Evaluating the Predictive Accuracy of Association Football Forecasting Systems

A. Constantinou[1], N.E. Fenton

Department of Electronic Engineering and Computer Science,
Queen Mary, University of London, UK

Abstract

Despite the increasing importance and popularity of association football forecasting systems there is no agreed method of evaluating their accuracy. We have classified the evaluators used into two broad categories: those which consider only the prediction for the observed outcome; and those which consider the predictions for the unobserved as well as observed outcome. We highlight fundamental inconsistencies between them and demonstrate that they produce wildly different conclusions about the accuracy of four different forecasting systems (Fink Tank/Castrol Predictor, Bet365, Odds Wizard, and pi-football) based on recent Premier league data. None of the existing evaluators satisfy a set of simple theoretical benchmark criteria. Hence, it is dangerous to assume that any existing evaluator can adequately assess the performance of football forecasting systems and, until evaluators are developed that address all the benchmark criteria, it is best to use multiple types of predictive evaluators (preferably based on posterior validation).

Keywords: scoring rules, probability forecasting, football predictions, sports forecasting

---

1. Introduction

Evaluating the predictive accuracy of any forecasting system is a critical part of its validation. Even within a narrow application domain the nature of forecasting systems can vary greatly, so there are no universally accepted predictive evaluator methods. In the absence of an agreed method of evaluation, it is difficult to reach a consensus about which of two or more competing systems is 'best'. Moreover, given the potential of different evaluators to be highly sensitive to very small differences between forecasting systems, the choice of evaluator is clearly critical in determining the success or otherwise of any 'validation'.

Association football (referred to henceforth simply as *football*) is the world's most popular spectator sport and it constitutes an ever-increasing share of sports gambling worldwide (Dunning, 1999). Consequently, football forecasting systems have also attracted increasing attention. For the purpose of this paper we assume a football forecasting system is one that provides outputs in the form of three probability values (p_H, p_D, p_A) corresponding to the three possible outcomes of any given match - home win, draw, and away win. Systems like FinkTank/Castrol Predictor (Castrol Football, 1999) and the various online betting websites (Bet365, 2001) that make their predictions publicly available, are typical examples of popular forecasting systems.

The need to evaluate the predictive accuracy of football forecasting systems is evident. Given the simplicity of the outputs of such systems, it is not unreasonable to expect there to be an agreed satisfactory evaluator. Yet surprisingly, as we will show in this paper there is none. That is not to say that there is a shortage of different evaluators. We will review the 10 that have been proposed and used and will highlight the problems associated with them. The fundamental concern is lack of consistency whereby one evaluator might conclude that system α is more accurate than system β, whereas another may conclude the opposite. Given the critical need to 'validate' such systems it is clear that the selection of the evaluator can be as important as the development of the forecasting system itself. Otherwise outstanding systems might be erroneously rejected while poor systems might be erroneously judged as outstanding.

To motivate our concerns with some very simple examples, consider the predictions provided by systems α and β for matches 1, 2 and 3 presented in Table 1.

[1] Corresponding author.
E-mail addresses: constantinou@eecs.qmul.ac.uk (A. Constantinou), norman@eecs.qmul.ac.uk (N. Fenton)

Match	System	P(H)	P(D)	P(A)	Result
1	α	0.6	0.2	0.2	H
	β	0.7	0.2	0.1	
2	α	0.5	0.45	0.05	H
	β	0.5	0.05	0.45	
3	α	0.35	0.30	0.35	D
	β	0.6	0.30	0.10	

Table 1. Hypothetical predictions by systems α and β for matches 1, 2 and 3.

These examples raise the following issues, which we regard as benchmark criteria for any evaluator:

i. **Predicted value of the observed outcome**: In match 1 both α and β 'correctly' assign the highest probability to the winning outcome (H). But should we also have to take into consideration the fact that the value for β is higher, and therefore intuitively, more accurate?

ii. **Predictions of unobserved probabilities**: For match 2, both systems α and β again 'correctly' assign the highest probability to the winning outcome (H). Moreover, not only do they assign the same probability to the winning outcome (both are 0.5), they also assign the same probability values – albeit in a different order - to the unobserved outcomes (0.45 and 0.05). So should we conclude that α and β have identical accuracy? If we did it would contradict the intuition of any football fan who would confirm that α is more accurate than β since a draw is 'closer' to a home win than an away win.

iii. **Distribution of unobserved probabilities**: For match 3, both systems predicted the draw with probability 0.3. Do we consider the predictions equal or are we supposed to take into consideration the distribution of the unobserved probabilities as well? Intuitively α is superior because its distribution of probabilities is far more indicative of a draw than is β, which strongly predicts a home win. Formally, if we regard the match result as being on a scale from 0 to 1 (with Home win at the 0 end and Away at the 1 end) then system α generated a distribution of high variance with a mean of 0.5, whereas system β generated a distribution of lower variance with mean pointing towards the home win.

iv. **Goal difference**: Although the outcomes are presented simply as H, D or A, can we really afford to ignore the actual final score? If the home team wins 5-0 does this make us more certain that model β is superior to α on match 1? And if the score was 4-3 does this lead us to conclude that maybe α was more accurate?

v. **Red card effect**: What if one or more red cards have dramatically impacted the outcome? For example, in match 3 if the home team had a player sent off near the beginning of the match does that imply that model β is superior to α after all? There might be cases in which a bad prediction will look better because of the red card effect (Vecer, Kopriva, & Ichiba, 2009). Is this especially important when evaluating a small number of matches?

To help structure our review of the various evaluators found in the literature, we have classified them into two broad categories: those which consider only the prediction for the observed outcome (e.g. informational loss); and those which consider the predictions for the unobserved as well as observed outcome (e.g. Brier score); In Section 2 we present a detailed overview of the evaluators in the context of this classification. In Section 3 we describe the four different forecasting systems and match results that were used in the analysis of the evaluators. Section 4 presents the results, which demonstrate spectacular inconsistency between evaluators. The discussion in Section 5 summarises our concerns over the results and the rationality of each of the evaluators.

The paper provides a number of novel contributions:

- The first comprehensive review of evaluators of football forecasting systems.
- A useful classification of evaluators and a simple unifying way to describe how they are defined.
- A comprehensive empirical study that applies all of the evaluators to four very different forecasting systems using the most recent data from the UK Premiership.
- Full online access to the data including all of the model predictions.
- Conclusive evidence that none of the existing evaluators can be reliably used to judge the performance of football forecasting systems (hence casting doubt on all previous evaluations).
- A set of simple benchmarking criteria that evaluators need to satisfy
- Constructive guidelines on how to develop improved evaluators.

