# why spain will win...

**...maybe? Dr Ian Hale**, senior lecturer in statistics at the University of Salford, discusses how mathematical models of football matches are used in the gambling industry – and sportingly puts his neck on the line by supplying his own predictions for the World Cup 2010

PREDICTING football results is a rapidly growing area of academic interest. Economists use models to assess the efficiency of betting markets, operational researchers use models to experiment with the various effects of tournament design, and statisticians showcase their proficiency with advanced statistical techniques by modelling the intricacies of football data.

It is not, of course, just academics who are mining the archives of football scores. Bookmakers live and breathe football prediction models – as do the more committed flutterers. Mistakes cost money and jobs, whilst finding a small advantage can carry great rewards.

## BETTING MARKETS

In academia, the most common application of football forecasting models is to test for betting market efficiency. The Efficient Markets Hypothesis (EMH) is a cornerstone of financial theory and, in its simplest form, states that an investor should not be able to consistently obtain returns above the average. Finding a forecasting model of football that can generate better-than-average – or even positive – returns usually results in a publication for the academic as an example of a violation of the EMH, but the proprietary nature of the models means that the published ones rarely (if ever) represent the very best models, and even less often generate positive returns consistently.

**online**

**Meet E&T's panel of predictors in our exclusive video at http://kn.theiet.org/magazine/videos**

**Follow their World Cup match predictions day-by-day from 11 June, and add your own contributions, by visiting http://engtechmag.wordpress.com**

> 'The proprietary nature of prediction models means that the published ones rarely (if ever) represent the very best models'

The best performing models are the reserve of the gambling industry. It is paramount for a bookmaker to set odds at a value that realistically represent the probabilities of a match being won, drawn or lost. If the bookmaker fails to do this, it will risk huge losses.

For instance, Asian bookmakers would think nothing of taking an individual bet of US$200,000 – and regularly receive bets of $400,000 – and a typical weekend in the English Premier League typically attracts $500m turnover in Asia. With such tides of cash being wagered, it is not surprising that bookmakers make use of every possible tool at their disposal – one of them being mathematical models.

A mathematical model is not typically used on its own to set odds. An expert odds-setter is employed to adjust the model-generated odds given any extra information. For example, a typical model might take into account recent results of a team and each team's position in the league only. The odds-setter will then adjust these predicted odds to account for, say, an injury to a star player.

On the other side of the market, are the bettors. Specialist companies offer services to advise clients which bets to make. Maciej Jarowek is a betting consultant for AsianConnect88.com, a betting broker. I asked him for his thoughts on the use of mathematics in the industry. As a bookmaker does, he uses a model to give him a starting point and then adjusts the odds as he sees fit.

Jarowek is an expert on Polish football, and when he finds a discrepancy between his odds and those of the bookmakers, he can choose whether the discrepancy is enough to advise placing a bet. He says the biggest advantage professional gamblers have over the bookmaker is that they do not have to bet on every match – they can pick and choose. Given that Jarowek is one of many professional gamblers working for AsianConnect88.com, and that there are numerous such agencies in operation, it appears there is considerable scope for beating the market.

Compared with other sports, the result of a football match is relatively difficult to predict. Some academics believe this is one of the sources of the popularity of football – fans never really know the outcome of a match before it has happened making watching the game an exciting proposition.

Pundits make careers based on this uncertainty of outcome and spend endless hours contemplating each and every possible event in a match. There are countless upsets in domestic football every weekend, and the World Cup throws up its fair share of giant-killings. Given this inherent unpredictability, how can we make predictions?

There are two broad approaches to modelling football match outcomes. First, one can predict the result directly – whether the result will be a win, a draw or a loss. In this case, the outcome (win, loss, draw) is an ordinal variable (a win is better than a draw which is better than a loss). The model of choice here is an ordinal regression one, such as ordered probit. The output from such a model is the probability of each outcome, so that for an upcoming match one can use the model to calculate the probability of a win, draw or loss.

The second approach for modelling a match outcome is an indirect method where the analyst models the exact score of the game. In this case, the analyst estimates the probability of each possible number of goals scored by each team. One can then infer the probability of a team winning, drawing or losing the match by summing the relevant exact score probabilities.

## AN ORDERED PROBIT

Evidence suggests very little difference in performance of each approach in modelling match outcome. Here, I build a model of the first type, namely an 'ordered probit' model.

