Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models

Anthony Constantinou* and Norman E. Fenton

Risk Assessment and Decision Analysis Research Group (RADAR), Department of Electronic Engineering and Computer Science, Queen Mary, University of London, London, UK

[Submitted for publication in the Journal of the Royal Statistical Society, Series C: Applied Statistics, August 2011**]**

Summary. Despite the massive popularity of probabilistic (association) football forecasting models, and the relative simplicity of the outcome of such forecasts (they require only three probability values corresponding to home win, draw, and away win) there is no agreed scoring rule to determine their forecast accuracy. Moreover, the various scoring rules used for validation in previous studies are inadequate since they fail to recognise that football outcomes represent a ranked (ordinal) scale. This raises severe concerns about the validity of conclusions from previous studies. There is a well-established generic scoring rule, the Rank Probability Score (RPS), which has been missed by previous researchers, but which properly assesses football forecasting models.

Keywords: association football forecasting, forecast assessment, forecast verification, predictive evaluation, probability forecasting, rank probability score, sports forecasting

1. Introduction

If a problem has a fixed set of possible outcomes (such as a football match where the outcomes are H, D, A corresponding to Home win, Draw, Away win), a *probabilistic forecast model* is one that provides predicted probabilities (such as p_H, p_D, p_A in the case of football) corresponding to the outcomes. Probabilistic forecasting has become routine in domains as diverse as finance, macroeconomics, sports, medical diagnosis, climate and weather. Some forecasts are conceptually simple (involving a single binary outcome variable) while others are complex (involving multiple possibly related numeric variables). To determine the accuracy of forecast models we use so-called *scoring rules***,** which assign a numerical score to each prediction based on how 'close' the probabilities are to the (actual) observed outcome. For a detailed review on the theory of scoring rules and probability assessment in general see (Jolliffe & Stephenson, 2003; Gneiting & Raftery, 2007)

Defining suitable scoring rules has proven to be extremely difficult, (Murphy & Winkler, 1987; Garthwaite, Kadane, & O'Hagan, 2005), even for apparently 'simple' binary forecasts (Jolliffe & Stephenson, 2003), or even when restricted to a specific application domain. For example, in macroeconomics serious concerns over the various scoring rules have been exposed, (Leitch & Tanner, 1991; Armstrong & Collopy, 1992; Hendry, 1997; Fildes & Stekler, 2002). As a result, many have suggested the use of more than one rule in an attempt to obtain an informed picture of the relative merits of the forecasts (Jolliffe & Stephenson, 2003).

The work in (Gneiting & Raftery, 2007) addresses many of the above problems, by recognising that the underlying measurement scale type of the outcomes for a specific problem should drive the type of scoring rule used. For example, if the problem is to predict a winning lottery number then although the possible outcomes appear to be an ordered set $\{1,2,\ldots,49\}$ the relevant scale type is only *nominal*; if the winning number is 10 then a prediction of 9 is no 'closer' than a prediction of 49 – they are both equally wrong and any scoring rule should capture this. On the other hand, if the problem is to predict tomorrow's temperature in degrees centigrade the relevant scale type is (at least) *ordinal (ranked)* since if the actual temperature is 10 then a prediction of 9 must be considered closer than a prediction of 49, and any scoring rule should capture this.

The crucial observation we make about football forecasting is that the set of outcomes $\{H, D, A\}$ must be considered as an ordinal scale and not a nominal scale. The outcome D is closer to H than A is to H; if the home team is leading by a single goal then it requires only one goal by the away team to move from H to D. A second goal by the away team is required to move the result on to an A. It follows that if the result is H any scoring rule should penalise the probability assigned to A more heavily than that assigned to D. It turns out that, as obvious as this observation appears, we will show in Section 2 that it has been missed in every previous study of football forecasting systems. To

