September 2009
Regulars

Understanding uncertainty: Football crazy

by David Spiegelhalter and Yin−Lam Ng

This article is adapted from material on the Understanding Uncertainty website.

Who gets the trophy? Image: Jameboy.

On May 22nd 2009 the English Premier league had ten more matches left to play, with West Bromwich Albion at the bottom of the league with 31 points, and Manchester United at the top with 87 points. The bottom three teams would be relegated: West Brom were certain to be one of them, but four teams could possibly join them. ManU were certain to end up the top team and so were not expected to play their strongest team in their away match against Hull City, who were one of the teams up for relegation and so had everything to play for.

The BBC Radio 4 programme *More or Less* had heard of the work we had been doing on modelling European football results, and they asked us to produce predictions for these final ten matches using a statistical method that could be explained on the radio. Quite a tricky challenge, particularly knowing that the predictions would be announced before the matches and then afterwards compared to what really happened and how well other pundits did. But using some basic probability theory we can quite easily produce a reasonable probability for all the possible results of a game.

We can start by looking at the state of the league on May 22nd 2009, with goals for and goals against.

Team	Points	Goals for	'Attack strength'	Goals against	'Defence weakness'
Man United	87	67	1.46	24	0.52
Liverpool	83	74	1.61	26	0.57
Chelsea	80	65	1.41	22	0.48
Arsenal	69	64	1.39	36	0.78
Everton	60	53	1.15	37	0.80
Aston Villa	59	53	1.15	48	1.04
Fulham	53	39	0.85	32	0.70
Tottenham	51	44	0.96	42	0.91
West Ham	48	40	0.87	44	0.96
Man City	47	57	1.24	50	1.09
Stoke	45	37	0.80	51	1.11
Wigan	42	33	0.72	45	0.98
Bolton	41	41	0.89	52	1.13
Portsmouth	41	38	0.83	56	1.22
Blackburn	40	40	0.87	60	1.30
Sunderland	36	32	0.70	51	1.11
Hull	35	39	0.85	63	1.37
Newcastle	34	40	0.87	58	1.26
Middlesbrough	32	27	0.59	55	1.20
West Brom	31	36	0.78	67	1.46

The Premier League table on May 22nd 2009 after 37 games each.

The average number of goals scored, and therefore also the average number of goals conceded, is 46. If we divide the number of goals scored by 46, we get a measure of the attack strength of a team: Arsenal's 64 goals divided by 46 gives 1.39, which shows they have scored 39% more goals than average. If we divide the number of goals conceded by 46 we get a measure of defence weakness: Stoke City's 51 conceded goals divided by 46 gives 1.11, which shows they let in 11% more goals than average.

We also need two other pieces of information: the average number of goals scored per match by a home team is 1.36, while for an away team it's 1.06. Now suppose we want to predict the result of Hull vs Manchester United. We start by estimating how many goals Hull will score. They are playing at home, so if they were an average team, we would expect them to score 1.36. But they are not average: over the season they have scored

only 85% of the average number of goals, and so their attack strength is 0.85. Multiplying up we get 1.36 × 0.85 = 1.16. And their opposition is not average either: ManuU's defence weakness is 0.52, since they have conceded only 52% of the average. So we get a total of 1.36 × 0.85 × 0.52 = 0.60 expected goals by Hull, which does not look too good.

For Manchester United, the baseline is 1.06, the average number of goals scored by an away team. But by the time we adjust this for ManU's attack strength and Hulls' defence weakness, we get 1.06 × 1.46 × 1.37 = 2.12.

But, just like nobody has 2.4 children, nobody scores 2.12 goals this is only an expected value. It's the average number of goals scored by ManU if the match were played again and again, heaven forbid. But we can use what is known as the *Poisson probability distribution* to distribute 100% of probability across the possible number of goals. The distribution expresses the probability of a number of events occurring in a given time period, if the average rate of the occurrence is known and the events are independent. Thus we get probability distributions shown in the table below.

Team	Expected goals	0	1	2	3	4	5
Hull City	0.60	55	33	10	2	0	0
Man U	2.12	12	25	27	19	10	4

The percentage probability of each team scoring a specified number of goals in the match on May 24th 2009, using a simple Poisson model.

So, if the next match follows past performance, there is a 55% probability that Hull won't score at all, and 63% (100 − 25 −12) probability ManU will get at least 2 goals, even though playing away.

To get the probability of an actual result, we might assume that the goals scored by each team are independent, in the sense that if we knew how many ManU scored, it would not give us any additional information about Hull's performance. This is a strong assumption and we'll come back to it in a moment, but it means that to find, for example, the probability of a 0−2 result, which is the most likely outcome, we multiply 55% by 27% to get 15% (55/100 × 15/100 = 0.1485), so even the most likely result is still not very likely!

In fact there tends to be some correlation between teams' results, in the sense that matches have a tendency to be either high or low scoring, which we might call a "pitch effect". Estimating probabilities allowing for correlations is more complicated and requires special software: the *bivariate Poisson model* is popular and can be fitted using free programs. Yin−Lam Ng, in her Cambridge MPhil in Statistical Science project, fitted models to all major league results in Europe over the last 15 years, and the predictions below are based on the best model found.

Statistical models assume that past performance predicts future results, and do not take into account new factors. For example, Hull City are trying to avoid relegation, Manchester United are conserving their strength having already topped the league, and so it is possible that Hull City may stand a much better chance of winning than the 9% we have given them some people obviously thought so, as the odds offered by the bookies were more like 2 to 1 against, or a 33% chance of Hull winning.

