	−0.516**	0.199	−0.536*	−0.496*
Betsafe	−0.620**	0.194	−0.507*	−0.621**
bwin	−0.037	0.408*	−0.555*	−0.458*
DOXXbet	−0.588**	0.176	−0.514*	−0.618**
Expekt	−0.572**	0.122	−0.487*	−0.661**
Interwetten	−0.406*	−0.450*	−0.537*	−0.439*
NordicBet	−0.629**	0.040	−0.584**	−0.537*
Betfair	−0.790**			

* Significant at the 0.01 level.
** Significant at the 0.001 level.

of competition between bookmakers, thus driving them to lower their take and produce more accurate forecasts. Similarly, a larger volume should also make prediction markets more accurate. Therefore, it is reasonable to assume that the market size will affect how bookmakers compare.

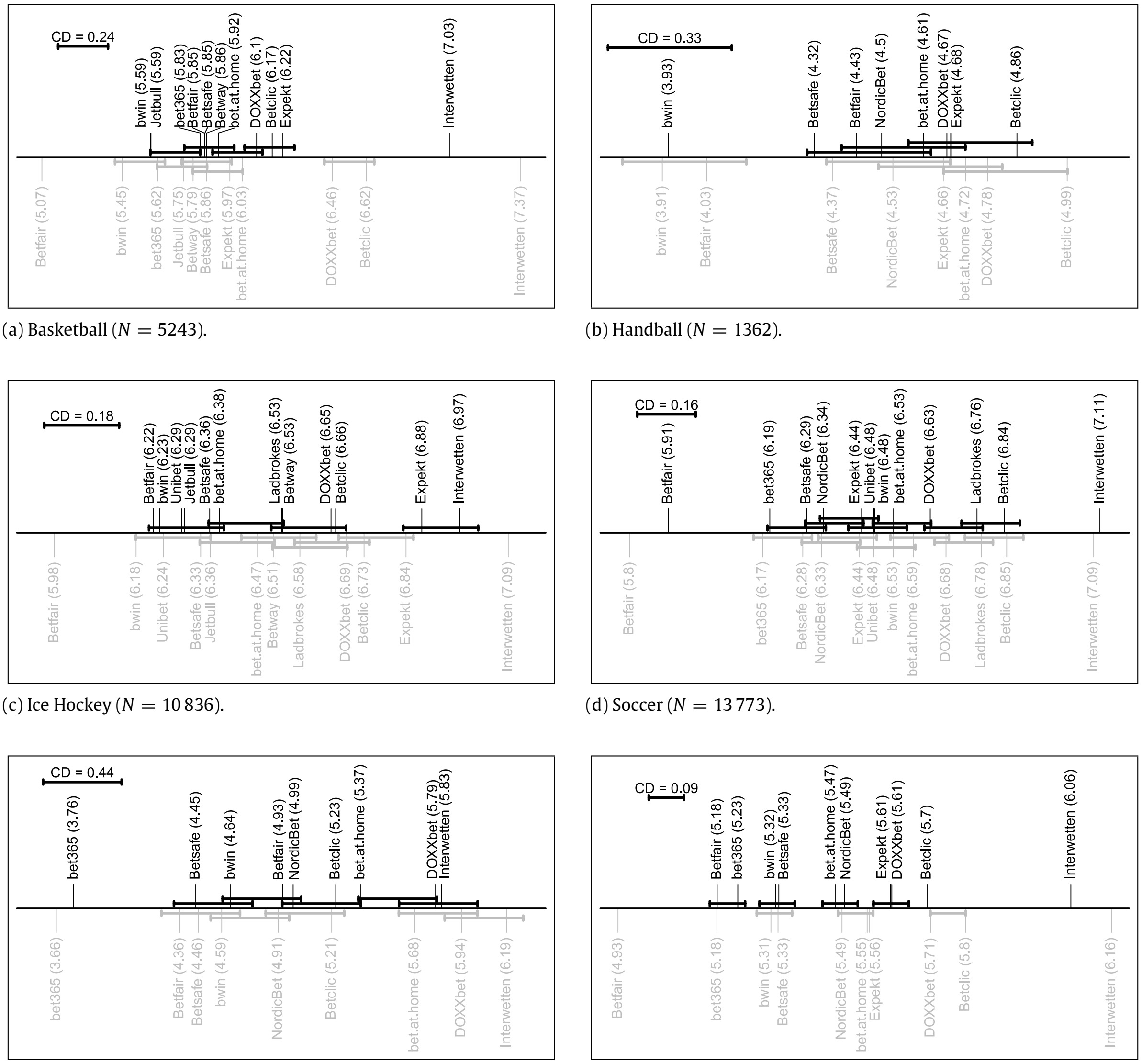
For each competition, we use the median volume of bets matched on Betfair as a proxy for the market size (see Table 1). Fig. 4 explores the relationship between market size (median total volume of bets matched) and bookmakers' takes (booksums) for two fixed-odds bookmakers and Betfair. The logarithm of market size is used, because it results in a better linear fit.