* *Address for correspondence*: Risk Assessment and Decision Analysis Research Group (RADAR), Department of Electronic Engineering and Computer Science, Queen Mary, University of London, London, UK, E1 4NS.
E-mail: constantinou@eecs.qmul.ac.uk (A. Constantinou), norman@eecs.qmul.ac.uk (N. Fenton)

demonstrate this we introduce some simple benchmark scenarios along with the result required of any valid scoring rule and show that none of the previously used scoring rules satisfies all of the benchmarks. It follows that all of the previous studies on football forecast models have used inadequate scoring rules. In Section 3 we show that there is a well established standard scoring rule for ordinal scale outcomes, called *The Rank Probability Score* (RPS), which satisfies all the benchmark examples for football forecast models. The implications of this are discussed in Section 4.

2. Scoring rules used in previous football forecast studies

We have reviewed all of the previously published studies in which one or more explicit scoring rule was used to evaluate the accuracy of one or more probabilistic football forecasting model. There were nine such studies. The various scoring rules are defined in the Appendix. They fall into two categories:

a) Those which consider only the prediction of the observed outcome (also known as local scoring rules). They are: Geometric Mean, Information Loss, and Maximum Log-Likelihood
b) Those which consider the prediction of the observed as well as the unobserved outcomes. They are: Brier Score, Quadratic Loss function, and Binary decision

At least one of the category (a) scoring rules was used in (Dixon & Coles, 1997; Rue & Salvesen, 2000; Hirotsu & Wright, 2003; Goddard, 2005; Karlis & Ntzoufras, 2003; Goddard, 2005; Forrest, Goddard, & Simmons, 2005; Joseph, Fenton, & Neil, 2006; Graham & Stott, 2008; Hvattum & Arntzen, 2010). At least one of the category (b) scoring rules was in (Forrest et al., 2005; Joseph et al., 2006; Hvattum & Arntzen, 2010).

In addition to the above scoring rules, some researchers have proposed and applied different 'ranking' methods of validation, such as the error in cumulative points expected for a team after a number of matches, the RMS and Relative Rank Error of the final football league tables, and pair-wise comparisons between probabilities. At least one of these types of methods has been found in (Pope & Peel, 1988; Dixon & Pope, 2004; Min et al., 2008; Baio & Blangiardo, 2010). However, these methods are beyond the scope of this paper; they do not represent an actual scoring rule, since they cannot provide a measure of accuracy for the prediction of a particular game.

The work in (Gneiting & Raftery, 2007) demonstrates that none of the above scoring rules are suitable for a problem whose outcomes are on a ranked scale, and hence they are unsuitable for assessing football forecast models. We demonstrate this informally by introducing five 'benchmark' scenarios (Table 1) in each of which we have a match which has two 'competing' prediction models α and β together with an actual outcome. In each case it is clear intuitively which of the predictions should be scored higher.

Table 1. Hypothetical forecasts by models α and β, and results for matches 1 to 5.

Match	Model	$p(H)$	$p(D)$	$p(A)$	Result	'Best model'
1	α	1	0	0	H	α
	β	0.9	0.1	0		
2	α	0.8	0.1	0.1	H	α
	β	0.5	0.25	0.25		
3	α	0.35	0.3	0.35	D	α
	β	0.6	0.3	0.1		
4	α	0.6	0.3	0.1	H	α
	β	0.6	0.1	0.3		
5	α	0.5	0.45	0.05	H	α
	β	0.55	0.10	0.35		

This is because:

a) Match 1: (Taking account of perfect accuracy): Model α predicts the actual outcome with total certainty and hence must score better than any other, less perfect, predicted outcome.
b) Match 2: (Taking account of predicted value of the observed outcome) Both models α and β assign the highest probability to the winning outcome H, with the remaining two outcomes evenly distributed. Since the observed value of α is higher than that of β, it must score higher.
c) Match 3: (Taking account of distribution of the unobserved outcomes) Given that the observed outcome here is D, both of the unobserved outcomes are equally distanced from

the observed one. Hence, the ordering concern here is eliminated. Still, a scoring rule must identify that model α is more accurate since its overall distribution of probabilities is more indicative of a draw than that of β (which strongly predicts a home win).

