Visual Analysis of Forecasts of Football Match Scores

Flavio Fontanella
FGV/EMAp, Rio de Janeiro, Brazil
Email: flavio.fontanella@gmail.com

Asla Medeiros e Sá
FGV/EMAp, Rio de Janeiro, Brazil
Email: asla.sa@fgv.br

Moacyr Alvim H. B. da Silva
FGV/EMAp, Rio de Janeiro, Brazil
Email: moacyr@fgv.br

***Abstract*—The present work describes the application of information visualization techniques to better understand the behaviour of a forecasting model designed to predict football match scores based on past confrontations. We have run the forecasts during the 2019 season of Brazilian National Championship and, once it was over, we gathered data to observe the performance of the forecasts, based on a set of evaluators we propose. The main contribution of this paper is the introduction of visual devices, each attached to one of these evaluators, designed to enhance interpretation of the performance of forecasting models in general and sports forecasts in particular. We also present other visualizations that help taking notice of particular features of the championship and the forecasts.**

I. Introduction

Football is arguably the most popular sport in the world. It is played by more than 270 million people all over the globe and watched by billions. [1]. As such, sports and football forecasting have long been in use for multiple purposes. Trying to make profit in gambling is probably the most popular one. Sports betting market has steadily gained popularity over time [2] and football, in particular, has experienced the fastest growth in recent years [3]. In Europe, online platforms revenues have reached 20 billion euros a year, presenting annual growth rates of up to 15% [2].

Other applications for sports forecasting are the information enhancement for mediatic match coverage as well as in supporting team management decisions [4]. In addition, the availability of plenty of data to be explored [5] make it possible to test scientific hypothesis, given the biases that naturally arise, as well as the scrutiny of the strengths and weaknesses of both teams and individual players. A forecasting model for football match results has potential uses in all such applications, among others, and thus, it's of great interest that it is tailored to become as accurate as can be.

In the present work, we propose a set of evaluators, each with a corresponding visual tool, aiming to better understand the behaviour of a forecasting model that predicts football match scores based on past confrontations. Information visualization techniques are useful to communicate forecasts as well as to evaluate the forecasting model and to analyze some aspects of the matches being predicted.

This paper is structured as follows: in Section II we review recent literature related to football forecasting, forecasting model evaluation and the usage of information visualization in sports data. In Section III we describe the football scores forecasting model, the context of the match forecasting and its databasis. In Section IV the goals and main issues of evaluating a forecasting model are discussed and visual tools reveals some characteristics of the model behaviour when applied to our dataset. Further visual analysis of complementary aspects is conducted in Section V. Finally, in Section VI we draw some conclusions and comment on future work.

II. Context Overview

A. Football Match Forecasting

Previous works from several authors observed that the number of goals scored in a football match can be estimated by a Poisson distribution. In his 1982 paper, Maher [6] wrote: *"There are good reasons for thinking that the number of goals scored by a team in a match is likely to be a Poisson variable"* He then proposed the use of independent Poisson random variables (henceforth treated as r.v.'s) to model football match scores. The mean of the Poisson r.v.'s were determined by the teams attack and defence parameters, which were inferred based on actual match scores. In 1997 Lee [7] used a similar model to simulate the 95/96 English Premier League matches and compared actual to simulated final league standings. In the same year, Dixon and Coles [8] argued against the independence assumption of the Poisson r.v.'s. and introduced a time weighting function to estimate team parameters, applying their model to test a betting strategy against market odds.

Over the years, other techniques have been proposed to model football results behaviour, including probit regressions, rating systems and machine learning models [9]. Most of the proposed approaches were not desiged to predict match scores; instead, they attribute probabilities to the three possible match outcomes: a win for the home team, a win for the away team and a draw, which is ultimately more relevant than the match score itself. Nevertheless, a model that forecasts match scores can easily convert to match outcomes forecasting, for the probability of an outcome equals the sum of the probabilities for all scores that corresponds to it. For instance, to assess the probability of a draw, it just takes adding the probabilities for the scores 0-0, 1-1, 2-2, 3-3 and so on.

B. Evaluating Forecasting Models

The quality evaluation of a forecasting model is not an easy task. Amongst several proposed approaches, none has been undisputed acknowledged as the best one. In 2011,

Constantinou an Fenton [10] highlighted the importance of evaluating a model as a critical part of its validation: *"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, (. . .) there is none."* The authors listed a few metrics that had been in use and presented some tests and simulations to show that applying them to evaluate two or more systems could lead to conflicting findings about which one is best. One year later, they [11] stated that the *Ranked Probability Score* (RPS) is an appropriate evaluation metric based on their criteria. In 2019, Wheatcroft [12] debated Constantinou and Fenton's arguments and compared the RPS to two other evaluation metrics, the *Brier Score* (BS) and the *Ignorance Score* (IS), concluding that both perform better than the RPS in some contexts. In all cases, the evaluation metrics are based on the triplet of outcome probabilities (home win, draw, away win).

In 2019, Reade et al [13] evaluated score forecasts from TV experts, tipsters crowd, betting market bookmakers and a statistical model, in the case where most of the metrics were based on *point forecasts*, i.e., when a particular score is selected among all possible scores. As for *probability forecasts*, where probability is distributed over all possible scores, they tested for efficiency and encompassment between models, indicating that scoring rules designed for point forecasts cannot evaluate probability forecasts. The authors cited Foulley and Celeux [14] as the only other study to evaluate football match score forecasts.

