

## Life conference and exhibition 2010 Greg Becker

# Who will win the Premier League?



7-9 November 2010



---

# Team talk – setting expectations

---

- Practical statistics
  - Showing how a complex problem that can't be solved using traditional methods, can be solved using Monte Carlo methods
  - Touching on Bayesian statistics, Monte Carlo methods
- Model development process
  - Theoretical foundation
  - Data & data problems
  - Testing and refining a model – an iterative process
- Practical application in other areas of actuarial work

# Fixture List – a tale of two halves

## Theory

- What could be in a model?
- What data could be used?
- Lessons learnt from the World Cup
  - Article Written in The Actuary
  - Compared to reality
- Half time – with half-time entertainment

## Model in practice

- Model proposed
  - Why and how?
  - How would it have done in 2009-2010?
  - How would it have done in 2008-2009?
- Betting stats
- Actuarial lessons

# They say you should know your audience: Please clap or cheer when your team logo comes up



# After 100 matches so far this season, this is the points table:

- In 2007/8, Arsenal and Manchester United were leading, and while Manchester United went on to win, but both Arsenal and Chelsea had led the table later in the season
- In 2008/9, Chelsea and Liverpool were leading at this stage, and neither went on to win!
- In 2009/10, Chelsea was already leading by 2 points, although Manchester United was leading the table as late as 2/4/2010

Team	P	GD	PTS
1 Chelsea	10	24	25
2 Arsenal	10	12	20
3 Man Utd	10	10	20
4 Man City	10	3	17
5 Tottenham	10	1	15
6 West Brom	10	-3	15
7 Newcastle	10	5	14
8 Everton	10	2	13
9 Blackpool	10	-6	13
10 Fulham	10	1	12
11 Bolton	10	-1	12
12 Sunderland	10	-3	12
13 Liverpool	10	-4	12
14 Aston Villa	10	-4	12
15 Birmingham	10	-2	11
16 Stoke	10	-4	10
17 Wigan	10	-11	10
18 Blackburn	10	-3	9
19 Wolves	10	-6	9
20 West Ham	10	-11	6

**Betting on the leader at this stage of the season is not a sure bet!**

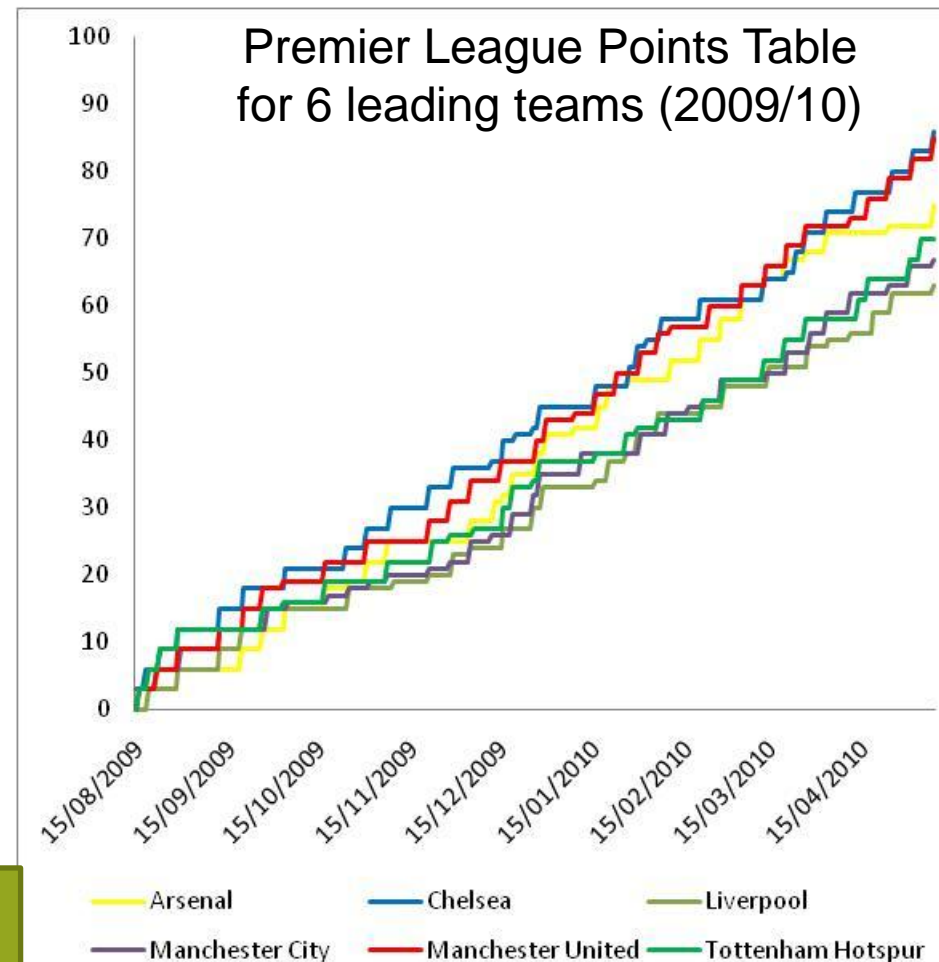
# Simplest prediction tool: The team leading is almost certainly the best team to bet on...