2. Overview of the proposed evaluators

We start with some definitions and assumptions. Football forecasting is a special case of any forecasting problem in which there is a (fixed) finite number *k* of outcomes for each of a sequence of problem instances. A forecasting system is assumed to assign probability values to each of the outcomes in each problem instance, i.e. a vector:

$$p = (p_1, p_2, \dots, p_k)$$

where p_i is the probability of the ith outcome predicted by the forecasting system. We will also assume that, for each problem instance, we know the corresponding vector of actual observed outcome probabilities

$$e = (e_1, e_2, \dots, e_k)$$

For football forecasting we make some especially simple assumptions:

- The outcomes are restricted to the elementary outcomes $\{H, D, A\}$, so $k = 3$. We do not consider here non-elementary outcomes like "H or D", "A or D" (even though these are important in football betting). Because of this assumption the vector of actual observed outcome 'probabilities' will always be of the form $(1,0,0)$ in the case of H, $(0,1,0)$ in the case of D, or $(0,0,1)$ in the case of A. Although some of the researchers in our review have constructed models that generate predictions about the final score, those researchers have used the scores to generate predictions for the three possible outcomes as the basis for evaluation. Whether this is appropriate or not is out of the scope of this paper. Our concern relies only on the integrity and consistency of evaluators of the predictions, and not on how the predictions are generated. The same applies for cases in which the probabilities were given by experts instead of a model.
- The problem instances are matches between an identifiable home team and identifiable away team. Because of the universal importance of home advantage in football matches, and the shortage of data on matches played at neutral venues, we do not consider matches played at neutral venues as being within scope here.

By an ***evaluator*** we mean a method that specifies for any forecasting system α:

1. *An accuracy score* $f_i(\alpha)$ for each individual problem instance i; and
2. *A cumulative accuracy score* $C(\alpha)$, i.e. a means of combining the accuracy scores of a set of n problem instances $f_1(\alpha), f_2(\alpha), \dots, f_n(\alpha)$

It should be noted that, in some of the evaluators we consider, the individual accuracy score is a special case of the cumulative score and so it is sufficient to define the latter. With these assumptions an evaluator is able to compare two forecasting systems α and β simply by comparing their respective cumulative accuracy scores over a set of problem instances.

To give a simple example of the above assumptions, if we wanted to use the standard Bayesian model assessment approach (Open University Course Team, 2007) as the basis for an evaluator we would define:

1. *Accuracy score* $f_i(\alpha)$ is the probability value a model α assigns to the actual outcome. So, considering the models α and β in Table 1 we get:

$$\begin{array}{ll} f_1(\alpha) = 0.6, & f_1(\beta) = 0.7 \\ f_2(\alpha) = 0.5, & f_2(\beta) = 0.5 \\ f_3(\alpha) = 0.3, & f_3(\beta) = 0.3 \end{array}$$

   The score in this case is just the probability of observing the data assuming the model is correct; this is called the likelihood. By Bayes theorem, if we have no prior reason to assume one model is more correct then another, then the model that results in the highest likelihood is the one most likely to be 'correct'. So, on the basis of match 1, β is deemed to be more accurate than α, since it has a higher probability of being the 'correct' model. For matches 2 and 3 the models are of course indistinguishable.

2. *Cumulative accuracy score* $C(\alpha)$ is the product of the individual accuracy scores, so $C(\alpha) = 0.09$ $C(\beta) = 0.105$. Since individual matches are assumed to be independent, it also follows from Bayes that the probability of observing a set of match results given a model is correct, is simply the product of the individual likelihoods.

It is surprising that, given the importance of the chosen evaluator in validating and comparing the performance of football forecasting systems, we have found no published articles that focus on the issue of evaluators. As described below, there has been little consistency in which evaluators are used. Although most studies have chosen a single evaluator, some of the researchers, notably (Hvattum & Arntzen, 2010), have identified the potential problems that popular evaluators could create, and this led them into proposing more than one evaluator.

With the above assumptions we have classified the evaluators into two categories:

1. Those which consider only the prediction for the observed outcomes (Section 2.1).

2. Those which consider the predictions for the unobserved as well as observed outcomes Section 2.2)

2.1 Evaluators which consider only observed outcomes

The evaluators in this category may all be viewed as some kind of variant of the Bayesian evaluator described in the above example, since they take into consideration only the assigned likelihood of the observed outcomes. Table 2 provides a summary of the evaluators in this category applied to the hypothetical predictions and results in Table 1. In all of the evaluators here we assume that, for an individual match i the number p_i is the probability value that the forecaster assigns to the actual outcome. So for example, in match 1, p_1 is 0.6 for model α as p_1 is 0.7 for model β.

- **Geometric mean**

  For an individual match i the score is simply p_i.

  Cumulative score:

$$\left(\prod_{i=1}^{n} p_i\right)^{\frac{1}{n}}$$

  This can be interpreted as a normalised pseudo-likelihood measure. We have found three cases where the geometric mean has been used as the definitive single evaluator:

  - Rue and Salvesen (2000) used it to compare their own model predictions with the odds provided by the bookmaker Intertops (Rue & Salvesen, 2000).
  - Graham and Stott (2008) used it to compare their proposed ordered probit model, with the predictions provided by the bookmaker William Hill (Graham & Stott, 2008).
  - John Goddard (2005) used it to compare the one forecasting model that depended on goals scored and conceded, and another that predicted only the outcomes H, D and A.

- **Informational loss**

  This is defined in (Witten & Frank, 2005):

  For an individual match:

$$-\log_2 p_i$$

  Cumulative score:

$$\sum_{i=1}^{n} -log_2 p_i$$

  The informational loss generates identical assessment with that of the negative log-likelihood function which is optimised by logistic regression (Witten & Frank, 2005).

  Hvattum and Arntzen (2010) have proposed informational loss as one of their six evaluators for assessing the performance of their two proposed models. Both of their models were based on the ELO rating system, and were used to derive covariates that were then used in ordered probit regression models. The resulting performance was then compared against other economical measures and statistical tests (Hvattum & Arntzen, 2010).

  In order to avoid the *zero-frequency problem* (whereby the information loss is minus infinity when any p_i is zero) Witten and Frank (2005) suggest that non-zero probabilities have to be assigned to every outcome. Since none of the predicted outcomes is logically false, this is sensible since assigning a probability of zero to any of the outcomes contradicts *Cromwell's rule* (Lindley, 1985). However, in cases where a model really does predict a zero probability outcome there are simple and plausible solutions, such as the standard *Laplace estimator* technique (see (Laplace, 1951) for a translation of Laplace's original paper).