d) Match 4: (Taking account of ordering when the set of unobserved outcomes are equal) Both models *α* and *β* assign the same probability to the winning outcome H. This time, however, they also assign the same probability values (but in a different order) to the unobserved outcomes (0.30 and 0.10). But, a scoring rule must identify that model α is more accurate since its overall distribution of probabilities is more indicative of a home win.

e) Match 5: (Taking account of overall distribution) Although *α* predicts the actual outcome *H* with a lower probability than *β* the distribution of α is more indicative of a home win than β. This match is the most controversial, but it is easily explained by considering a gambler who is confident that the home team will not lose, and so seeks a *lay* bet (meaning explicitly that the bet wins if the outcome is H or D). Assuming that *α* and *β* are forecasts presented by two different bookmakers, bookmaker α will pay much less for the winning bet (this bookmaker considers that there is only 5% probability the home team will lose, as opposed to bookmaker β who considers it a 35% probability).

Table 2 presents the results of the previously used football scoring rules for the benchmark scenarios and determines the extent to which they satisfy the benchmark for each of those forecasts presented in Table 1. A tick means that the scoring rule correctly scores model α higher than β. A double cross means that the scoring rule incorrectly scores model β higher than α. A single cross means that the scoring rule returns the same value for both models. Where necessary, the score is rounded to 4 decimal points. For the rules Binary Decision, Geometric Mean and MLLE, a higher score indicates a better forecast; whereas for the rules Brier Score and the Information Loss a lower score indicates a better forecast.

Table 2. Applying the specified scoring rules to each benchmark presented in table 1

Match (Model)	Binary Decision Score	Brier Score	Geometric Mean Score	Information Loss Score	MLLE Score
1	✓	✓	✓	✓	✓
(α)	1	0	1	0	0
(β)	0	0.02	0.9	0.152	-0.1054
2	✗	✓	✓	✓	✓
(α)	1	0.06	0.8	0.3219	-0.2231
(β)	1	0.375	0.5	1	-0.6931
3	✗	✓	✗	✗	✗
(α)	0	0.735	0.3	1.7369	-1.2039
(β)	0	0.86	0.3	1.7369	-1.2039
4	✗	✗	✗	✗	✗
(α)	1	0.26	0.6	0.7369	-0.5108
(β)	1	0.26	0.6	0.7369	-0.5108
5	✗	✗ ✗	✗ ✗	✗ ✗	✗ ✗
(α)	1	0.455	0.5	1	-0.6931
(β)	1	0.335	0.55	0.8625	-0.5978

None of the scoring rules returns the 'correct' outcome for all 5 scenarios. Indeed, all of the scoring rules fail to correctly identify model *α* as superior for scenarios 4 and 5.

3. The Rank Probability Score (RPS)

The RPS was introduced in 1969 (Epstein, 1969). It is both *strictly proper* (Murphy, 1969) and *sensitive to distance* (Murphy, 1970). The RPS has been described as a particularly appropriate scoring rule for evaluating probability forecasts of ordered variables (Murphy, 1970). In general the RPS for a single problem instance is defined as:

$$RPS = \frac{1}{r-1}\sum_{i=1}^{r}\left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j\right)^2$$

where r is the number of potential outcomes, and p_j and e_j are the forecasts and observed outcomes at position j. The RPS represents the difference between the cumulative distributions of forecasts and observations, and the score is subject to a negative bias that is strongest for small ensemble size (Jolliffe & Stephenson, 2003). Since the scoring rule is sensitive to distance, the score penalty increases the more the cumulative distribution forecasted differs from the actual outcome (Wilks, 1995). For a detailed analysis on the RPS see (Epstein, 1969).