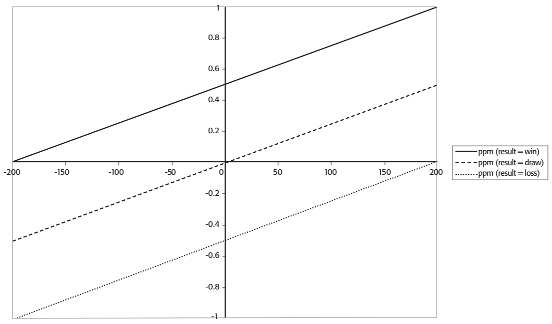
The ordered probit model can be used to estimate the probability of the three outcomes of a match. To do this, it uses information on each team. For instance, it seems reasonable that a team that has won its last three matches has a higher probability of winning its next match than a team that has lost its last three matches. Similarly, a team that is ranked higher than the opposition has a higher probability of winning the match than does the lower-ranked team. The amount of the win probability changes, given these nuggets of information are governed by the model parameters. In a statistical model, the parameters are estimated to best explain what has happened in the past.

My model here is based on nearly 9,000 international match results over an eight-year period. The information and the effects on the probability of a team winning a match are:

- **Venue – home/away or neutral.** A team is more likely to win if at home.
- **Distance from capital city to game location.** A team is more likely to win the shorter the distance travelled (by fans and the team itself).
- **Difference in world ranking between the teams.** A higher ▶

**Below: A formula to create a past-performance metric, which takes into account the fact that past results should be weighted relative to the result and relative to the strength of the opposition**
**Right: The past performance metric plotted on a graph.**

$$ppm(result) = \begin{cases} 1 + \dfrac{rankdiff - 204}{408} & \text{if } result = win \\ 0.5 + \dfrac{rankdiff - 204}{408} & \text{if } result = draw \\ 0 + \dfrac{rankdiff - 204}{408} & \text{if } result = loss \end{cases}$$



**Past Performance Metric**

◀ ranked team is more likely to win.

■ **Change in world rankings for each team during the previous 12 months.** An improving team is more likely to win.

■ **Type of match, namely: major tournament (World Cup or confederation championship), minor tournament (other FIFA-sanctioned one), qualifier or friendly.** Better teams take major tournaments more seriously.

■ **Past match results.** For each game, I use the previous eight results for each team (on average a national team plays eight matches in a 12-month period). In addition to the results, I know the world rankings for each of the eight opposition teams at the time of the game.

Using the information of past results needs a little extra thought. A 1-0 win for a team ranked 200 versus a team ranked 4 clearly represents a better performance than if the teams were ranked 200 and 198 respectively. As a consequence, the past results should be weighted relative to the result and relative to the strength of the opposition. A past performance metric (ppm), which captures this relationship, is given by the formula at the top (left) of this page.

This ppm goes from '–1' to '1', where '–1' represents the worst result possible – the top-ranked team losing to the bottom-ranked team – and '1' represents the worst-ranked team beating the best-ranked team. Note that 204 is the maximum value of the ranking of any team during the period under consideration. The past performance metric is plotted at the top (right) of this page.

## Results of 100,000 World Cup 2010 simulations

| Team | Number of wins | My Estimated Probability | FIFA ranking | Bet365 odds | Bet365 scaled probabilities |
|---|---|---|---|---|---|
| SPA | 11,633 | 0.1163 | 2 | 4-1 | 0.1751 |
| BRA | 10,256 | 0.1026 | 1 | 5-1 | 0.1459 |
| FRA | 9,108 | 0.0911 | 10 | 18-1 | 0.0461 |
| HOL | 8,480 | 0.0848 | 4 | 12-1 | 0.0673 |
| ARG | 7,787 | 0.0779 | 7 | 7-1 | 0.1094 |
| ENG | 5,044 | 0.0504 | 8 | 6-1 | 0.1251 |
| GER | 4,772 | 0.0477 | 6 | 12-1 | 0.0673 |
| NIG | 3,939 | 0.0394 | 20 | 150-1 | 0.0058 |
| POR | 3,779 | 0.0378 | 3 | 28-1 | 0.0302 |
| GRE | 3,723 | 0.0372 | 12 | 150-1 | 0.0058 |
| ITA | 3,689 | 0.0369 | 5 | 12-1 | 0.0673 |
| MEX | 3,432 | 0.0343 | 17 | 80-1 | 0.0108 |
| CAM | 3,330 | 0.0333 | 19 | 100-1 | 0.0087 |
| SLN | 2,751 | 0.0275 | 23 | 250-1 | 0.0035 |
| AUS | 2,457 | 0.0246 | 20 | 150-1 | 0.0058 |
| SER | 2,291 | 0.0229 | 16 | 66-1 | 0.0131 |
| IVC | 2,271 | 0.0227 | 27 | 33-1 | 0.0257 |
| URU | 1,613 | 0.0207 | 18 | 100-1 | 0.0087 |
| PAR | 1,613 | 0.0161 | 30 | 80-1 | 0.0108 |
| GHA | 1,551 | 0.0155 | 32 | 66-1 | 0.0131 |
| SWI | 1,506 | 0.0151 | 26 | 200-1 | 0.0044 |
| DEN | 1,250 | 0.0125 | 35 | 150-1 | 0.0058 |
| ALG | 1,132 | 0.0113 | 31 | 350-1 | 0.0025 |
| USA | 786 | 0.0079 | 14 | 80-1 | 0.0108 |
| SLK | 396 | 0.0040 | 38 | 250-1 | 0.0035 |
| HON | 334 | 0.0033 | 40 | 1000-1 | 0.0009 |
| SA | 245 | 0.0025 | 90 | 150-1 | 0.0058 |
| SK | 170 | 0.0017 | 47 | 200-1 | 0.0044 |
| JAP | 142 | 0.0014 | 45 | 350-1 | 0.0025 |
| CHI | 31 | 0.0003 | 85 | 66-1 | 0.0131 |
| NZ | 26 | 0.0003 | 78 | 2000-1 | 0.0004 |
| NK | 3 | 0.0003 | 106 | 1500-1 | 0.0006 |