- **Maximum Likelihood and Maximum Log-Likelihood Estimations**

  Proposed by Sir Ronald Aylmer Fisher (see e.g. (Aldrich, 1997)) as an approach to parameter estimation and inference in statistics, Maximum Likelihood Estimation (MLE) states that the desired probability distribution is the one that makes the observed data "most likely". As in most cases, we are

presenting the Maximum Log-likelihood estimation over the MLE since it generates identical behaviour while reducing the computational costs significantly (Myung, 2003). For a likelihood test, the Binomial distribution has been widely used in order to tabulate the different parameter values.

For an individual match:

$$ln\frac{n!}{h!\,(n-h)!}\,p_i^h(1-p_i)^{n-h}$$

Cumulative score:

$$\sum_{\mathrm{i}=1}^{\mathrm{j}} ln\frac{n!}{h!\,(n-h)!}\,p_i^h(1-p_i)^{n-h}$$

where p_i represents the probability for observing the specified outcome, n represents the total number of trials, and h represents the number of successes for the given prediction. Consequently, for a given observation there is always a MLE and thus, the closer to p, the better the prediction is considered. However, since we only have one observation for every match prediction (n and h are always equal to 1) then there is no need in calculating the likelihood. Thus, the likelihood always equals p_i and hence, the Log-likelihood $ln(p_i)$. As a result, the informational loss generates identical assessment with that of the negative Log-likelihood function which is optimised by logistic regression (Witten & Frank, 2005). Therefore, in order to avoid unnecessary calculations in this paper we simply introduce

For an individual match:

$$ln\,(p_i)$$

Cumulative score:

$$\sum_{i=1}^{n} ln\,(p_i)$$

MLE has been used as the basis for evaluating football forecasting systems as follows:

- Hirotsu and Wright (2003) used MLE it to examine the use of Markov process models for evaluating the characteristics of the English Premier League (EPL) teams by means of parameters representing home advantage, offensive and defensive strength, and their interactions.
- Forrest Goddard and Simmons (2005), used it as one of their evaluators to compare the effectiveness of forecasts based on published bookmaker odds and the forecasts made using a benchmark statistical model which incorporated a large number of quantifiable variables, as well as the experts' views which were represented by published odds (Forrest, Goddard, & Simmons, 2005).
- Karlis and Ntzoufras (2003) used the Maximum Log-likelihood ratio for a bivariate Poisson distribution model analysing sports data. They also proposed the Bayesian Information Criterion (BIC) developed by Gideon E. Schwarz (1978) and its closely related Akaike Information Criterion (AIC), which was modified by Akaike (1977) for model fitting. In this paper, we ignore both the BIC and AIC since they also take into consideration the number of parameters for each model. We simply do not have this information for the selected forecasting systems under assessment.
- Graham and Stott (2008), who have already been mentioned for making use of the geometric mean function, have also used the Maximum Log-likelihood as part of their predictive evaluation described above.

- **Pair-wise comparison evaluator**

Although it does not strictly fit into our definition of an evaluator (since it only makes sense for comparing two forecasting systems) we include in this classification the simple pair-wise comparison used by Dixon and Pope (2004).

For two systems α and β this is defined as: for an individual match assign a score of 1 to the model which has the highest probability value assigned to the actual outcome (otherwise 0, including if there is a tie).

The cumulative score is simply the sum for each model of the individual matches and can be presented as follows for two given systems α and β:

$$C(a)=\sum_{i=1}^{n}\begin{cases}1, & p_{ai}>p_{\beta i}\\ 0, & \text{otherwise}\end{cases},\qquad C(\beta)=\sum_{i=1}^{n}\begin{cases}1, & p_{\beta i}>p_{ai}\\ 0, & \text{otherwise}\end{cases}$$

Using the example of Table 1, the overall score for the 3 matches results in system α equal to 0 and system β equal to 1 (since system β 'wins' match 1 and matches 2 and 3 are tied), as demonstrated in table 2. Dixon and Pope (2004) used this method to evaluate the predictions of two bookmakers.

Evaluator	**Formula** (as expressed in section 2.1)
Geometric Mean	e.g. when $match = 1$ and $system = \alpha$ then: $$f_1(a) = (\prod_{i=1}^{n} p_i)^{\frac{1}{n}} = (0.6)^{\frac{1}{1}} = 0.6$$ Cumulative: $$C(a) = \left(\prod_{i=1}^{n} p_i\right)^{\frac{1}{n}} = (0.6 \times 0.5 \times 0.3)^{\frac{1}{3}} = \mathbf{0.4481}$$
	Cumulative: $$C(\beta) = \left(\prod_{i=1}^{n} p_i\right)^{\frac{1}{n}} = (0.7 \times 0.5 \times 0.3)^{\frac{1}{3}} = \mathbf{0.4717}$$
Informational Loss	e.g. when $match = 1$ and $system = \alpha$ then: $$f_1(a) = -\log_2 p_i = -\log_2 0.6 = 0.7379$$ Cumulative: $$C(a) = \sum_{i=1}^{n} -\log_2 p_i = \left[[-\log_2 0.6] + [-\log_2 0.5] + [-\log_2 0.3]\right] = \mathbf{3.4739}$$
	$$C(\beta) = \sum_{i=1}^{n} -\log_2 p_i = \left[[-\log_2 0.7] + [-\log_2 0.5] + [-\log_2 0.3]\right] = \mathbf{3.2515}$$
Maximum Log-likelihood Estimation	e.g. when $match = 1$ and $system = \alpha$ then: $$f_1(a) = ln\,(p_1) = ln(0.6) = -0.5108$$ Cumulative: $$C(a) = \sum_{i=1}^{j} ln\,(p_i) = [-0.5108] + [-0.6931] + [-1.2039] = \mathbf{-2.4079}$$
	$$C(\beta) = \sum_{i=1}^{j} ln\,(p_i) = [-0.3566] + [-0.6931] + [-1.2039] = \mathbf{-2.2537}$$
Pair-wise comparison	e.g. when $match = 1$, $system = \alpha$ and $team = h$ then: $$f_1(a) = \begin{cases} 1, & p_{a1} > p_{\beta 1} \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 1, & 0.6 > 0.7 \\ 0, & \text{otherwise} \end{cases} = 0$$ Cumulative: $$C(a) = \sum_{i=1}^{n} \begin{cases} 1, & p_{ai} > p_{\beta} \\ 0, & \text{otherwise} \end{cases} = 0 + 0 + 0 = \mathbf{0}$$
	$$C(\beta) = \sum_{i=1}^{n} \begin{cases} 1, & p_{ai} > p_{\beta} \\ 0, & \text{otherwise} \end{cases} = 1 + 0 + 0 = \mathbf{1}$$

Table 2. The demonstration (4d.p.) of evaluators which consider only observed outcomes (Geometric Mean, Informational Loss, Maximum Log-Likelihood Estimation, pair-wise comparison) according to the examples presented in table 1 for systems α and β. For the Geometric mean and Pair-wise comparison, a higher score indicates better performance, whereas for the Informational Loss and Maximum Log-likelihood Estimation an error closer to 0 indicates better performance.