Table 3. Score generated by the RPS for each hypothetical forecast presented in table 1, along with the respective cumulative distributions (forecasted and observed).

Match	**Model**	$\sum_{j=1}^{i=1,2,3} p_j$	$\sum_{j=1}^{i=1,2,3} e_j$	**RPS**
1	α	1,1,1	1,1,1	(0.0000)
	β	0.90,1,1	1,1,1	0.0050
2	α	0.80, 0.90, 1	1, 1, 1	(0.0250)
	β	0.50, 0.75, 1	1, 1, 1	0.1562
3	α	0.35, 0.65, 1	0, 1, 1	(0.1225)
	β	0.60, 0.90, 1	0, 1, 1	0.1850
4	α	0.60, 0.90, 1	1, 1, 1	(0.0850)
	β	0.60, 0.70, 1	1, 1, 1	0.1250
5	α	0.50,0.95,1	1, 1, 1	(0.1262)
	β	0.55,0.65,1	1, 1, 1	0.1625

Table 3 presents the generated score for each of the scenarios presented in table 1, along with the respective cumulative distributions (forecasted and observed). A lower score (rounded to 4 decimal points) indicates a better forecast. Unlike the previous metrics the RPS correctly scores α as 'best' for all 5 matches.

4. Implications and conclusions

Measuring the accuracy of any forecasting model is a critical part of its validation. In the absence of an agreed and appropriate type of scoring rule it might be difficult to reach a consensus about: a) whether a particular model is sufficiently accurate; and b) which of two or more competing models is 'best'. In this paper, the fundamental concern is the inappropriate assessment of forecast accuracy in association football, which may lead in inconsistencies, whereby one scoring rule might conclude that model α is more accurate than model β, whereas another may conclude the opposite. In such situations the selection of the scoring rule can be as important as the development of the forecasting model itself, since the score generated practically judges the performance of that model. On the one hand, an outstanding model might be erroneously rejected while on the other hand a poor model might be erroneously judged as acceptable.

We have shown that, by failing to recognise that football outcomes are on an ordinal scale, all of the various scoring rules that have previously been used to assess the forecast accuracy of football models are inadequate. They fail to correctly determine the more accurate forecast in circumstances illustrated by the benchmark scenarios of Table 1. This failure raises serious concerns about the validity and conclusions from previous studies that have evaluated football forecasting models. What makes the failure of all previous studies to use a valid scoring rule especially surprising is that there was already available (before any of the studies were conducted) a well-established scoring rule, the RPS, that avoids the inconsistencies we have demonstrated.

With the relentless increase in interest in football forecasting it will become more important than ever that effective scoring rules for forecast models are used. Although we are not suggesting that the RPS is the only valid candidate for such a scoring rule, we have shown that (unlike the previous scoring rules used) it does satisfy the basic benchmark criteria expected.

Given the massive surge in popularity of the sport and its increasing dominance in sport betting internationally, it is important to note that we have only considered the assessment of forecast accuracy and not profitability. We cannot claim that a forecasting model assessed as more accurate than a bookmaker by such a rule will necessarily indicate profitability. After all, profit is not only dependent on the accuracy of a model but also on the specified betting methodology. Other

researchers have already concluded that it might be best to combine profitability methodologies with a proper forecast assessment rule for that matter (Wing et al., 2007). Yet, it is evident that profitability is dependent on accuracy and not the other way around. Accordingly, higher forecast accuracy indicates a higher prospective profit which denotes the importance of propriety in forecast assessment.

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

Armstrong, J., & Collopy, F. (1992). Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting* , 8, 69-80.

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

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

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.

Epstein, E. (1969). A Scoring System for Probability Forecasts of Ranked Categories. *J. Appl. Meteor.* , 8, 985-987.

Fildes, R., & Stekler, H. (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics* , 24, 435-468.

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

Garthwaite, P., Kadane, J., & O'Hagan, A. (2005). Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association* , 100, 680-700.