Evaluating score forecasts seems quite harder than it is for outcome forecasts since the forecasting space is of much higher dimension. While the outcome consists of only three possible results, scores can, potentially take value at any pair of non-negative integers. In practice, integers bigger than 10 are low probability events in football, thus, it is possible to simplify the output by limiting the number of goals scored by a team to a single digit, that can be represented by a squared matrix of scores from 0-0 to 9-9. This 100-point probability distribution is to be compared to the actual score after the match is played, which can be done adopting distinct approaches: (i) by checking the probability assigned to the actual final score alone; alternatively, (ii) by taking into account the whole distribution of score probabilities to estimate how good or bad the prediction has been.

In the present work, we approach the problem by taking a *look* at these comparisons between forecasts and actual scores from several perspectives, aiming to shed some light on the performance of the forecasting model adopted. To the best of our knowledge, we are not aware of any other work on the subject, and previous works concentrate on evaluating point forecasts, which is quite different from the present proposal.

C. Information Visualization in Football and Related Fields

In 2018, a state-of-the-art paper focusing in visualization in sports [15] has been published. The authors proposed the categorization of sports visualizations by its type of data into three classes: *box-score data*, *tracking data* and *meta-data*. The term box-score designates the statistical summary of a game including any discrete data referencing in-game events. The authors separate the visualization of box-score data into *time-evolving championship tables and rankings* that are also extensively used in data storytelling and published in news media; while *scores, goals, and points* relate to the general sports rules form another group of visualization solutions. Tracking and sensing technologies are used to collect spatio-temporal information within the match. Recently it has been widely explored in football as, for instance, in [16], [17]. Additionally, data that surrounds sports can enrich the context information and may be also considered being referred to as meta-data, as in [18].

The popularity of box-score data is argued to be due to its cheap and technology-independent means of acquisition. Although box-score data is often simple and small-scale, it is diverse and highly sport-specific. The present work fits in this box-score category, although the aim is somewhat different, in the sense that it does not intend to tell a story of what happened. Instead, we propose to use visualization as an evaluation tool to the probability forecasting model, detailed in Section III, by comparing forecasts to the actual results. The phenomenon being observed (and predicted) is discrete and the events are sparse. Such characteristics narrow the set of works that our approach could be compared with.

Information visualization as a field is growing rapidly and it is increasingly difficult to follow the growing body of literature within the field. In 2017, a Survey of Surveys (SoS) on the subject has been published [19] where the authors classifies survey papers into natural topic clusters which enable readers to find relevant literature and develop the first classification of classifications. The SoS paper gives a glimpse of the amount of research being conducted in the subject. Surprisingly, visualization in sports is not present in the SoS as a real world application field, and a state-of-the-art on the subject was to appear only one year later [15]. Nevertheless, we could identify the discussion of visualization of forecasting models within the spatio-temporal applications.

A discussion of great relevance is the visual display of uncertainty. For instance, in [20], the authors test whether different graphical displays of a hurricane forecast containing uncertainty would influence a decision about storm characteristics. This discussion calls the attention to the importance of conveying information about model uncertainty in visual displays. Probabilistic visualizations have various designs and are used to communicate uncertainty information in many domains, but only recently this discussion is acheiving maturity. One recent contibution of relevance is the proposal of a Probabilistic Grammar of Graphics (PGoG) [21]. The authors argued that visualizations depicting probabilities and uncertainty are vastly adopted in many fields of applications, yet these probabilistic visualizations are difficult to specify, prone to error, and their designs are cumbersome to explore. To address the issue, the authors proposed a grammar and provided a proof-of-concept implementation of PGoG in R.

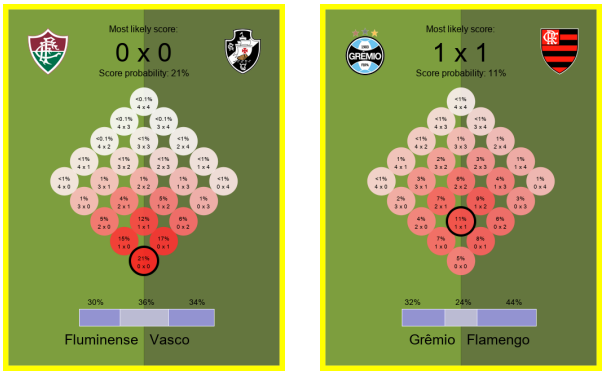
III. Football Forecasting Model

The forecasting model we use is based on Poisson distributions and independence assumption in a similar way to the ones proposed by Maher [6] and Lee [7] and, additionally, incorporates a time weighting function in the estimation of teams parameters, as in Dixon and Coles' model [8] .

The number of goals scored by a home team A against an away team B is modelled as a Poisson random variable with mean $\lambda = \frac{\alpha_h(A)}{\beta_a(B)}$, where $\alpha_h(A)$ is a parameter related to team A home attacking power and $\beta_a(B)$ is related to team B away defence power.

Each of the 20 teams in the championship have two attack and two defence parameters $(\alpha_h, \alpha_a, \beta_h, \beta_a)$ that are estimated through maximum likelihood using all the results from previous matches of that championship edition. Each round of the championship is weighed by the time weighting function taking values within the interval $[0.2, 1]$, simulating a loss of memory.

Then, the parameters (α_h, β_h) from the home team, and (α_a, β_a) from the visiting team, are used to forecast a match. Once the means of the two Poisson random variables have been determined, their individual distributions are combined to generate a probability for each possible final score of the match. In order to communicate the match forecast, the probabilities of each score, calculated by the model, are displayed as in Figure 1, before the match is played.