2.2 Evaluators that consider both observed and unobserved outcomes

In all of the evaluators in this section we assume that, for an individual match i $(p_{i1,}\, p_{i2,}\, p_{i3})$ is the vector of predicted probability values that the forecaster assigns to (H, D, A) and $(e_{i1,}\, e_{i2,}\, e_{i3,})$ is the actual outcome (so each e_{ij} is either 0 or 1). Table 3 shows the evaluators applied to the hypothetical predictions and results in Table 1.

- **Brier score**

  Also known as Quadratic Loss (Hvattum & Arntzen, 2010), the Brier Score measures the average squared deviation between predicted probabilities for a set of their events and their outcomes. For an individual match i the score is (Brier, 1950)

  $$\sum_{j=1}^{e}(p_{ij} - e_{ij})^2$$

  The cumulative score is:

  $$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{e}(p_{ij} - e_{ij})^2$$

  Forrest, Goddard, & Simmons (2005) used the Brier score as one of their evaluators in the study described above. Hvattum and Arntzen (2010), who have already been mentioned for making use of the Informational loss function, have also proposed the Brier Score as part of their predictive evaluation described above.

- **Binary decision**

  This is a very simple evaluator defined for a single match i as:

  1 if one of p_{i1}, p_{i2}, p_{i3} is greater than the others and corresponds to the actual outcome e

  0 otherwise

  The cumulative score is simply the sum of the scores for single matches i

  Joseph, Fenton and Neil (2006) used this evaluator to compare the performance of their proposed expert Bayesian network model against the predictive performance generated by other well known machine learning techniques. Their model was explicitly developed to handle the predictions of Tottenham Hotspur football club.

- **Overall posterior validation**

  This is actually a class of evaluators that compare the predicted performance of specific teams over a number of matches with their actual performance or involve a comparison of posterior summary statistics. For an individual match i we simply compute the expected number of points for each team. So, assuming 3 points for a win and 1 for a draw we get:

  $$f_{ih}(a) = 3 \times p_{i1} + p_{i2}$$

  $$f_{iu}(a) = p_{i2} + 3 \times p_{i3}$$

  for a single match i where $h = homeTeam$ and $u = awayTeam$. For example, for match 1 of Table 1 model α scores 2 for the home team $(3 \times 0.6 + 0.2)$ and 0.8 for the away team $(0.2 + 3 \times 0.2)$, while model β scores 2.3 for the home team $(3 \times 0.7 + 0.2)$ and 0.5 for the away team $(0.2 + 3 \times 0.1)$.

  For the cumulative-point score we simply add the individual scores for each team and also determine the team rankings based on each team's score to be used by the evaluators RMS and Relative rank error. This 'predicted' ranking can then be compared with the actual ranking of the team based on actual points using one of the following proposed metrics:

  $$Relative\ rank\ error \sum_{teams} \frac{\left|rank_{actual} - rank_{predicted}\right|}{rank_{actual}}$$

$$RMS\ error = \sqrt{\sum_{teams} \frac{\left(rank_{actual} - rank_{predicted}\right)^2}{rank_{actual}}}$$

The following studies have used one or more of these approaches:

- Baio and Blangiardo (2010) compares a simple application of Bayesian hierarchical modelling against the bivariate Poisson structure model that was proposed by Karlis and Ntzoufras (2003).
- Byungho et al (Min, Kim, Choe, Eom, & McKay, 2008) used it to evaluate their proposed model, described as FRES, using Bayesian inference and rule-based reasoning along with an in-game time series approach in an aim to predict the results of World Cup matches.
- Dixon and Pope (2004) extended the analysis of a previous study (Pope & Peel, 1988), in addition to using a Poisson distribution model derived from another study (Dixon & Coles, 1997) in an attempt to test the efficiency of the odds provided by the betting market with respect to that model. Their evaluation and comparison of the models was focused on the summary statistics retrieved by each of the models. Particularly, they have taken into consideration the mean, median and standard deviation of the probable outcomes of a home win, a draw and an away win. Additionally, Dixon and Pope (2004) have presented a histogram comparison demonstrating the frequency of the specified odds and probability estimates, including comparisons of Kernel density estimated of aggregated match probabilities for models, and odds from bookmakers.