Gneiting, T., & Raftery, A. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* , 102(477), 359-378.

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.

Hendry, D. (1997). The Econometrics of Macroeconomic Forecasting. *The Economic Journal* , 1330-1357.

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)* , 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.

Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecast Verification. A Practitioner's Guide in Atmospheric Science.* West Sussex, England: Wiley.

Joseph, A., Fenton, N., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems* , 7, 544-553.

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

Leitch, G., & Tanner, J. E. (1991). Economic Forecast Evaluation: Profits Versus The Conventional Error Measures. *American Economic Association* , 580-590.

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.

Murphy, A. (1969). On the "ranked probability score". *J. Appl. Meteor.* , 8, 988-989.

Murphy, A. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review* , 98, 917-924.

Murphy, A., & Winkler, R. (1987). A general framework for forecast verification. *Monthly Weather Review* , 115, 1330-1338.

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

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

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

Wilks, D. (1995). Statistical Methods in the Atmospheric Sciences. *International Geophysics Series* , Vol. 59, Academic Press, 467 pp.

Wing, C., Tan, K., & Yi, L. (2007). *The Use of Profits as Opposed to Conventional Forecast Evaluation Criteria to Determine the Quality of Economic Forecasts.* Singapore: Nanyang Business School, Nanyang Technological University, Nanyang Avenue.

Appendix: Description of previously proposed scoring rules

In what follows we assume that:

a) In each instance (such as a single football match) there are r possible outcomes (so $r = 3$ for football matches)
b) The model predictions for the outcomes are respective probabilities $p_1, p_2, \dots, p_n$
c) The respective actual observed outcomes are $e_1, e_2, \dots, e_n$. So for football matches the e_is are either 1 or 0 and in such cases the index of the actual observed outcome will be denoted w, so $e_i = 1$ if $i = w$ and 0 if $i \neq w$.

A scoring rule is actually defined in terms of two components:

1. A score for an individual instance given a forecast for that instance (e.g. a single football match)
2. A method for defining the cumulative scores over a set of instances. With the exception of the geometric mean, method used is either the arithmetic mean of the individual scores or the total of the individual score over those instances. The geometric mean, in contrast uses a multiplicative function for the cumulative score; meaning that it punishes individual bad predictions heavier.

For the purposes of this paper it is sufficient to consider only how the scoring rule is defined for individual instances. These definitions are:

a) **Binary Decision**: The score is 1 if $p_w > p_i$ for each $i \neq w$ and 0 otherwise. The Binary Decision rule takes into consideration all of the probabilistic values since it seeks the highest one for assessment. However, the rule does not generate a score based on the values.

b) **Brier Score**: also known as Quadratic Loss, the Brier Score (Brier, 1950) is

$$\sum_{i=1}^{r} (p_i - e_i)^2$$

c) **Geometric Mean** For an individual instance the score is simply p_w.

d) **Information Loss**: For an individual instance this is defined as $-\log_2 p_w$. In order to avoid the *zero-frequency* problem (whereby the informational loss is minus infinity when p_w is zero) it is expected that non-zero probabilities are assigned to every outcome.

e) **Maximum Log-Likelihood Estimation (MLLE)**: Maximum likelihood estimation is known as an approach to parameter estimation and inference in statistics, which states that the desired probability distribution is the one that makes the observed data 'most likely'. Informational Loss and Maximum Log-Likelihood estimation differ in equation but generate identical forecast assessment. As in most cases, we present the MLLE over the MLE since it generates identical assessment while reducing the computational cost significantly (Myung, 2003). For a likelihood test, the Binomial distribution with parameters n (trials), and s (successes) could be used. However, since we only have one observation for every match prediction (n and s are equal to 1), the Log-likelihood is $ln(p_w)$. Therefore, in order to avoid unnecessary calculations in this paper, we simply define the MLLE to be $ln\,(p_w)$ for an individual match.