(a) Case I (b) Case II

Fig. 1. Forecasted score probability matrix with additional summary and context information for two matches from the second half of the season. In case I high probability is predicted for one or a few scores. In case II probability is spread over several scores, depicting high uncertainty.

The proposed chart displays different pieces of information, such as outcomes probability distribution and scores probability distribution. The chosen representation is a score probability matrix with values mapped into luminance. On top of this central chart representation, several additional elements are added to produce the final presentation contextualising the match and resuming the probability distribution to facilitate reader's interpretation. It also highlights the most likely score (highest probability forecasted) and includes team badges arranged to resemble the two sides of the pitch, aiming to ease the task of catching the balance of distributions between the teams. The score probabilities representation is similar to a 538's figure [22] used to explain their process to generate football match forecasts, but conveys more information and context.

The same process is repeated for each round of the second half of the season (rounds 20 to 38). All match scores from previous rounds are used to estimate teams' parameters at the time, recalculating all the λ's to forecast each match in next round. The first half of the season (rounds 1 to 19) is not predicted by the model for once we wouldn't have enough data to estimate attack and defence parameters reliably. Instead, its information is gathered and used to forecast the second half of the season. The forecasted matrices for the season are available at www.fgv.br/emap/campeonato-brasileiro/previsoes.html.

IV. Forecast Evaluation through Visual Tools

In the present work, the main goal is to judge the quality of a forecast, or best, of a set of forecasts, given the actual match scores. The analysis is based on the dataset consisting of forecasts and results for all the 190 matches played in the second half of 2019 Brazilian National Championship for men's football. We will work with three different approaches that differ in the way they compare the forecast to the score. For each approach we shall propose one or more metrics and develop corresponding visual displays to help assessing the information. In the end, our methodology will consist of four evaluators, each one with a visualization that presents the results for every match being evaluated.

A. Actual Score Position in the Forecast Ranking

One simple evaluation approach considers ranking the forecasted scores according to their probability and verifying the position of the actual score within the forecasted rank. This approach is very practical and may be appealing, although it ignores the exact probability (trust) placed on the score. We simplify the implementation of this approach by adopting a point-reward system that awards ten points if the actual score was ranked first; nine points if it was ranked second and so on, until no point is awarded if the actual score did not rank among top ten predicted scores.

Figure 2 illustrates the reward of the actual score for its forecasting rank for each match of the second half of the season. Top chart presents the information respecting the order of the match in the championship table in x axis, while in the bottom chart, the same data is reordered for grouping matches by reward, which reveals the pattern we seek for analysis. The annotation in Figure 2 summarizes the number of matches that received each grade (10 to 0); 24 out of the 190 matches were top rated (12.6%), with 16 (8.4%) making second and other 24 getting third. 164 matches (86.3%) got at least 1 point in the reward system, with the median match being awarded 6 points and the average 5.45. The teams with best and worst forecasting performances according to this proposed metric had their matches emphasized as illustrated in Figure 3.

B. Probability Forecasted to the Actual Score

Our second evaluation approach considers the probability forecasted by the model to the actual score. For instance,

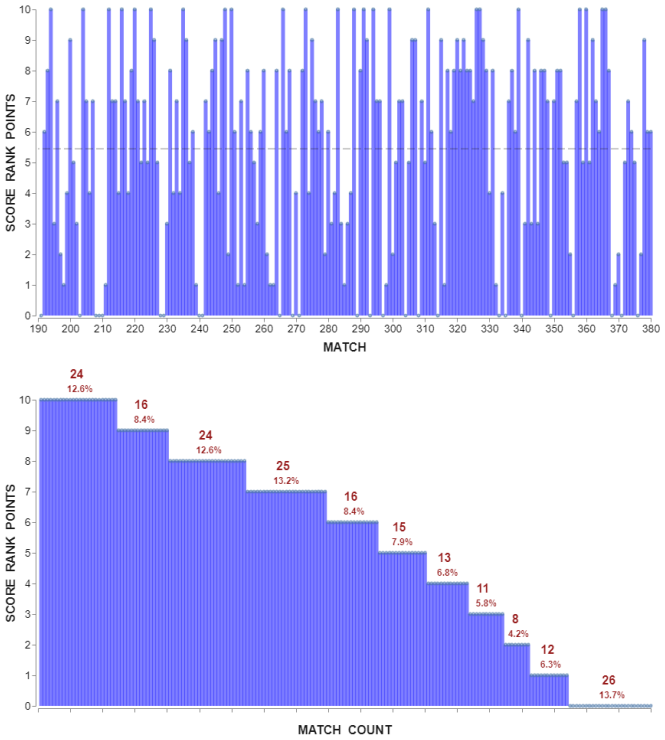


Fig. 2. Actual Score Points in the Forecast Ranking. *Top*: ordered by match occurrence in championship table in time. *Bottom*: same data grouped by actual score reward.