Evaluator	Formula (as expressed in section 2.2)
Brier Score	e.g. when $match = 1$ and $system = \alpha$ then: $$f_1(a) = \sum_{j=1}^{3}(p_{ij} - e_{ij})^2 = (0.6-1)^2 + (0.2-0)^2 + (0.2-0)^2 = 0.24$$ Cumulative: $$C(a) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{3}(p_{ij} - e_{ij})^2 = \begin{bmatrix}[(0.60-1)^2 + (0.20-0)^2 + (0.20-0)^2] \\ +[(0.50-1)^2 + (0.45-0)^2 + (0.05-0)^2] \\ +[(0.30-1)^2 + (0.35-0)^2 + (0.35-0)^2]\end{bmatrix}\frac{1}{3} = \mathbf{0.4766}$$
	$$C(\beta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{3}(p_{ij} - e_{ij})^2 = \begin{bmatrix}[(0.70-1)^2 + (0.20-0)^2 + (0.10-0)^2] \\ +[(0.50-1)^2 + (0.05-0)^2 + (0.45-0)^2] \\ +[(0.30-1)^2 + (0.60-0)^2 + (0.10-0)^2]\end{bmatrix}\frac{1}{3} = \mathbf{0.485}$$
Binary Decision	e.g. when $match = 1$ and $system = \alpha$ then: 1, if one of p_{i1}, p_{i2}, p_{i3} is greater than the others and corresponds to the actual outcome e 0, otherwise p_{11}corresponds to outcome e (home win, 60%) p_{12}, p_{23}correspond to the unobserved outcomes (draw 20% and away win, 20%) $$0.60 > 0.20 \text{ and } 0.60 > 0.20, \text{therefore } f_1(a) = 1$$ Cumulative: $$C(a) = 1 + 1 + 0 = \mathbf{2}$$
	Cumulative: $$C(\beta) = 1 + 1 + 0 = \mathbf{2}$$
Overall Posterior Validation (Cumulative Points)	e.g. when $match = 1$, $system = \alpha$ and $team = h$ then: $$f_{1h}(a) = (3 \times p_{11}) + (p_{12}) = (3 \times 0.60) + (0.20) = 2p.$$ Cumulative: $$C_h(a) = (3 \times 0.60) + (0.20) + (3 \times 0.50) + (0.45) + (3 \times 0.35) + (0.30) = \mathbf{5.30}\boldsymbol{p}.$$ $Actual\ points\ =\ 7$ thus, $error = \lvert 5.30 - 7 \rvert = 1.7$ $$C_u(a) = (3 \times 0.20) + (0.20) + (3 \times 0.05) + (0.45) + (3 \times 0.35) + (0.30) = \mathbf{2.75}\boldsymbol{p}.$$ $Actual\ points\ =\ 1$ thus, $error = \lvert 2.75 - 1 \rvert = 1.75$ where $h = homeTeam$ and $u = awayTeam$
	$$C_h(\beta) = (3 \times 0.70) + (0.20) + (3 \times 0.50) + (0.05) + (3 \times 0.60) + (0.30) = \mathbf{5.95}\boldsymbol{p}.$$ $Actual\ points\ =\ 7$ thus, $error = \lvert 5.95 - 7 \rvert = 1.05$ $$C_u(\beta) = (3 \times 0.10) + (0.20) + (3 \times 0.45) + (0.05) + (3 \times 0.10) + (0.30) = \mathbf{2.50}\boldsymbol{p}.$$ $Actual\ points\ =\ 1$ thus, $error = \lvert 2.50 - 1 \rvert = 1.5$ where $h = homeTeam$ and $u = awayTeam$

Table 3. Demonstration of evaluators Brier Score, Binary Decision and posterior cumulative-point comparison according to the examples presented in table 1. For the Brier Score a lower score indicates better performance, whereas for the Binary Decision a higher score indicates better performance. For the posterior cumulative-point comparison a lower error between expected and actual points indicates a better performance.

3. Data

In order to demonstrate the behaviour and the results of the proposed evaluators, it was important to find predictions that come from diverse approaches to football forecasting. Diverse approaches are more likely to generate predictions with greater variance between them, thus serving as better test data for the evaluators. Based on diversity and also availability of relevant data, we have chosen four forecasting systems:

- *Fink Tank* (also presented as *Castrol Predictor*) is highly respected in the UK and its predictions are published weekly by *The Times* and the *CastrolRankings*
- *Bet365* is representative of probabilities given by the betting industry
- *Odds Wizard* is an industrial professional based forecasting software where users have to pay in order to retrieve any live predictions, or predictions before the matches are played.
- *pi-football*, which is a recent Bayesian network based model (Probabilistic Intelligence in Football, 2010) incorporating data and expert judgement.

The data used is for the first 221 matches of the English Premier League (EPL) during season 2010-2011. The reason for this choice is because at the time of writing this paper, the 2010-2011 season was still in progress and the full relevant data for previous seasons did not exist. Nevertheless, 221 predictions by each of the forecasting systems is more than enough information in order to demonstrate potential inconsistencies between the proposed evaluators. The full data used in the study is presented at www.pi-football.com

4. Results

All of the evaluators presented in section 2 have been used to evaluate the predictions generated from each of the four forecasting systems discussed in section 3. Table 4 presents the results.

For each evaluator we have computed:

- *Overall score* (as defined in Section 2) for each of the four forecasting systems
- *Relative performance* of the four forecasting systems. To calculate this, the system with the best overall score is given the value of 0%, and each subsequent system is given the proportional increase in error from the model ranked first. For example, for the Brier score, BET365 is best (it has the lowest value 0.6227) so it gets 0%, while Odds Wizard with value 0.6247 is 0.33% worse and so gets assigned the value 0.33%. Fink Tank with value 0.6289 is 1.01% worse and so is assigned the value 1.01%.
- *Ranking*. This is simply the ranking of the four systems based on the overall score (the system with the best overall score is ranked 1 etc.)

The results clearly demonstrate the inconsistency between the evaluators. There is not a single system of the four on which the evaluators agree on the ranking. For example, the pi-football system varies from best to worst, being ranked best four times and worst twice out of 9 distinct evaluators. The results demonstrate the deeply worrying extent to which our confidence in the performance in a forecasting system is wholly at the mercy of the choice of particular evaluator.

If the four systems produced very similar predictions then it could be argued that differences in the evaluator rankings might be due to random noise. But it is easy to show that the systems produce ***very different*** predictions:

- Figure 1 demonstrates the posterior predictive validation of cumulative points generated by each of the systems, as well as the actual cumulative points for each of the teams. For the teams Manchester City, Aston Villa, Everton, Newcastle, Sunderland, Bolton, West Brom, Blackpool, Wolves and Wigan, a very important variation in performance is observed between at least two of the forecasting systems for each case.
- Table 5 presents the posterior predictive validation of the outcomes H, D and A, illustrating the overall inconsistency between the forecasting systems. The largest variation appears to occur for the outcomes of a draw, particularly with pi-football and Fink Tank models with posterior probabilities of 27.12% and 22.53% respectively; almost up to an incredible difference of 5% of the total rate.
- Table 6 presents the actual and estimated league table between the forecasting systems, illustrating the significant differences in team-ranking. The most important variation regarding high quality teams has to be Manchester City, with systems ranking the team at positions 3, 4, 5 and 6. A much greater variation is observed for lower quality teams. For example, Sunderland had been estimated at positions 8, 9, 11 and 17, a staggering 9 positions in variation, as well as West Brom at positions 10, 14, 15 and 18.