But many things should be taken into account

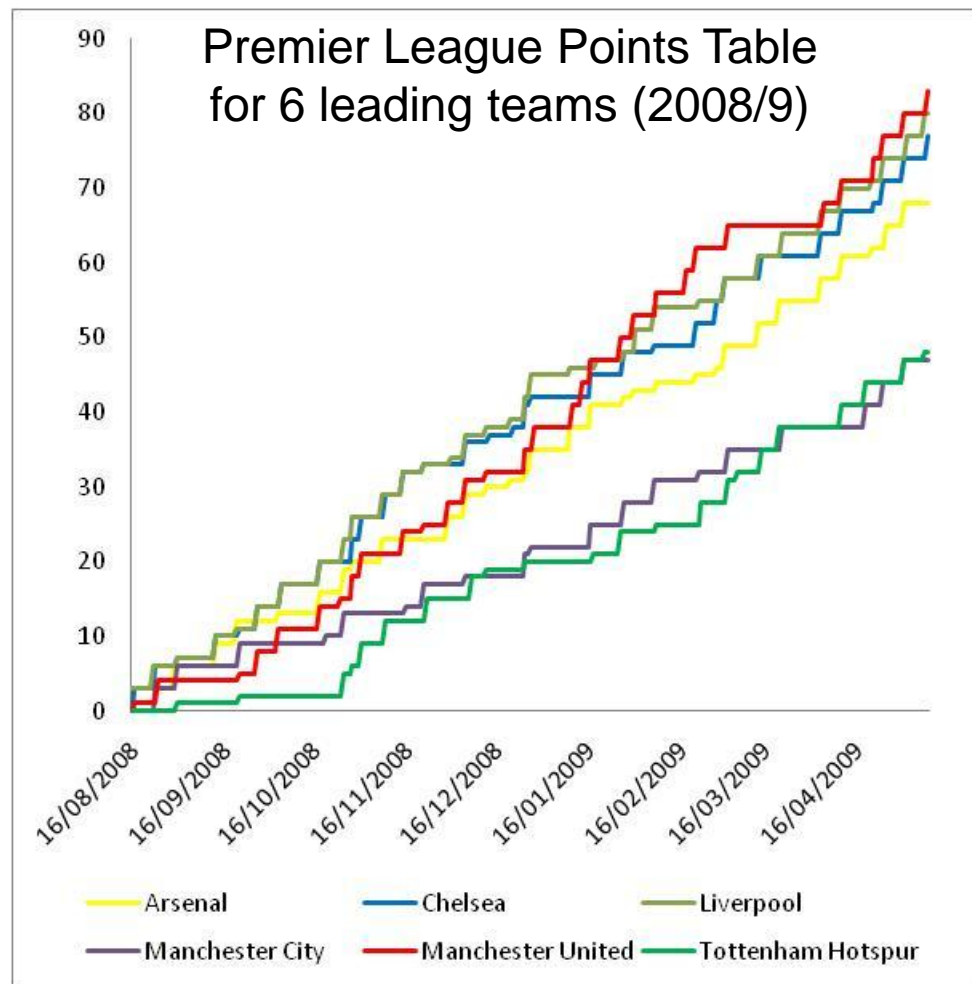
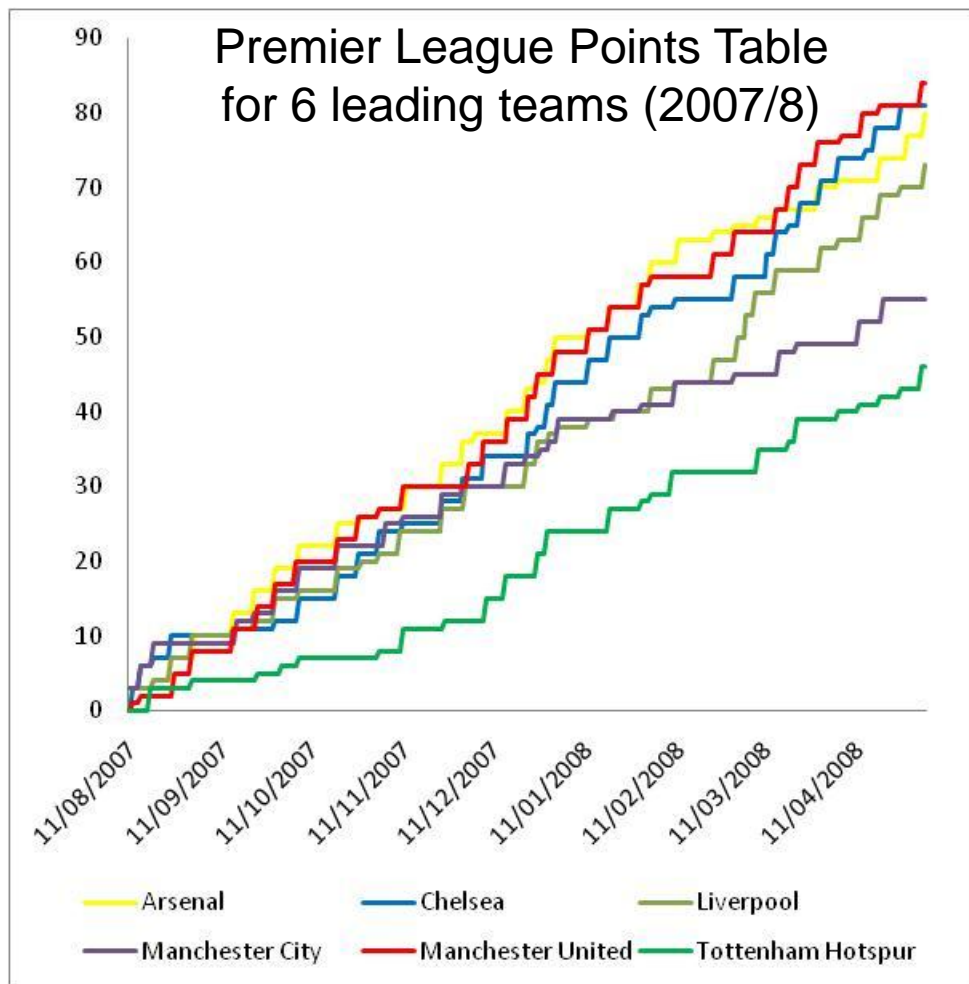
- games in hand
- the opposition to come
- the injury list
- involvement in other competitions and
- playing home or away

which all can influence the outcome

**Manchester United had more points than Chelsea on 24/4/2010, although Chelsea won the league by a point 16 days later!**