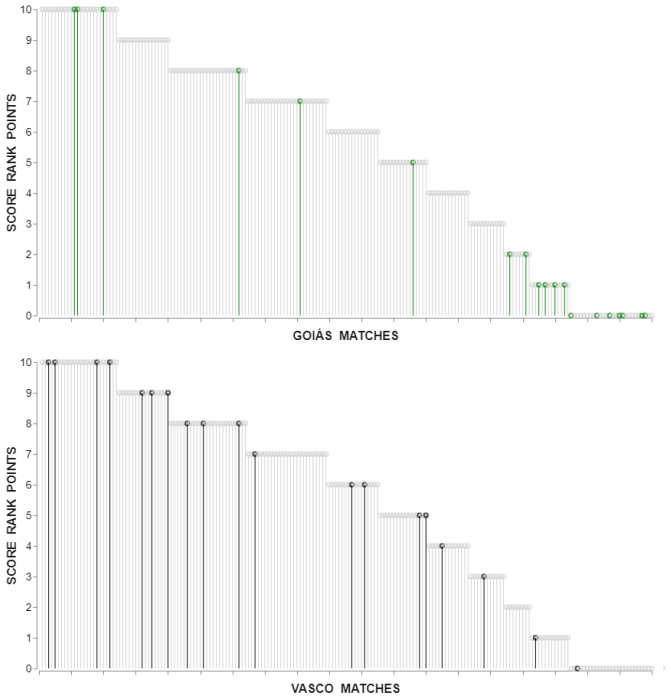


Fig. 3. Actual Score Points in the Forecast Ranking for Teams with Worst and Best Performances According to this Metric. *Top:* Goiás' Matches. *Bottom* Vasco's Matches.

Actual Score Probability	# of matches	% of matches
>0.20	5	2.6
0.20 - 0.15	17	8.9
0.15 - 0.10	54	28.4
0.10 - 0.05	63	33.2
< 0.05	51	26.8

TABLE I
SUMMARY OF ACTUAL SCORE FORECASTED PROBABILITIES

if the model predicted a probablility of 20% to the actual score it is considered '20% correct'. The implementation is straightforward, for each match we simply check the forecasted probability attributed to the actual score. This approach may seem more informative than the previous, but it still lacks information about how the remaining probability is distributed through all possible scores.

Figure 4 maps into a blue bar and circle the probability assigned to the actual score of each match during the second half of the season. Top chart is ordered by match occurrence in the championship table in time, while in the bottom chart the same data is reordered to reveal the visual pattern. Additionally, the charts deliver one more piece of information, the red bars, that represent the probability assigned to the most likely score of each match and may be regarded as a measure of uncertainty for the forecast. The teams with best and worst forecasting performances according to this proposed metric had their matches emphisized as illustrated in Figure 5.

Table I summarizes the number of matches with predicted probabilities for the actual score in each interval of 5 pct. The higher probability assigned to an actual score was 25.3%, while the lowest was 0.007%. The median match score prediction was 7.8% and the average was 8.8%.

C. Actual Score Distance to Probability Distribution

Our third approach is to consider all the forecasted scores in order to define a distance between the actual score and the forecasted distribution. Formally it's necessary to define a metric once the actual score is a point of the match scores space while the forecast is a distribution on that discrete space. Thus, in order to measure distances, we need either to weight the distances over the whole distribution or to summarize the distribution into one representative point and take the distance from it to the actual score. We took the second path. Since there is more than one way to summarize the distribution into a representative point, we decided for two implementations: how far was the actual score from (A) the mean of the forecasted distribution (mean predicted score) and (B) the mode of the forecasted distribution (top rated prediction). The distance is unfolded in two dimensions: distance in home team goals and distance in away team goals.

We computed, for each match, the mean predicted home score $E[h]$ and the mean predicted away score $E[a]$ and compared them with the actual home (hs) and away (as) scores by taking the differences:

$$Err_h = E[h] - hs$$

$$Err_a = E[a] - as$$

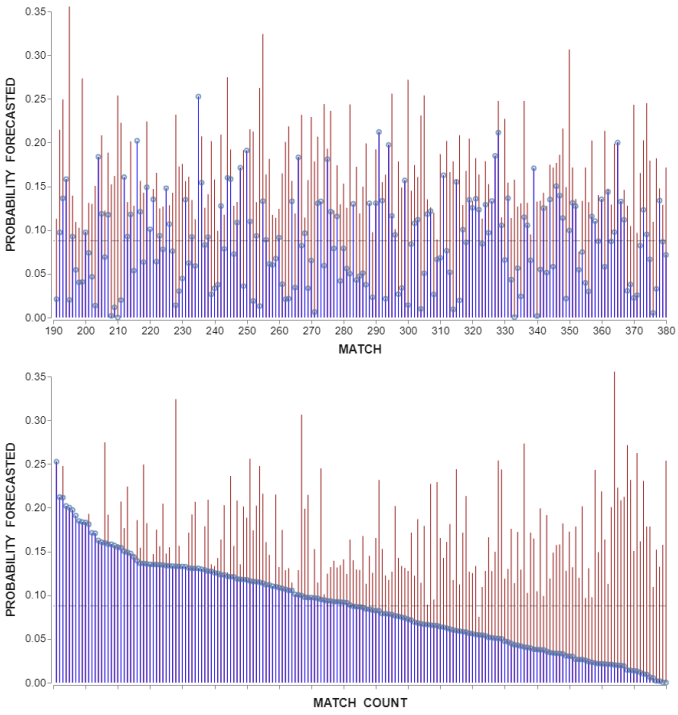


Fig. 4. Actual Score Forecasted Probabilities mapped into blue bars and circles contextualised within the championship. Red bars represent most likely scores forecasted probabilities. *Top*: ordered by match occurrence in championship table in time. *Bottom*: same data with actual score forecasted probability in descending order.