Evaluators	pi-football	BET365	Fink Tank	Odds Wizard
Type 1: (section 2.1)				
1) Geometric Mean	0.3504	0.3540	0.3516	0.3537
... score (higher better)	1%	0%	0.66%	0.09%
... relative performance	**4**	**1**	**3**	**2**
... ranking				
2a) Informational Loss (or MLE)	334.28	331.07	333.17	331.36
... score (lower better)	0.97%	0%	0.63%	0.09%
... relative performance	**4**	**1**	**3**	**2**
... ranking				
2b) Maximum Log-likelihood Estimation	-231.71	-229.48	-230.94	-229.68
... score (higher better)	0.97%	0%	0.63%	0.09%
... relative performance	**4**	**1**	**3**	**2**
... ranking				
3) Pair-wise Comparison Score	2.38	2.47	2.73	2.39
... score (lower better)	0%	3.79%	14.58%	0.19%
... relative performance	**1**	**3**	**4**	**2**
... ranking				
Type 2: (section 2.2)				
4) Brier Score	0.6256	0.6227	0.6289	0.6247
... score (lower better)	0.48%	0%	1.01%	0.33%
... relative performance	**3**	**1**	**4**	**2**
... ranking				
5) Binary Decision	0.4841	0.4932	0.4660	0.4751
... score (higher better)	1.87%	0%	5.50%	3.67%
... relative performance	**2**	**1**	**4**	**3**
... ranking				
6) Overall Posterior Validation	3.96	3.86	5.06	4.51
... score (lower better)	3%	0%	31.22%	16.90%
... relative performance	**2**	**1**	**4**	**3**
... ranking				
7) Relative Rank Error	7.75	8.03	13.14	9.38
... score (lower better)	0%	3.58%	69.53%	20.99%
... relative performance	**1**	**2**	**4**	**3**
... ranking				
8) Root Mean Squared Error (RMS)	5.81	5.89	8.62	7.09
... score (lower better)	0%	1.41%	48.29%	22.04%
... relative performance	**1**	**2**	**4**	**3**
... ranking				
9) Summary Stats – PD Posterior Error	7.29	10.13	16.47	12.37
... score (lower better)	0%	38.85%	125.73%	69.54%
... relative performance	**1**	**2**	**4**	**3**
... ranking				

Table 4. Score/Error (4d.p.) according to the specified evaluator for models pi-football, Bet365, Fink Tank and Odds Wizard. Evaluators 2a and 2b provide identical assessment. Data represents the first 221 observed results from the English Premier League matches, season 2010/2011.

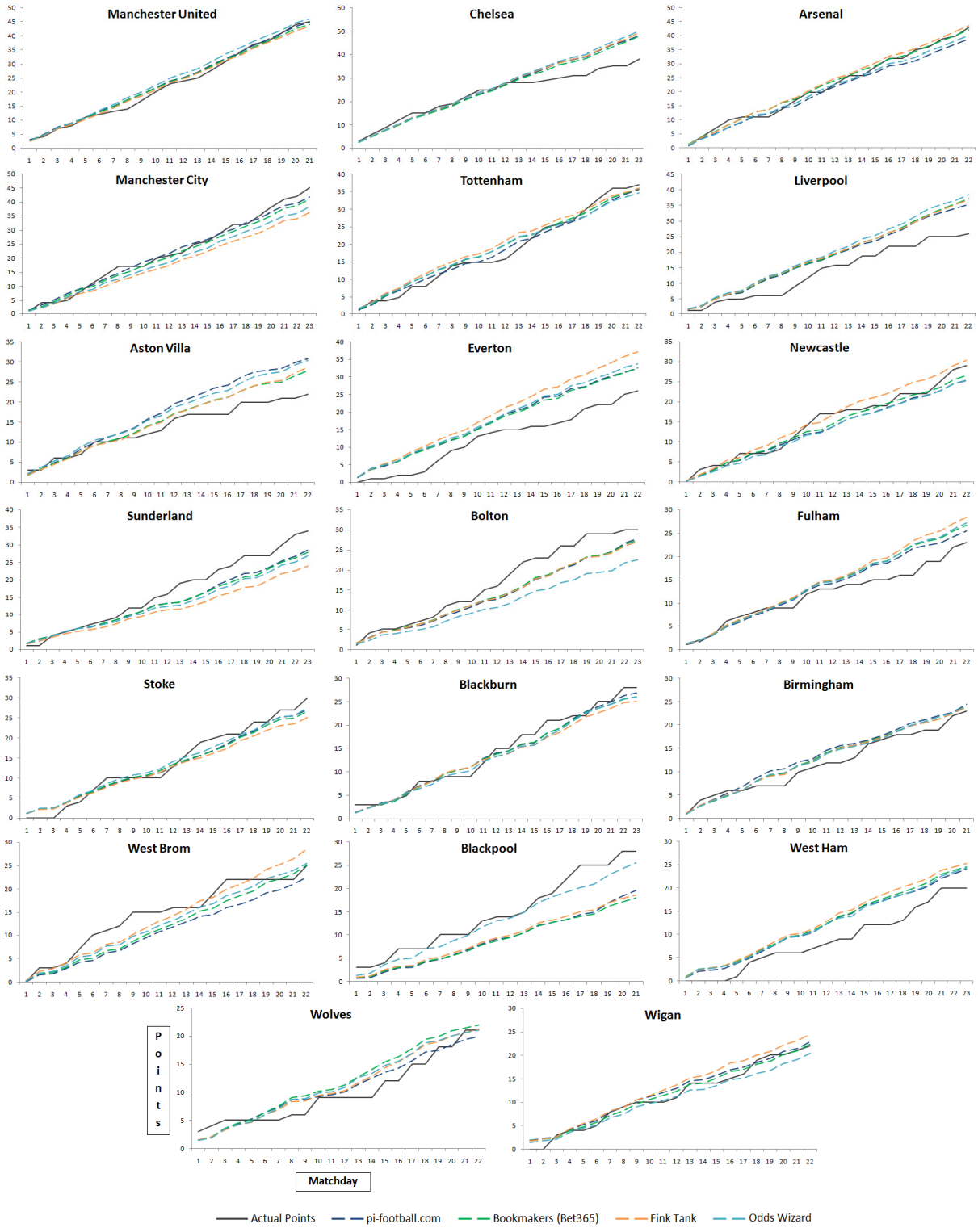

Figure 1. Overall Posterior validation of cumulative points expected by the each of the forecasting systems; pi-football, Fink Tank/Castrol predictor, Bet365 and Odds Wizard. Observed and estimated cumulative points for the first 221 matches of the English Premier League, season 2010/2011 (colour version online).

	Observed	pi-football (posterior)	Bet365 (posterior)	Fink Tank (posterior)	Odds Wizard (posterior)
Home Win	45.24	46.56	45.61	47.62	48.94
Draw	30.76	27.12	25.70	22.53	24.58
Away Win	23.98	26.31	28.67	29.83	26.46
Error – Home Win		1.31	0.36	2.38	3.69
Error – Draw	-	3.64	5.06	8.23	6.18
Error – Away Win		2.33	4.69	5.85	2.48
Total error	**-**	**7.29**	**10.13**	**16.47**	**12.37**

Table 5. Posterior predictive validation (2d.p.) of the outcomes home win, draw and away win. Observed and estimated outcomes for the first 221 matches of the English Premier League, season 2010/2011.