With the exception of bwin, the Pearson correlation coefficients are negative and significant for all bookmakers (see Table 4, first column). A substantial amount of the intra-bookmaker variability can be explained by the market size, but the case of bwin (see Fig. 4) illustrates how a bookmaker's take depends not only on the market size, but also on the sport. To partially account for that, we used all bookmaker/competition observations to fit the following linear model: $Booksum \sim \log(Volume) + Bookmaker + Sport$, where Bookmaker and Sport are dummy variables. We omit information about individual coefficients; the adjusted R^2 value of the model was 0.74 (p-value < 0.001).

Only a small proportion of the variability in the difference between a bookmaker and Betfair across different competitions (in terms of RPS scores) can be explained by the market size (see Table 4, second column). Of the two cases in which the correlation was found to be significant, the coefficient is positive in one and negative in the other (bwin, Interwetten) (see Fig. 5). When observing the absolute difference, the correlation coefficients are negative for all bookmakers (see Table 4, third column). As expected, the advantage of using Shin probabilities decreases with the market size (see Table 4, fourth column).

6. Conclusion

Our results show that Shin probabilities are better than basic normalization and regression model-based approaches for all bookmaker/sport pairs. Therefore, whenever the goal is to maximize the forecasting accuracy, Shin probabilities should be considered. The regression



(a) Basketball ($N = 5243$).

(b) Handball ($N = 1362$).

(c) Ice Hockey ($N = 10\,836$).

(d) Soccer ($N = 13\,773$).

(e) Volleyball ($N = 989$).

(f) All sports ($N = 33\,981$).

Fig. 3. Mean ranks of bookmakers' scores broken down by sport. Differences beyond the critical distance (CD) are significant at the 0.01 level. The results in black are for Shin probabilities, those in gray are for basic normalization.

models produce forecasts that are calibrated but are of poor resolution, and should not be used. Basic normalization, due to its simplicity, might still be preferred in applications where the forecasting accuracy is not crucial.

For all methods, the probabilities depart from calibration-in-the-small. The biases are relatively small, but a systematic underestimation of the probability of a draw is observed in sports with draws (soccer and especially hockey and handball). This suggests that none of the methods fully capture the way in which bookmakers set odds for sports with a draw outcome. It would be worth investigating whether this due to a draw being more difficult to predict (it is less likely that someone will have inside information) or to some other process, such as bookmakers sometimes offering more favorable odds on purpose (Franck, Verbeek, & Nüesch, 2013).

Our results show that, on average, two of the largest fixed-odds bookmakers (bwin, Bet365) and the world's largest betting exchange, Betfair, are sources of the best probability forecasts, while the fixed-odds bookmaker Interwetten is the worst. Overall, the differences between Betfair and the bookmakers (and between Shin probabilities and basic normalization) decrease with market size. In the larger markets (soccer), Betfair stands out as the best source, while other bookmakers are better sources of probability forecasts in the smaller markets. Therefore, in line with the forecasting literature, no single forecaster is best in all markets. The market size also accounts for most of the variation in bookmaker takes.