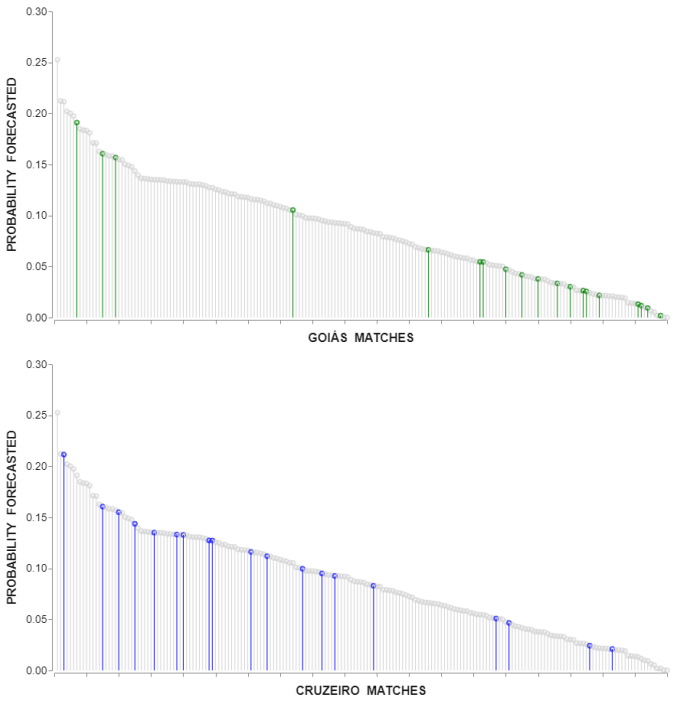


Fig. 5. Actual Score Forecasted Probabilities for Teams with Worst and Best Performances According to this Metric. *Top:* Goiás' Matches. *Bottom* Cruzeiro's Matches.

where the mean predicted values are given by:

$$E[h] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} P(h=i, a=j)\cdot i$$

$$E[a] = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} P(h=i, a=j)\cdot j$$

In practice, we have implemented a truncated version of those infinite sums by using only single digit scores.

The two calculated errors – Err_h and Err_a – were then plotted into a 2-dimensional target, with the center of the target being the actual score of the match, horizontal axis referring to home team score errors and vertical axis to away team score errors. Figure 6 presents the target and all 190 matches plotted. The colored areas on the target represent regions of similar total absolute errors. The central region (henceforth referred as "bullseye") corresponds to the matches with total absolute error up to 0.5 and each subsequent region adds 1 to that limit (second region from 0.5 to 1.5 total absolute error and so on).

Table II summarizes the number of matches plotted into each region of the target. 14 out of 190 matches were plotted into the bullseye (7.4%), while 4 matches (2.1%) missed the target (total absolute error above 4.5). The average total absolute error was 1.68, with 0.94 for home team errors and 0.74 for away team errors.

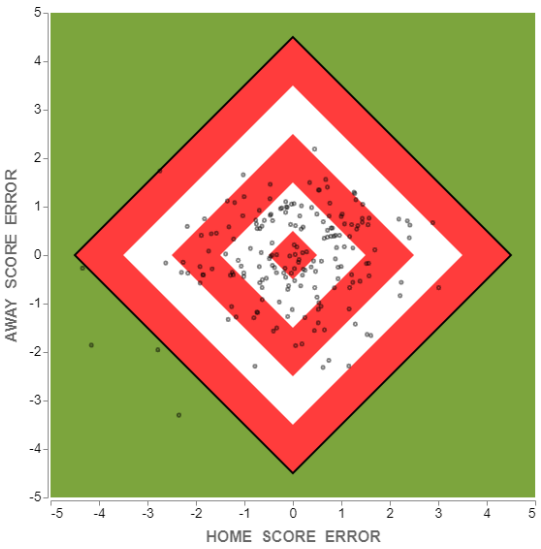


Fig. 6. Target featuring actual score as center with plots representing mean prediction errors

As the two r.v.'s are independent and, for each one, the error being computed as simple, not squared, it seems logical that the combined error would follow the same rule. Thus, the total error is determined by the 1-norm and the regions of equivalence are squares with diagonals parallels to the axes. Thus, we chose to represent the target with concentric squares instead of circles. One could argue that an error of 1 in one

Total Abs. Error to Mean Score	# of matches	% of matches
< 0.5	14	7.4
0.5 - 1.5	77	40.5
1.5 - 2.5	69	36.3
2.5 - 3.5	23	12.1
3.5 - 4.5	3	1.6
> 4.5	4	2.1

TABLE II
TOTAL ABSOLUTE ERROR TO MEAN SCORE

variable does not equal an error of 1 in the other. It could make sense if we consider their distribution. The (absolute) errors for away scores tend to be smaller than the (absolute) errors for home scores. It's easily seen by the target plots and we already know that it is true for the average absolute errors. That would shape the target as a diamond, and so the regions of equivalence. The drawback is that the correct proportion of the diamond (diagonals) would depend on some parameter for the distribution of errors (like variance) which would likely vary from round to round (and of course between tournaments, seasons...). Worse yet, if we were to compare different forecasting models for the same set of matches, each one would have its own target with particular shape, which is not at all effective when trying to compare the performances.

