

An evaluation of quantitative and qualitative methods of predicting the 2002 FIFA World Cup

P. O'Donoghue, W. Dubitzky, P. Lopes, D. Berrar, K. Lagan, D. Hassan, A. Bairner and P. Darby

University of Ulster, Jordanstown, UK

The purpose of the current study was to attempt to predict the 2002 FIFA World Cup using a broad spectrum of methods. The methods ranged from the qualitative to the quantitative. The predictions were evaluated by awarding 1 point for each round 2 participant identified, 1 point for each group winner or runner-up identified, 2 points for each quarter-finalist identified, 3 points for each semi-finalist identified and 5 points for correctly identifying the winner of the 3rd place play-off or the final. This gives a maximum mark of 78 for any prediction. It was not possible for all methods to use the same data (although this would have been desirable) due to the very nature of the methods. Some relied on objective facts, whereas others used more personal subjective knowledge and experience.

There were two purely qualitative methods used – an individual expert and a focus group. The individual expert had research experience in soccer culture, history and politics as well as an interest in the evolution of international soccer. The prediction was based on rich subjective information derived from a variety of media sources as well as perceptions relating to relevant factors such as home advantage, the importance of soccer within the nation, key individual player characteristics and performance in major tournaments. The focus group was chaired by another university academic with research interests in sport and politics. There were four other soccer enthusiasts. The knowledge used was similar to that of the individual expert, except that the focus group was based on shared knowledge with the need to negotiate a joint prediction. The individual expert predicted Italy to defeat France in the final, while the focus group predicted Argentina to defeat Italy in the final. A third prediction used bookmakers' odds quoted on 30 May 2002 for each team to win its group as well as the odds quoted to win the tournament.

First, the group winners and runners up were deemed to be the first and second favourites of the bookmakers to win the group. France at 4/1 was selected as the tournament winner, Italy at 5/1 was selected as runner up as the most favoured team in the other half of the draw to France. Argentina at 5/1 was selected as the third placed team in the tournament, with Spain at 8/1 being the most favoured team in the other quarter of the draw to be the 4th semi-finalist. For each quarter of the draw, the odds were used to select the most likely quarter finalists to complete the prediction.

A probabilistic neural network learned about the relationship between relevant match factors (including performance in the last 10 internationals) from 7965 international soccer matches including pre-tournament friendlies. The known factors of the group matches of the 2002 FIFA World Cup were then entered into the trained neural network to determine a prediction for each match. This allowed group outcomes to be predicted and results for expected knockout matches to be determined in the same manner. This method predicted that Germany would defeat France in the final.

A study of 282 matches from World Cups, European Championships and Copa Americas since 1993 (when the FIFA world rankings were established) was used to determine the effect of FIFA world ranking, distance travelled to the tournament, differences in recovery days since previous match, as well as switching countries in the middle of a tournament (this could only use data from Euro 2000) on the score of matches. These effects were programmed into a simulator that was run 2000 times, predicting Brazil as the most likely winner with Italy as the most likely runner-up.

Two statistical methods used FIFA world ranking and distance travelled to predict match outcomes. The other two factors used in the simulator could not be normalized and so were excluded from the statistical analyses. In the first approach, linear regression was used, predicting Japan to defeat Korea in the final. In the second approach, discriminant function analysis was used to predict wins, draws and losses in the group games, while binary logistic regression was used to predict the victors of knock-out matches. This resulted in a prediction that France would defeat Germany in the final.

A further prediction was made using random scorelines with respect to the higher ranked team by FIFA within matches from the 282 matches on which the simulator was based. This resulted in a prediction that Nigeria would defeat Mexico in the final. A final prediction ran a commercially produced computer game of the FIFA 2002 World Cup on a games console allowing the computer to play itself on 32 occasions. The system favoured France as the tournament winners with Spain as the most likely runners-up.

Using the evaluation scheme, the simulator was the most successful prediction (40), followed by linear regression (33), discriminant function analysis/logistic regression (32), the computer game (30), the probabilistic neural network (27), the focus group (27), bookmakers' odds (20), the individual expert (19) and random scores (11).

Given that these scores are out of 78, it is clear that predicting the outcome of the World Cup before it commences remains a difficult and complex task. Furthermore, both quantitative and qualitative methods had advantages and disadvantages. While the simulator successfully predicted the tournament winner, it only predicted three of the quarterfinalists. The focus group, on the other hand, predicted four of the quarter-finalists.