Related work comparing bookmakers and betting exchanges (Betfair) has found that Betfair is significantly better for NBA basketball than most fixed-odds bookmakers, but not all (bet356 was close second) (Štrumbelj &

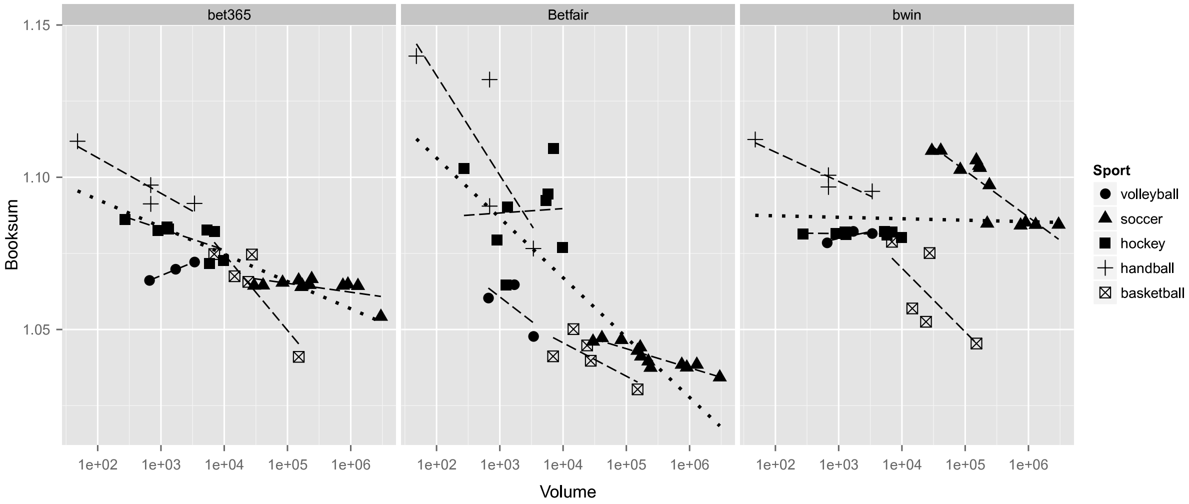


Fig. 4. The relationship between the market size and booksums. Each point represents one competition. Least-squares linear fits are included for each sport (dashed) and for all sports combined (dotted).

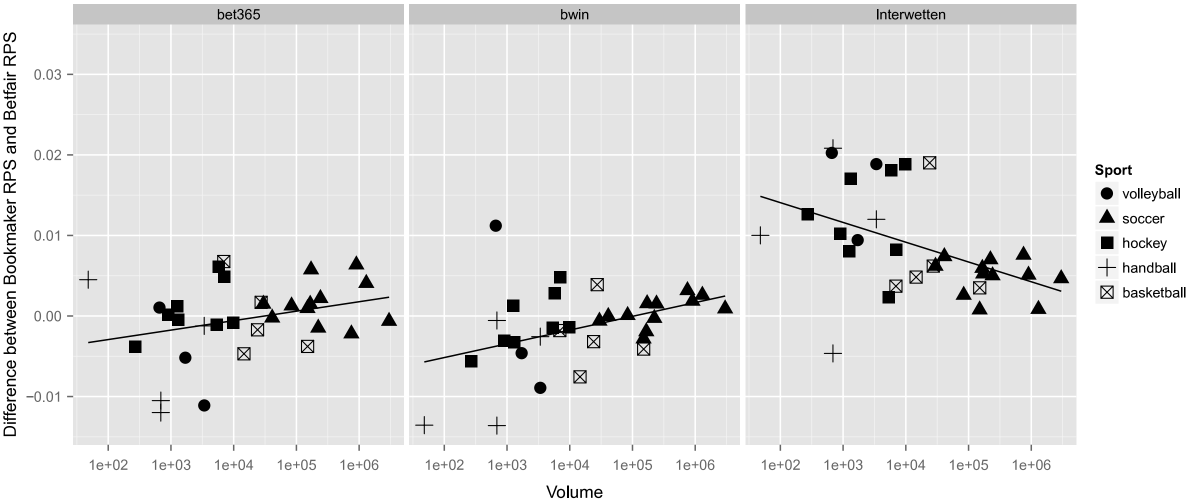


Fig. 5. The relationship between the market size and the difference in forecasting quality between each bookmaker and Betfair. Each point represents one competition. Least-squares linear fits are included.