Similar to comparison (A) is comparison (B), but while the former took differences between mean predictions and actual scores, the latter compares actual scores to the top rated predictions $T(h)$ and $T(a)$, such as

$$Err_h = T(h) - hs$$

$$Err_a = T(a) - as$$

where the top rated predicted values are given by:

$$T(h) = argmax_i(\sum_{j=0}^{\infty} P(h = i, a = j))$$

$$T(a) = argmax_j(\sum_{i=0}^{\infty} P(h = i, a = j))$$

Of course, every error comes to be an integer number in this comparison. As before, the two errors Err_h and Err_a calculated for all 190 matches were plotted into a 2-dimensional target as illustrated in Figure 7. The representation has pretty much the same structure as Figure 6 with the major difference being that we account for overplotting - as errors were all integer numbers - by mapping the frequence of occurence of each error into the radius of the circle. The biggest circle occurs into the bullseye, with 24 out of 190 matches (12.6%) getting the perfect score, as we had previously seen in comparison (1). Three other circles came second in size, representing 20 matches, each of them with a -1 error to home or away team score or both.

Table III summarizes the number of matches plotted into each region of the target. The average total absolute error was 1.80, with 0.99 for home team errors and 0.81 for away team errors.

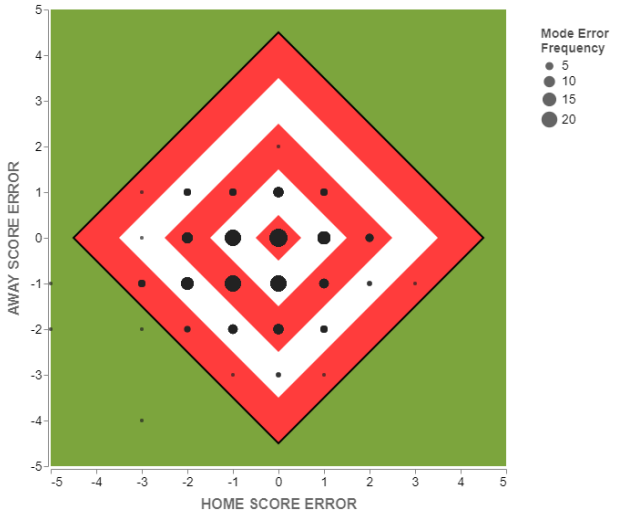


Fig. 7. Target featuring actual score as center with plots representing top rated prediction errors

Total Abs. Error to Top Rated Score	# of matches	% of matches
< 0.5	24	12.6
0.5 - 1.5	61	32.1
1.5 - 2.5	58	30.5
2.5 - 3.5	32	16.8
3.5 - 4.5	11	5.8
> 4.5	4	2.1

TABLE III
TOTAL ABSOLUTE ERROR TO TOP RATED SCORE

The circles plotted in Figure 7 do not correspond to the points in the same regions of the target in Figure 6. That's because errors relative to the mean score may be different from errors relative to the top rated score. Interesting enough, with forecasts being generated by independent Poisson r.v.'s, the top rated score always equals the truncated average score. This means that a 1.9-0.9 average score prediction would correspond to a 1-0 top-rated prediction, which, in turn, would account for a 1.8 difference between the total errors regarding comparisons (A) and (B). In that case, a match score of 1-0 would hit the bullseye in (B), but only make the third region in (A).

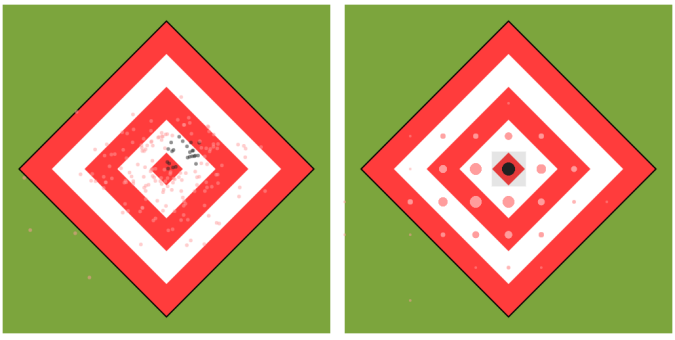


Fig. 8. Two targets and the corresponding regions

Figure 6 naturally compares 2-dimensional forecasts to results, with target center meaning perfect accurate forecasts

and errors being deviations from the center. Surprisingly, we haven't found this representation in previous works on forecasting evaluation and much certainly not in football score forecasting. Figure 7 doesn't seem so natural. It probably should not be represented as a target, except to enable comparison between the different types of errors. That connection takes place in Figure 8, which has both targets (A) and (B) side by side and illustrates the corresponding regions explained before.

A quick look at Figure 7 would suggest there is huge bias to the left and below. There are many more circles plotted to the lower left corner of the target than to the upper right, meaning both home and away top rated prediction errors tend to be negative, which, in turn, means that top rated prediction tends to underestimate the actual score. In fact, home score errors spread from -5 to 3, averaging -0.505, while away score errors spread from -4 to 2 and averaged -0.568.

When we look at Figure 6 though, the bias is not that clear. One may find some intuition on the same type of bias when notices all four out-of-target spots are located down and left, but the mass of points doesn't look biased either way. In fact, the average errors are 0.000 for home scores and -0.029 for away scores, no significant bias.

Top rated prediction bias can be easily explained (and expected) from the mean prediction lack of bias. As stated before, with independent Poisson r.v.'s the top rated prediction score always equals the truncated mean predicted score, implying that top rated prediction error will be lesser or equal than mean prediction error regardless of actual match result.