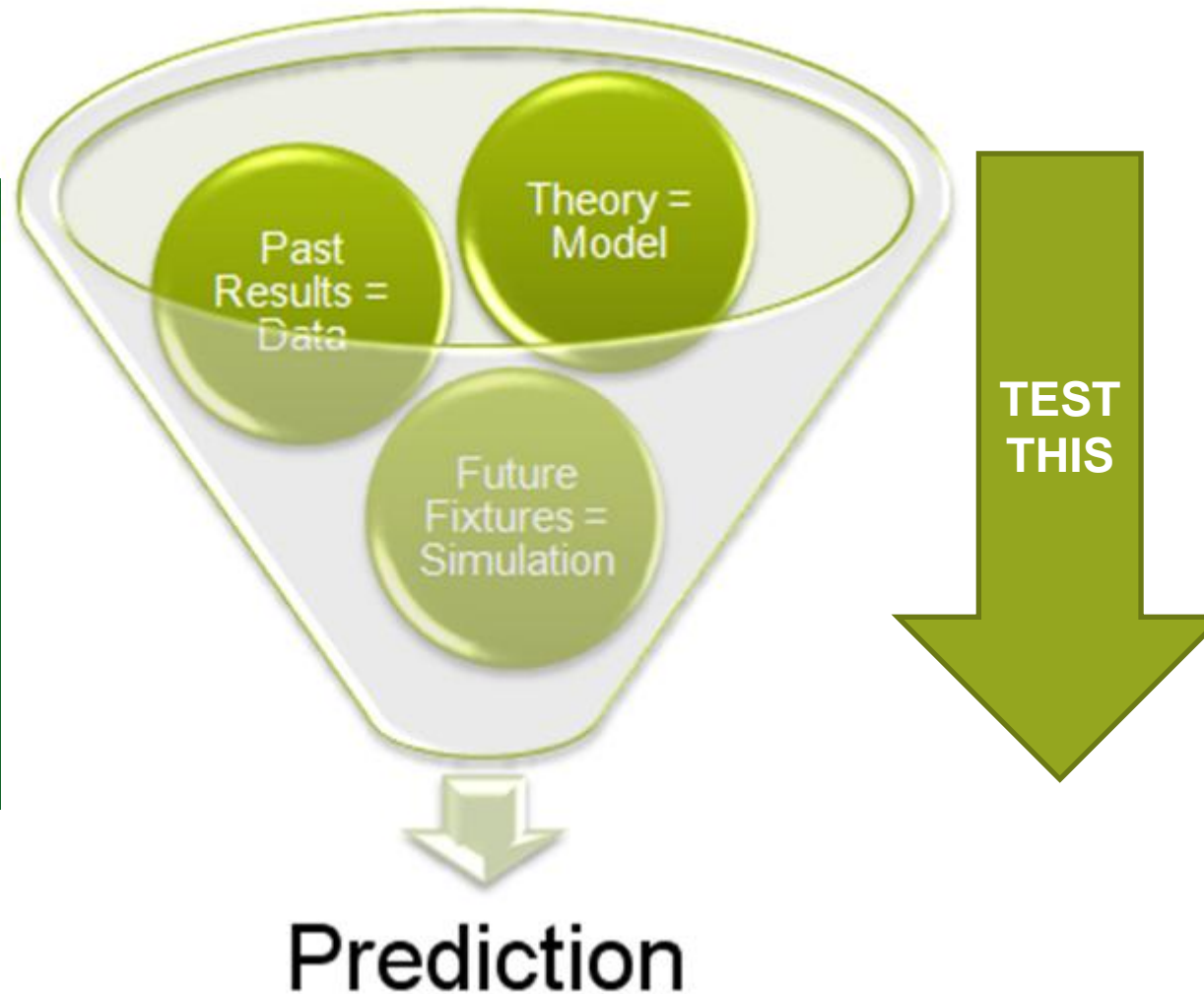
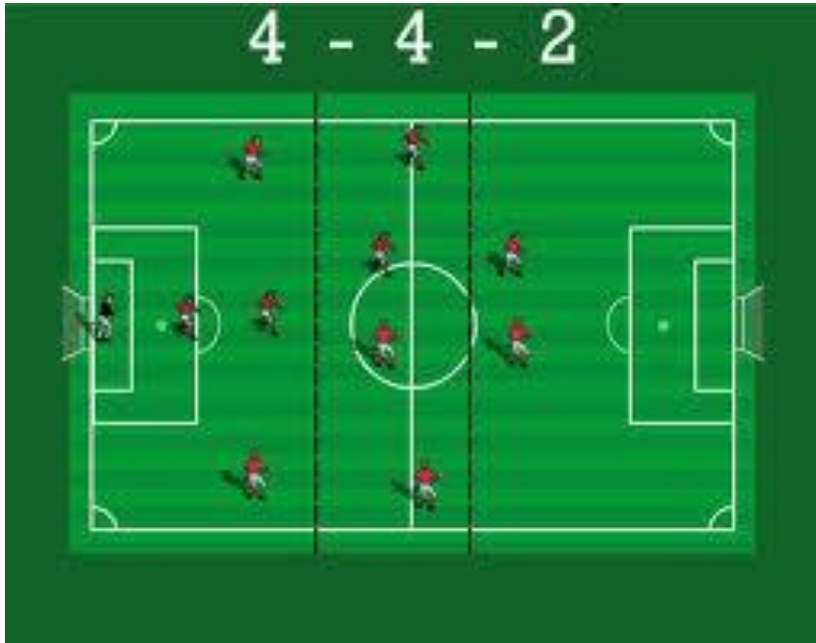