Vračar, 2012). In top European soccer, Betfair was found to be the best source (Franck et al., 2010). In German top-tier soccer, Betfair was found to be slightly better than the fixed-odds bookmaker (Spann & Skiera, 2009). Our results agree with these findings, which is to be expected for these larger markets, where the differences between basic normalization and Shin probabilities are smaller.

References

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., et al. (2008). The promise of prediction markets. *Science*, *320*(5878), 877–878.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *75*, 1–3.

Cain, M., Law, D., & Peel, D. (2002). Is one price enough to value a state-contingent asset correctly? Evidence from a gambling market. *Applied Financial Economics*, *12*(1), 33–38.

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

Constantinou, A. C., & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, *8*(1).

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Epstein, E. S. (1969). A scoring system for probability forecast of ranked categories. *Journal of Applied Meteorology*, *8*, 985–987.

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

Forrest, D., & Simmons, R. (2002). Outcome uncertainty and attendance demand in sports: the case of English soccer. *Journal of the Royal Statistical Society, Series D (The Statistician)*, *51*, 229–241.

Franck, E., Verbeek, E., & Nüesch, S. (2013). Inter-market arbitrage in betting. *Economica*, *80*(318), 300–325.

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

Goddard, J., Beaumont, J., Simmons, R., & Forrest, D. (2005). Home advantage and the debate on competitive balance in professional sports leagues. *Journal of Sports Sciences*, *23*(4), 439–445.

Graefe, A., & Armstrong, J. S. (2011). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting*, *27*(1), 183–195.

Hausch, D. B., Lo, V. S., & Ziemba, W. T. (Eds.) (2008). *Efficiency of racetrack betting markets*. World Scientific.

Humphreys, B. R., & Watanabe, N. M. (2012). *The Oxford handbook of sports economics: the economics of sports, vol. 1 (Ch. Competitive balance)*. Oxford University Press.

Jullien, B., & Salanié, B. (1994). Measuring the incidence of insider trading: a comment on Shin. *Economics Journal*, *104*(427), 1418–1419.

Lahiri, K., & Wang, J. G. (2013). Evaluating probability forecasts for GDP declines using alternative methodologies. *International Journal of Forecasting*, *29*(1), 175–190.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595–600.

Pachur, T., & Biele, G. (2007). Forecasting from ignorance: the use and usefulness of recognition in lay predictions of sports events. *Acta Psychologica*, *125*, 99–116.

Scheibehenne, B., & Broder, A. (2007). Predicting Wimbledon 2005 tennis results by mere player name recognition. *International Journal of Forecasting*, *23*(3), 415–426.

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

Song, C., Boulier, B. L., & Stekler, H. O. (2007). The comparative accuracy of judgmental and model forecasts of American football games. *International Journal of Forecasting*, *23*(3), 405–413.

Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, *28*(1), 55–72.

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

Stephenson, D. B., Coelho, C. A., & Jolliffe, I. T. (2008). Two extra components in the Brier score decomposition. *Weather and Forecasting*, *23*(4), 752–757.

Štrumbelj, E., & Robnik-Šikonja, M. (2010). Online bookmakers' odds as forecasts: the case of European soccer leagues. *International Journal of Forecasting*, *26*(3), 482–488.

Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, *28*(2), 532–542.

Train, K. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Tziralis, G., & Tsatsiopoulos, I. (2007). Prediction markets: an extended literature review. *The Journal of Prediction Markets*, *1*(1), 75–91.

Vaughan Williams, L. (Ed.) (2005). *Information efficiency in financial and betting markets*. Cambridge University Press.

Erik Štrumbelj obtained his B.Sc. and Ph.D. from the University of Ljubljana, Faculty of Computer and Information Science. He is currently an assistant professor at the University of Ljubljana. His research interests include probability, statistics, and machine learning.