Below is a table of the four most likely results for each match according to the statistical model.

Home	Away	Most likely	2nd most likely	3rd most likely	4th most likely	Actual result
Arsenal	Stoke	2-0 (14%)	1-0 (13%)	2-1 (9%)	3-0 (9%)	**4-1**
Aston Villa	Newcastle	**1-0 (10%)**	2-0 (10%)	2-1 (10%)	1-1 (10%)	**1-0**
Blackburn	West Brom	1-1 (10%)	2-0 (10%)	2-1 (10%)	1-1 (10%)	**0-0**
Fulham	Everton	0-0 (19%)	1-0 (16%)	0-1 (14%)	1-1 (13%)	**0-2**
Hull	Man United	0-2 (14%)	**0-1 (14%)**	1-2 (9%)	1-1 (8%)	**0-1**
Liverpool	Tottenham	1-0 (16%)	2-0 (15%)	3-0 (10%)	2-1 (9%)	**3-1**
Man City	Bolton	2-1 (10%)	1-1 (10%)	**1-0 (10%)**	2-0 (10%)	**1-0**
Sunderland	Chelsea	0-1 (20%)	0-2 (15%)	0-0 (13%)	1-2 (8%)	**2-3**
West Ham	Middlesbrough	1-0 (19%)	0-0 (14%)	2-0 (13%)	1-1 (11%)	**2-1**
Wigan	Portsmouth	**1-0 (17%)**	2-0 (14%)	0-0 (11%)	1-1 (10%)	**1-0**

The four most likely results for each match, with their percentage probability according to a bivariate Poisson model.

Note that the highest chance is 20%, and for most matches there's only a around 50% chance that any of these top four results will occur. So it's rather misleading to treat the "most−likely" results as predictions all this model does is produce (what we hope are) reasonable probabilities. If we add up the probabilities for results that lead to a win/draw/lose we get the probabilities shown below. Some of these become quite high, for example 72% for a home win in the Arsenal−Stoke match, but even these could not be considered as firm predictions.

Home	Away	Home win	Draw	Away win
Arsenal	Stoke	**72**	19	10
Aston Villa	Newcastle	**62**	21	17
Blackburn	West Brom	54	**23**	23
Fulham	Everton	35	35	**30**
Hull	Man United	9	19	**72**
Liverpool	Tottenham	**72**	20	9
Man City	Bolton	**59**	22	19
Sunderland	Chelsea	10	25	**65**
West Ham	Middlesbrough	**57**	28	15
Wigan	Portsmouth	**44**	32	25

The percentage probability of each result for the final ten matches of the Premier league, based on a bivariate Poisson model. The actual results are shown in bold.

The most likely results were read out on the *More or Less* broadcast on May 22nd, without any qualifying probabilities, somewhat to our consternation. They were also given on the BBC *More or Less* website , this time with probabilities (although we mistakenly said the Fulham−Everton most−likely 0−0 prediction with probability 10% whereas we should have said 19%, and Liverpool−Tottenham's most−likely prediction was given probability 10% instead of 16%.)

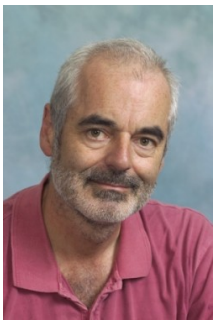
So what happened? The day of the matches was nerve−wracking, but when the results were announced we were very relieved to find that using our best predictions, we got nine results out of ten right in terms of win/draw/lose, and we also predicted two exact scores: Aston Villa−Newcastle (1−0) and Wigan−Portsmouth (1−0). This was particularly gratifying as Mark Lawrenson, the official BBC football expert, only got seven correct results, and only one exact score.

A contented Alex Ferguson whose ManU are current Premier League champions. Image: Austin Osuide

This is a very good result for statistics! But perhaps a bit lucky in particular it is very difficult to predict draws and it was rather fortunate that the most−likely 1−1 Blackburn−WestBrom score turned out to be a 0−0 draw, since a draw was not the most likely outcome. One possible advantage of the statistical method is that it is not influenced by emotion. For example, in the Hull−ManU match, Hull was considered as having some chance of a win, and Mark Lawrenson predicted a draw, but we went for a ManU win and were proved correct.

These types of models have been refined over the years and are now used by bookies and sports betting companies, who employ experienced statisticians and make use of the latest computational methods: in particular it is natural to extend our model to allow for a team's abilities to change over the season, and so discount historical evidence to allow recent performance to dominate. And, not surprisingly, they don't tell anyone exactly what they do! One thing you can bet on: simple models like those above will be very unlikely to out−perform the odds being offered by bookies, so you should not use them to spot good bets. We have heard that some people did make money from our predictions, and we have since been approached by people wanting to work with us on sports modelling, but I don't think we will take this up as a sideline it could be much too engrossing.

About this article

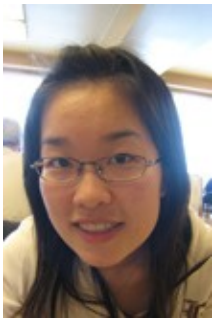


David Spiegelhalter

David Spiegelhalter is Winton Professor of the Public Understanding of Risk at the University of Cambridge.

Yin−Lam Ng was an MPhil student at Cambridge and has now joined Hong Kong Polytechnic University as a research assistant.

David and his team run the *Understanding uncertainty* website, which informs the public about issues involving risk and uncertainty.



Yin−Lam Ng

Plus is part of the family of activities in the Millennium Mathematics Project, which also includes the NRICH and MOTIVATE sites.