	Observed		pi-football (Posterior)		Bet365 (Posterior)		Fink Tank (Posterior)		Odds Wizard (Posterior)	
1	Man United	45	Chelsea	48.12	Chelsea	47.93	Chelsea	49.48	Chelsea	50.21
2	Man City	45	Man United	44.85	Man United	44.07	Arsenal	43.43	Man United	46.03
3	Arsenal	43	Man City	41.91	Arsenal	42.07	Man United	43.33	Arsenal	40.09
4	Chelsea	38	Arsenal	38.87	Man City	41.02	Everton	37.23	Liverpool	38.42
5	Tottenham	37	Tottenham	35.73	Liverpool	37.24	Liverpool	36.78	Man City	38.33
6	Sunderland	34	Liverpool	35.26	Tottenham	36.08	Man City	36.36	Tottenham	34.75
7	Bolton	30	Everton	32.63	Everton	32.50	Tottenham	35.99	Everton	33.69
8	Stoke	30	Aston Villa	30.81	Sunderland	27.87	Newcastle	30.29	Aston Villa	30.40
9	Newcastle	29	Sunderland	28.46	Aston Villa	27.76	Aston Villa	28.73	Stoke	27.58
10	Blackpool	28	Bolton	27.87	Bolton	27.47	West Brom	28.49	Fulham	27.14
11	Blackburn	28	Stoke	27.18	Fulham	26.68	Fulham	28.40	Sunderland	26.96
12	Everton	26	Blackburn	26.93	Stoke	26.65	Bolton	27.14	Blackburn	25.99
13	Liverpool	26	Fulham	25.51	Newcastle	26.61	West Ham	25.31	Blackpool	25.59
14	West Brom	25	Newcastle	25.34	Blackburn	26.03	Stoke	25.06	West Brom	25.50
15	Fulham	23	Birmingham	24.45	West Brom	25.08	Blackburn	25.03	Newcastle	25.40
16	Birmingham	23	West Ham	24.04	West Ham	24.56	Wigan	24.51	West Ham	24.31
17	Aston Villa	22	Wigan	22.87	Birmingham	24	Sunderland	23.89	Birmingham	24.22
18	Wigan	22	West Brom	22.58	Wigan	22.47	Birmingham	23.78	Bolton	22.56
19	Wolves	21	Wolves	19.96	Wolves	21.94	Wolves	21.29	Wolves	21.01
20	West Ham	20	Blackpool	19.61	Blackpool	18.05	Blackpool	18.59	Wigan	20.39

Table 6. Posterior predictive validation (2.d.p) of the models pi-football, Fink Tank, Bet365 and Odds Wizard. Observed and estimated league table for the first 221 matches, season 2010/2011.

5. Discussion

The empirical results suggest that none of the existing evaluators can be trusted to accurately assess the performance of a football forecasting model. In fact, returning to the examples in Table 1 it is easy to see why theoretically the evaluators are unsatisfactory, by reviewing how the evaluators address the benchmark issues that we already highlighted:

Predicted value of the observed outcome: Any evaluator that cannot identify system β as superior to α in match 1 clearly fails to preserve intuitive notions of accuracy. This rules out *binary decision* as an acceptable evaluator (all of the others preserve intuition here).

Consideration of the probabilities of the unobserved outcomes: The viewpoint that only the predicted probability of the observed outcome matters is best summed up by Ian Witten and Eibe Frank (2005) who stated (in defence of the *informational loss* evaluator) that

> "if you're gambling on a particular event coming up, and it does, who cares how you distributed the remainder of your money among the other events?"

It appears that much of the confusion here relates to the particular sport of the forecasting system under evaluation. Obviously in sports for which there are just two outcomes, we only need to consider the probability of the observed outcome since the other is automatically determined from it. This includes, for example, the US National Football League (NFL) where draws are eliminated. In such a situation, the problems of finding an 'optimal' evaluator may be feasible. For example, by ignoring the confidence of any given prediction, (Boulier & Stekler, 2003) showed that the Brier score performed optimally where they measured the predictive accuracy of the outcomes of NFL games for the 1994-2000 using power scores. Boulier and Stekler compared the generated forecasts from probit regression based on power scores published in The New York Times with those of a naive model, the betting market, and the opinions of the sports editor of The New York Times.

But Witten and Frank's viewpoint is also relevant to any sport in which we are trying to predict which one, from a set of competitors, wins. In particular, this covers any type of racing event.

However, the three outcomes of a football match are definitely ***not*** like the outcomes of a three horse race. In the event of a home win, a prediction of a draw is not equally as poor as a prediction of an away win, since a draw is 'closer' to being a home win than an away win is to a home win. So, in the case of match 2 Table 1, any rational evaluator should determine that α is more accurate than β despite the fact that they assign the same probability to the winning outcome (0.5). To see why, consider a person placing a double-chance 1X bet (home team not to lose). A bookmaker who used system α would have pay less than that of system β. This is because system α considers that there is only 5% probability for the home team to lose the match, as opposed to system β which considers a 45% probability for the same scenario.

This rules out, as potentially satisfactory evaluators, every evaluator that considers only the observed outcomes. However, what makes the match 2 predictions interesting is that the unobserved probabilities are both 0.45 and 0.05 in both cases – ***but they are in a different order***. It follows that an evaluator that does consider the unobserved outcomes ***but is indifferent to the order of these*** (such as Brier Score) is also inherently unsatisfactory (the latter evaluators will only be satisfactory if the unobserved outcomes are evenly distributed).

This rules out every one of the proposed evaluators except those based on posterior validation. Because posterior validation is based on comparing expected points with actual points, it correctly determines that α is more accurate than β. Moreover, unlike any of the other evaluators, posterior validation is also able to correctly determine that in match 3 α is more accurate than β since its distribution is 'closer' to the actual outcome (draw).

However, approaches based on overall posterior validation are far from being optimal in assessment. That is to say, comparisons based on these should be interpreted with care as they highlight the overall performance of the specified models rather than the distinct performance between matches. As an example, a difference of total points between observed and estimated outcomes during a posterior validation of either cumulative points or team-ranking provides inadequate information regarding the distinct predictive performance between matches and is thus inappropriate for devising ultimate conclusions.

Also, none of the evaluators can account for the issues of goal difference and red card effects raised in Section 1.