## (SSHHH – I PREDICT THE WORLD CUP WINNER)

The ordered probit model described above can be used to estimate the probability of each outcome (win/draw/loss) in one match. A bookmaker might use these probabilities to help inform odds-setters, whilst a bettor might use them to assess whether a bet should be placed.

In order to predict the winner of the World Cup 2010, one needs to predict the winner of a series of matches. To do this, I use simulation, and have written code that uses the fitted model of match outcome to simulate the entire tournament. The World Cup starts with eight mini-leagues (groups) of four teams, playing each other once. The first- and second-placed teams in each group then progress to the knockout stages. This tournament structure affects the probability of a team winning.

For example, in the upcoming World Cup, the most likely teams to progress from Group G are Brazil and Portugal. The most likely team to win Group H is Spain. The winners (runners-up) of Group H then meet the runners-up (winners) of group G in the first knockout round. Thus Spain are almost guaranteed a difficult match in the first knockout round. Compare this to Group A favourites, France, who will most likely face the runner-up of Group B. Given Group B favourites, Argentina, are much stronger than the other teams in the group, France (and Argentina) are likely to have less difficult matches in the first knockout stage than Spain, Brazil and Portugal.

It is clear then, that predicting the winners of the tournament overall is not just a case of picking the best team. One needs to take into account the effect of the tournament structure. The results of 100,000 simulated tournaments are shown in the table, left.

Despite their seemingly difficult draws, Spain and Brazil are still first and second in terms of number of wins. The number of wins can easily be converted into a predicted probability of winning the tournament for each team which is shown in the third column of the table (left).

It is interesting to compare the predicted win probabilities with the FIFA World Rankings. France are big movers upwards – possibly because of their easier draw, whilst Portugal fall to ninth favourites, despite being the third-ranked team in the world.

Given that the most frequent use of such a model is to compare the probabilities with those of bookmakers, the fifth and sixth columns of the table give the odds and implied probabilities from Bet365, as at the date I made my predictions (14 May 2010). I have to admit, when I saw the similarities, I was pleased – it confirms the model and simulation exercise give sensible results. However, closer inspection reveals some discrepancies. Disclaimer time: Please note I do not suggest you use these odds to place a bet. The bookies really do know what they are doing and incorporate far more information I have used to inform their odds – they are also paid much more than I am for doing this sort of thing!

It appears the bookmakers think Spain and Brazil are more likely to win than I do,

> 'I don't suggest you use these odds to place a bet. The bookies really do know what they are doing and incorporate far more information than I have!'

suggesting these are actually not good teams to back (at these odds). This might be because they are protecting themselves against the market which is keen to be on these teams – a similar story is true for England.

France, on the other hand, offers good value – the model suggests the probability of victory is nearly twice the probability assumed by the bookmaker. It will be hard to take for any Republic of Ireland fan, if France were to go on to win the World Cup, given that they were knocked out of the World Cup by France forward Thierry Henry's 21st-century spin on the 'hand of God' goal.