# The points table leader (at this stage of the season) did not remain on top of the league until the end of the season



In 2007/8 Manchester United remained on top of the table from 15/3/2008 onwards  
In 2008/9 Manchester United remained on top of the table from 7/2/2009 onwards

# How does everything fit together?



# What could be in a model to predict the season's league winner? Some ground rules...

## (Simplifying) Assumptions

- Teams don't change over the season
- Results in other competitions do not affect the premier league outcome
- Each match is independent of each other
- Teams perform the same way independent of the competition
- No “bankruptcy” point penalties

## Complications overlooked

- Players get traded, players get injured, or go into and out of form
- Players getting over-played by the end of the season due to success in other competitions
- Teams have streaks of form that affects their confidence
- Teams have derby's and particular rivalries

# Why a Bayesian statistical approach? If we regard 2010/11 results as the only relevant data then...

- At the start of the season we know nothing
  - Every team has the same chance of winning
- As the season progresses we gather more data
  - Goals scored and goals conceded by every team, home and away
- Our prediction needs to be continuously refined to fit the latest data

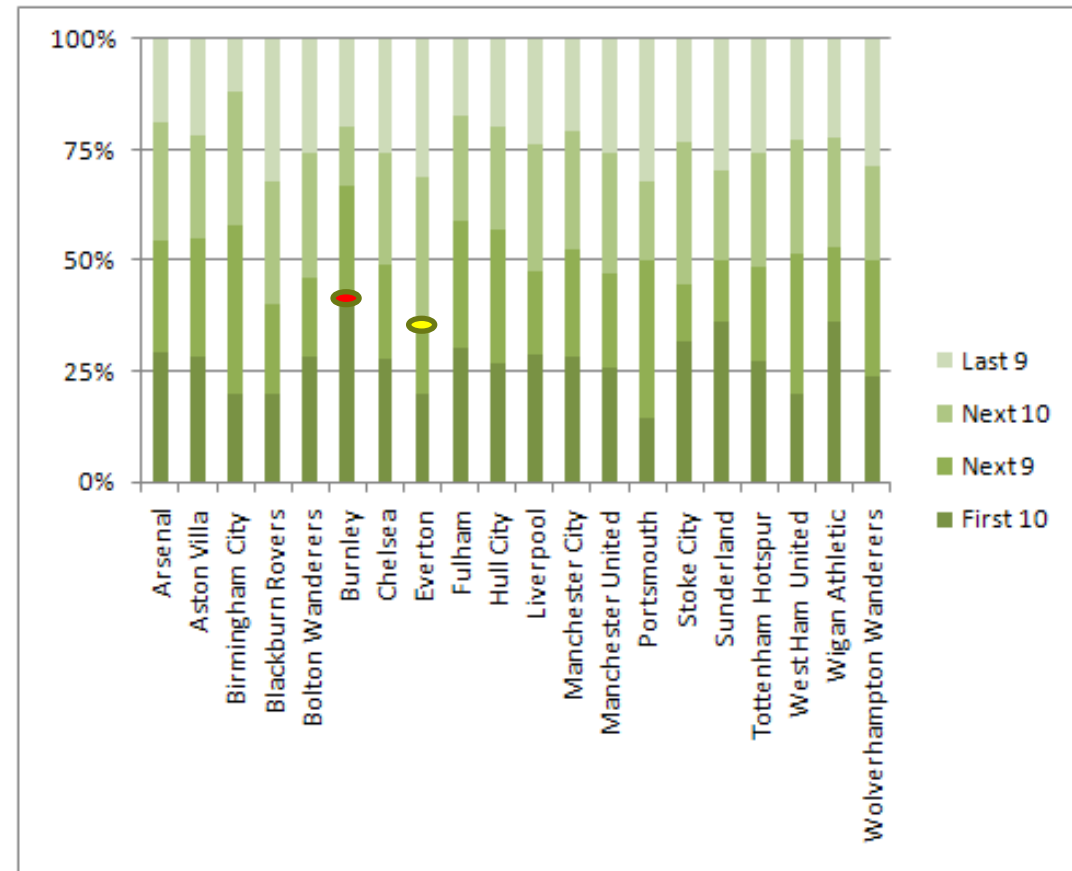
Sat	18/09/10	Stoke City	1 - 1	West Ham United	+	i
		Aston Villa	1 - 1	Bolton Wanderers	+	i
		Blackburn Rovers	1 - 1	Fulham	+	i
		Everton	0 - 1	Newcastle United	+	i
		Tottenham Hotspur	3 - 1	Wolverhampton ...	+	i