It's also easy to see in figures 6 and 7 that home score variance looks higher than away score variance. In fact they are. 1.461 to 0.853 regarding mean prediction errors and 1.606 to 0.945 regarding top rated prediction errors. That's the same intuition we had when discussing the shapes of the targets. If we were to standardize the errors, the resulting plot would be equivalent (in the sense of matching the corresponding plot regions) to plotting the original errors into a diamond target with horizontal diagonal larger than the vertical one.

V. Further Visual Analysis

After presenting our evaluating metrics and visual tools, we explore it further seeking for insight. For instance, we would like to check if all teams in the season were equally predictable.

For each team we check over each match it played for (i) the actual score rating, (ii) the probability forecasted to the actual score and (iii) the sum of the absolute errors for both home and visiting teams scores. Remind that the model uses a team's defence parameter to forecast the opposing team score. Thus, home and visiting scores depend on both teams' parameters.

Figures 3 and 5 were created from this analysis, filtering best and worst performing teams in the corresponding charts. Figures 9 and 10 are different from all the previously proposed charts in the sense that the plots represent teams and not scores nor any match score evaluator.

Figures 9 and 10 are represented with bubble charts, each team corresponding to a bubble, with x and y position and size determined by the three evaluators proposed. It's easy

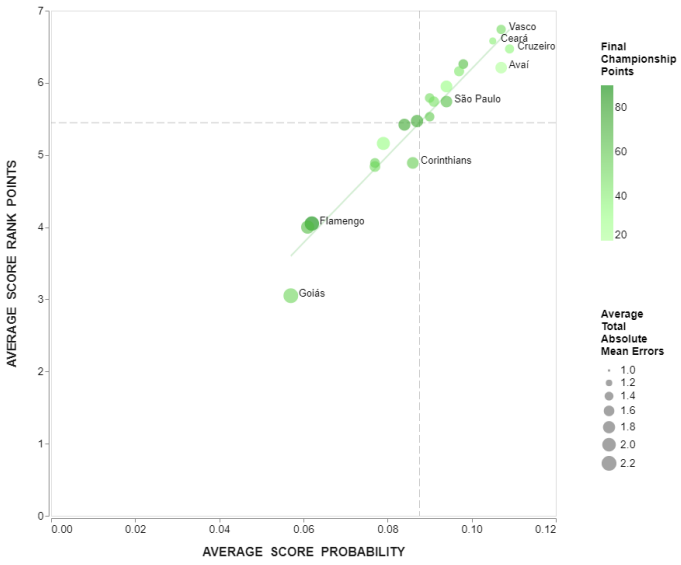


Fig. 9. 4-metric comparison by team. Color refers to championship points.

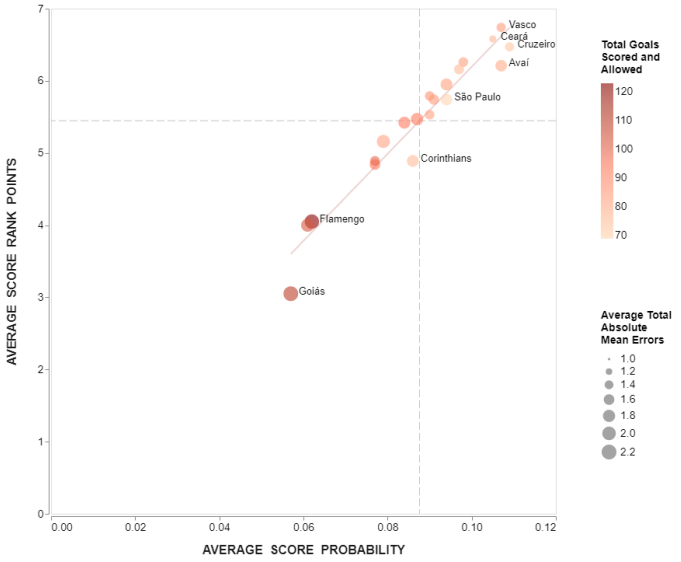


Fig. 10. 4-metric comparison by team. Color refers to goals scored and allowed.

to see a correlation between the three metrics, plots almost over a line, big circles to the bottom-left, small circles to the upper-right. The two figures differ only by the bubbles color, from which luminance is used to encode one championship stat: in Figure 9 it's the the number of championship points earned while in Figure 10 it's the total sum of goals scored and allowed by the team throughout the season. Interesting to note that the latter stat looks correlated to chart position (x and y variables) and size, while the former doesn't.

Analysing the charts we can conclude that the model performed quite differently among teams. Goiás has been the more unpredictable with honours, leading all three categories. As for the more predictable, Ceará, Cruzeiro and Vasco each led one of the three evaluators.

VI. Discussion and Conclusion

The results from the last couple of sections surely give us much information regarding the forecasts. As explained in section II, no single evaluator can determine the quality of a forecast and thus, those different evaluations may account for distinct aspects of its quality. Our evaluation methodology relies on four different metrics and four corresponding visual tools. The first one relates to how well the actual score was forecasted among all possible scores, with Figure 2 giving the visual support. The second metric corresponds to the exact probability predicted to the actual score, with Figure 4 as the visual analytic support. The other two metrics relate to how close/far the actual score was from some transformation of the forecast. Figures 6 and 7 plot distance and orientation and convey bias and variance for the corresponding metric.

That was the major goal of this work and we feel it has been achieved. The proposed set of metrics seem to give good information about forecasts performance and the developed visualization tools do help assessing and understanding those metrics. On the other hand, though we can use all the collected information as grades to the performance of the forecasting model, it is not possible to conclude whether the model performed well or not, because we lack both benchmark grades and other models to compare with. There is no other model, to our knowledge, that ran through this set of matches and forecasted (and published) each score probability.