6. Concluding Remarks

We have provided an analysis of the various evaluators that have been used to assess the performance of association football forecasting systems. Given the massive surge in popularity of the sport, and its increasing dominance of sports betting internationally, the need for widely accepted evaluators in this domain is evident. Moreover, given that the outputs of any football forecasting system are very simple (in this paper we have assumed that the outputs are simply three probability values corresponding to the three possible outcomes of any given match - home win, draw, and away win) it is not unreasonable to expect there to be an agreed satisfactory evaluator. Yet surprisingly, there is no such agreed method. We classified the various evaluators (10 in total) into two broad categories: those (such as the geometric mean and informational loss) which consider only the prediction for the observed outcome; and those (such as the Brier score and overall posterior validation) which consider the predictions for the unobserved as well as observed outcome.

We have demonstrated both empirically and also theoretically that the existing evaluators are all inadequate to more or lesser degrees. In our empirical study we used the evaluators to assess the accuracy of four different forecasting systems (Fink Tank/Castrol Predictor, Bet365, Odds Wizard, and pi-football) based on data for the first 221 matches of the 2010-2011 Premier league. We found fundamental inconsistencies between the evaluators and demonstrated that they produce wildly different conclusions about the accuracy of the four forecasting systems. What is quite shocking is the extent to which the four systems in the study were ranked differently by the different evaluators. For example, one forecasting system pi-football, was ranked 3rd or 4th by all of the common simple evaluators, but was ranked 1st by most of the posterior validation based evaluators.

From the theoretical perspective we have provided some simple benchmark validity criteria that need to be satisfied by any reasonable evaluator. Many of the existing evaluators fail to satisfy the most basic of these criteria. Only those methods based on posterior validation come close to satisfying the most important criteria.

With the relentless increase in interest in football forecasting it will become more important than ever that effective evaluators of forecasting systems are used. We have highlighted the areas for improvement and recommend that our benchmark criteria should be used as a basic validity check for any new proposed evaluator. In the mean time, in the absence of a satisfactory evaluator, we recommend that multiple types of predictive evaluators (preferably based on posterior validation) be used in order to avoid most seriously erroneous conclusions.

Acknowledgements

We would like to thank the Engineering and Physical Sciences Research Council (EPSRC) for funding this research, Martin Neil for his assistance and Agena Ltd for software support.

References

Akaike, H. (1977). On entropy maximazation principle. *Applications in Statistics* , 27-41.

Aldrich, J. (1997). R.A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science* , 12: 3, 162-176.

Baio, D., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics.* , 37: 2, 253-264.

Bet365. (2001). Retrieved February 27, 2011, from Bet365: http://www.bet365.com

Boulier, B. L., & Stekler, H. O. (2003). Predicting the outcomes of National Football League games. *International Journal of forecasting.* , 19: 257-270.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* , 78: 1-3.

Castrol Football. (1999). Retrieved February 24, 2011, from Castrol Football: http://cn.castrolfootball.com/

Dixon, M. J., & Pope, P. F. (2004). The value of statistical forecasts in the UK association football betting market. *International journal of forecasting* , 20, 697- 711.

Dixon, M., & Coles, S. (1997). Modelling association football scores and inefficienies in the football betting market. *Applied Statistics* , 46, 265-80.

Dunning, E. (1999). *Sport Matters: Sociological Studies of Sport, Violence and Civilisation.* London: Routledge.

Forrest, D., Goddard, J., & Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International journal of forecasting* , 21, 551-564.

Goddard, J. (2005). Regression models for forcasting goals and match results in association football. *International Journal of forecasting.* , 21: 331-340.

Graham, I., & Stott, H. (2008). Predicting bookmaker odds and efficiency for UK football. *Applied Ecomonics* , 40, 99-109.

Hirotsu, N., & Wright, M. (2003). An evaluation of characteristics of teams in association football by using a Markov process model. *Journal of the Royal Statistical Society. Series D (The Statistician)* , 52: 4, 591-602.

Hvattum, L., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of forecasting.* , 26, 460-470.

Joseph, A., Fenton, N., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge Based Systems* .

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician* , 52: 3, 381-393.

Laplace, P. (1951). *A Philosophical Essay on Probabilities, translated from the 6th French edition by Frederick Wilson Truscott and Frederick Lincoln Emory.* New York: Dover Publications.

Lindley, D. (1985). *Making Decisions.* London, UK: John Wiley. Second Edition.

Min, B., Kim, J., Choe, C., Eom, H., & McKay, R. (2008). A compound framework for sports results prediction: A football case study. *Knowledge-Bases Systems* , 21, 551-562.

Myung, I. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* , 47, 90-100.

Open University Course Team. (2007). *Bayesian Statistics.* The Open University.

Pope, P., & Peel, D. (1988). Information, prices and efficiency in a fixed-odds betting market. *Economica* , 56, 323-341.

Probabilistic Intelligence in Football. (2010). Retrieved February 24, 2011, from Probabilistic Intelligence in Football: http://www.pi-football.com

Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *The Statistician* , 49, Part 3, pp. 339-418.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* , 6: 2, 461-464.

Stael Von Holstein, C., & Murphy, A. (1978). The Family of Quadratic Scoring Rules. *Monthly Weather Review.* , 106: 917-924.

Vecer, J., Kopriva, F., & Ichiba, T. (2009). Estimating the Effect of the Red Card in Soccer. *Journal of Quantitative Analysis in Sports* , 5: Iss. 1, Article 8.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* San Francisco, CA: Elsevier.

Biographies

Anthony Constantinou received his BSc (honours) in Computer Science and his MSc (with Distinction) in Artificial Intelligence with Robotics from the University of Hertfordshire, UK. He is currently an EPSRC PhD researcher at Queen Mary, University of London (QMUL) within the Risk Assessment and Decision Analysis research group since October 2009, where he also works as a teaching assistant for Decision and Risk Analysis and Software Engineering modules. His primary research interests focus on intelligent decision making processes under uncertainty.

Norman Fenton is a Professor of Computer Science at Queen Mary, University of London (QMUL) since 2000, where he teaches Software Engineering, he is the Head of the Risk Assessment and Decision Analysis Research group (RADAR) and also the Chief executive Officer of Agena Ltd, a company that specialises in risk management for critical systems. Norman's work typically involves analysing and predicting the probabilities of unknown events using Bayesian statistical methods including especially causal, probabilistic models. This type of reasoning enables improved assessment by taking account of both statistical data and also expert judgment. Norman's experience in risk assessment covers a wide range of application domains such as legal reasoning (he has been an expert witness in major criminal and civil cases), medical trials, vehicle reliability, embedded software, transport systems, and financial services. Norman has a special interest in raising public awareness of the importance of probability theory and Bayesian reasoning in everyday life (including how to present such reasoning in simple lay terms). He has been a season ticket holder at Tottenham Hotspur for many years.