		West Bromwich ...	3 - 1	Birmingham City	+	i
		Sunderland	1 - 1	Arsenal	+	i
Sun	19/09/10	Manchester United	3 - 2	Liverpool	+	i
		Wigan Athletic	0 - 2	Manchester City	+	i
		Chelsea	4 - 0	Blackpool	+	i
Sat	25/09/10	Manchester City	1 - 0	Chelsea	+	i
		Arsenal	2 - 3	West Bromwich ...	+	i
		Birmingham City	0 - 0	Wigan Athletic	+	i
		Blackpool	1 - 2	Blackburn Rovers	+	i
		Fulham	0 - 0	Everton		i
		Liverpool	2 - 2	Sunderland	+	i
		West Ham United	1 - 0	Tottenham Hotspur	+	i
Sun	26/09/10	Bolton Wanderers	2 - 2	Manchester United	+	i
		Wolverhampton ...	1 - 2	Aston Villa	+	i
		Newcastle United	1 - 2	Stoke City	+	i
Sat	02/10/10	Wigan Athletic	12 : 45	Wolverhampton ...		i
		Birmingham City	15 : 00	Everton		i
		Stoke City	15 : 00	Blackburn Rovers		i
		Sunderland	15 : 00	Manchester United		i
		Tottenham Hotspur	15 : 00	Aston Villa		i
		West Bromwich ...	15 : 00	Bolton Wanderers		i
		West Ham United	15 : 00	Fulham		i
Sun	03/10/10	Manchester City	13 : 30	Newcastle United		i
		Liverpool	15 : 00	Blackpool		i
		Chelsea	16 : 00	Arsenal		i

$$P(H|E) = \frac{P(E|H) P(H)}{P(E)}$$

# This model relies on teams being consistent. Is this a valid assumption?

As we can see, in the 2009/2010 season, there were some outliers

- Burnley scored 40% of their seasons points in their first 10 matches
- Everton scored 64% of their points in the second half of the season



What about home advantage?

# Home and Away: do we need to take this into account?

- Different for different teams?
- Home ground has bigger or smaller impact/difference?
- Which home fans are the best? Or is it teams not travelling well?