As for the other visualizations, they are not used in an 'evaluation framework', but have been embedded here for storytelling purpose. Figure 1 helps introducing the forecasting model and depicts the ideas of uncertainty and probability distribution. Figure 8 explores the claim that different evaluators may lead to opposing conclusions on the goodness of a forecast. The other figures give tournament contextualization and uncover the high variation on performance when forecasting scores from different teams.

Finally, some considerations on future works. First (and of major relevance to this work) is the need to evaluate other forecasting models for comparison. We shall run the other models for the same set of matches to compare their performances using our proposed metrics. Better yet, would be to have one or more benchmark models, so that all other models could get compared to already known benchmark results and not only among themselves. Second, we should run the models on other sets of games, either from other tournaments or seasons, and check the performances. Of course performances may vary a lot not only between models but also between those sets of matches. One would expect some tournaments to be 'more predictables' than others, which could lead to consistent better performances from the forecasting models. By doing that we may shed some light on how the models perform over time and space, which could yield yet more visual designs. Third, the proposed visualization tools could be adapted to outcomes forecasts, the win-draw-loss triplet. There's much more information available and forecasting models running in that context. And while there's already plenty use of different metrics to compare the models, we have found few (if any) interesting visual tools to help assessing those comparisons.

Acknowledgment

The authors would like to thank the reviewers for their valuable comments. This work has been funded by FGV.

References

[1] "top-10 most popular sports in the world 2020" [online], available: https://sportytell.com/sports/most-popular-sports-world/. (accessed Sep.5, 2020).

[2] L. Rebeggiani and J. Gross, "Chance or ability? the efficiency of the football betting market revisited." MPRA, Jun 2018.

[3] G. Angelini and L. De Angelis, "Parx model for football match predictions," Journal of Forecasting, vol. 36, no. 7, pp. 795–807, 2017.

[4] P. Robberechts, J. V. Haaren, and J. Davis, "Who will win it? an in-game win probability model for football," CoRR, vol. abs/1906.05029, 2019.

[5] A. Carvalho, "An overview of applications of proper scoring rules," Decision Analysis, vol. 13, Nov 2016.

[6] M. J. Maher, "Modelling association football scores," Statistica Neerlandica, vol. 36, no. 3, pp. 109–118, 1982.

[7] A. J. Lee, "Modeling scores in the premier league: Is manchester united really the best?" CHANCE, vol. 10, no. 1, pp. 15–19, 1997.

[8] M. J. Dixon and S. G. Coles, "Modelling association football scores and inefficiencies in the football betting market," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 46, no. 2, pp. 265–280, 1997.

[9] A. Constantinou, "Dolores: a model that predicts football match outcomes from all over the world," Machine Learning, May 2018.

[10] A. Constantinou and N. Fenton, "Evaluating the predictive accuracy of association football forecasting systems," 2011.

[11] ——, "Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models," Journal of Quantitative Analysis in Sports, vol. 8, Jan 2012.

[12] E. Wheatcroft, "Evaluating probabilistic forecasts of football matches: The case against the ranked probability score," Aug 2019.

[13] J. J. Reade, C. Singleton, and A. Brown, "Evaluating strange forecasts: The curious case of football match scorelines," SSRN Electronic Journal, Mar 2019.

[14] J. Foulley and G. Celeux, "A penalty criterion for score forecasting in soccer," Jun 2018.

[15] C. Perin, R. Vuillemot, C. D. Stolper, J. T. Stasko, J. Wood, and S. Carpendale, "State of the art of sports data visualization," Computer Graphics Forum, vol. 37, no. 3, pp. 663–686, 2018.

[16] J. L. S. Malqui, N. M. L. Romero, R. Garcia, and H. A. ana João L. D. Comba, "How do soccer teams coordinate consecutive passes? a visual analytics system for analysing the complexity of passing sequences using soccer flow motifs," Computers Graphics, vol. 84, pp. 122 – 133, 2019.

[17] M. Stein, J. Häussler, D. Jäckle, H. Janetzko, T. Schreck, and D. A. Keim, "Visual soccer analytics: Understanding the characteristics of collective team movement based on feature-driven analysis and abstraction," ISPRS Int. J. Geo-Information, vol. 4, pp. 2159–2184, 2015.

[18] D. B. Coimbra, T. T. d. A. Tiburtino, A. C. Telea, and F. V. Paulovich, "The shape of the game." Conference on Graphics, Patterns and Images, 31. (SIBGRAPI), Oct-Nov 2018.

[19] L. McNabb and R. S. Laramee, "Survey of surveys (sos) - mapping the landscape of survey papers in information visualization," Computer Graphics Forum, vol. 36, no. 3, pp. 589–617, 2017.

[20] I. Ruginski, A. Boone, L. Padilla, L. Liu, N. Heydari, H. Kramer, M. Hegarty, W. Thompson, D. House, and S. Creem-Regehr, "Non-expert interpretations of hurricane forecast uncertainty visualizations," Spatial Cognition Computation, vol. 16, pp. 154–172, Jan 2016.

[21] X. Pu and M. Kay, "A probabilistic grammar of graphics," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2020, pp. 1–13, [online] https://doi.org/10.1145/3 313 831.3 376 466, (accessed Jul. 21, 2020).

[22] J. Boice. "how our club soccer predictions work", fivethirtyeight.com [online], available: https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/. (accessed Jul. 7, 2020).