So here it is – I put my neck on the line (at least the model does) to give you some tips, based purely on statistics (not my opinion):

■ Spain are the tournament favourites, but may not offer value for money. (The more astute reader will recognise this as a bit of a politician's answer: if Spain win, I can say "I told you they were favourites", whilst if they lose, I can say "I told you they weren't a good bet – there was an 88.4 per cent probability of them not winning"!)

■ France to win is a good bet.

I also persuaded Jarowek to give me a tip:

■ Brazil not to win. (Typically reserved and cautious).

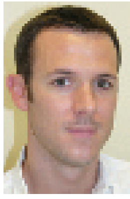
The model's predictions for the latter stages of the tournament are:

**QF1:** Holland vs Brazil
**QF2:** France vs England
**QF3:** Germany vs Argentina
**QF4:** Italy vs Spain
**Semi-final line up:**
**SF1:** Brazil vs France
**SF2:** Argentina vs Spain
**Final:**
Brazil vs. Spain

Pretty obvious really! I should say that, although my reputation as a statistician rests on these tips being reasonably good, I will be cheering on my native England no matter whom the opposition. I wait in anticipation for kick-off on 11 June... ■

## HOT TIPS: THE E&T PANEL GIVES ITS WORLD CUP PREDICTIONS

***E&T* has gathered a panel of four experts – people with a professional or personal interest in the science of predictions – to kindly stick their neck out and state who they judge will be involved at the latter stages of the tournament – and who, above all, will claim the cup.**



### Dr Ian McHale

Dr Ian McHale is a Senior Lecturer in Statistics at the University of Salford. His research interests include statistics in sports and the study of gambling markets. If he had known footballers and golfers earned so much money, he would have ignored his parents and concentrated less on maths at school and more on his sporting activities. Still, mathematical models no doubt came in handy for the predictions here.

www.salford.ac.uk

**Prediction**
**Winner:** Spain
**Finalists:** Spain and Brazil
**Semi Finalists:** Spain, Brazil, Argentina and France.

### David Williams

David Williams is head of PR and official spokesman at Ladbrokes plc, a leading retail bookmaker in the UK, Ireland, Belgium and Spain with over 2,700 high-street betting shops. Ladbrokes' share of the global betting and gaming market brings in revenues of over £1bn, with £15bn worth of stakes placed with the company in 2009. Utilising state-of-the-art technology, Ladbrokes is able to react quickly to fluctuations in the global gambling market, and Williams tapped in to this technology to select his final four teams.

http://sports.ladbrokes.com/en-gb

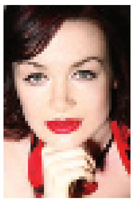
**Prediction**
**Winner:** Spain
**Finalists:** Spain and Brazil
**Semi Finalists:** Spain, Brazil, England and Argentina

### Chris Day

Goalkeeper Chris Day started his career at Tottenham Hotspur before moving to Crystal Palace and then to Watford under one-time England manager Graham Taylor. Spells with QPR, Oldham and Millwall followed before Day joined Stevenage Borough FC. Day has based his predictions on his football knowledge – including the surprise elimination of tournament favourites Spain, reasoning that the runner-up in Group G is likely to be Portugal, who would then give Spain a real test in the next match. With Ronaldo in their side, Portugal will be no pushovers.

www.stevenageborofc.com

**Prediction**
**Winner:** Brazil
**Finalists:** Brazil and Argentina
**Semi Finalists:** Brazil, Argentina, England and Holland



### Laura Daligan

'Psychic, artist and pin-up' Laura Daligan specialises in tarot card readings and dream interpretations. She is also a colour healer, using her artistic skills to help people find the right colours for their wardrobes, homes and personal style. "I feel England will do well," she says. "As semi finalists, I am getting England, Brazil; Argentina will do well but their game is not stable. Spain will also do well. I'm getting an already-established world cup team, the colour red and a macho manager for the winning team. That suggests Spain. Korea and Japan will also have some interesting games."

www.lauradaligan-psychic.com

**Prediction**
**Winner:** Spain
**Finalists:** Brazil and Spain
**Semi Finalists:** England, Brazil, Argentina and Spain

**How did they do?**
When the World Cup kicks off on 11 June, so will the ***E&T* World Cup blog**. Visit the blog to see all four experts on our panel pit their skills against **one game a day**, from the opening match through to the final. **Follow the action online and see how they do. Why not join in and make your own prediction?**

http://engtechmag.wordpress.com/

*Disclaimer*
*All predictions are for entertainment only. Any mortgages lost or other sums squandered based on the information supplied are the reader's own responsibility*