Season	Goals scored at home	Goals scored away	Home goals per match	Away Goals per match	Difference	Share of goals scored away
2007/8	581	421	0.76	0.55	0.21	42%
2008/9	532	410	0.70	0.54	0.16	44%
2009/10	645	408	0.85	0.54	0.31	39%
2010/11*	153	106	1.39	0.96	0.43	41%

We could estimate that just over 40% of goals are scored by the away team

# Why are we using the Monte Carlo method?

**We can't solve the problem analytically!**

- The winner of the premier league will be the result of the remaining 280 matches
- Since each match can have one of three outcomes

**Win home + loss away**

**Draw**

**Loss home + Win away**

modeling the rest of the season deterministically would result in  **$3^{280}$**  different possible outcomes being calculated – which is a number that has **133** digits!

**We know that the best answer should reflect our uncertainty, and the Monte Carlo method reflects this, generating as a result, a distribution of the relative likelihood of the alternate outcomes**

# How can we set the model up to be run using the Monte Carlo method?

- Since the remaining matches all happen independently, we can model each independently
- Since each match has a home team and an away team we can reflect that too
- Since each team has played a series of matches, and has scored and conceded goals, we can model the probability of all possible results, where for instance the result

*Home scores  $H$  and Away scores  $A$*

can be reflected as

$P(\text{Home}=H \text{ and Away}=A) = P(\text{Home score } H)P(\text{Away concede } H)P(\text{Home concede } A)P(\text{Away score } A)$

**All we need now is estimates of these probabilities... DATA**

# Does it matter who plays who in each individual match?

Since our measure of how good the model is, is the likelihood estimate, and the likelihood estimate is of the following form:

$$\prod_{i \text{ matches}} (\text{Home team scoring } x_i)(\text{Away team conceding } x_i)(\text{Home team conceding } y_i)(\text{Away team scoring } y_i)$$

it can be shown that the estimates are independent of who played who, but rather dependent on how many goals were scored or conceded by the home and away team each game:

$$\prod_{i \text{ matches}} (\text{Home team scoring } x_i) \prod_{i \text{ matches}} (\text{Away team conceding } x_i) \prod_{i \text{ matches}} (\text{Home team conceding } y_i) \prod_{i \text{ matches}} (\text{Away team scoring } y_i)$$

**This may be counter intuitive, but reflects our underlying assumption that a team has a constant ‘average scoring rate’ and ‘average conceding rate’ – which is constant across the season irrespective of the opposition**

# What do I mean by data?

## Retrospective

- Past results
  - Goals scored
  - Goals conceded
  - Who played
  - Current league points

## Prospective

- Fixture list
  - Home and away
  - Playing against whom

**Clean the data, validate the data.  
Luckily there are no reporting delays...  
but one is almost immediately out of date...**

# So what is our data?

## 100 Matches completed

- We know where we've been

BBC Mobile News Sport Weather iPlayer TV Radio More Search the BBC

**SPORT FOOTBALL** Watch Sport news bulletin

Sport Homepage  
Football  
Premier League  
Live Scores  
Results  
Fixtures  
Table  
Predictor  
Top Scorers  
A-Z of Sports  
Related BBC sites  
News  
Weather  
Sport Relief

### Barclays Premier League table

Barclays Premier League : Table  
Monday, 1 November 2010 22:05 UK

Team	P	Home					Away					GD	PTS
		W	D	L	F	A	W	D	L	F	A		
1 Chelsea	10	5	0	0	16	0	3	1	1	11	3	24	25
2 Arsenal	10	4	0	1	15	5	2	2	1	7	5	12	20
3 Man Utd	10	4	1	0	13	4	1	4	0	9	8	10	20
4 Man City	10	3	1	1	7	5	2	1	2	6	5	3	17
5 Tottenham	10	2	2	1	6	4	2	1	2	5	6	1	15
6 West Brom	10	3	2	0	8	4	1	1	3	6	13	-3	15
7 Newcastle	10	2	1	2	14	7	2	1	2	5	7	5	14
8 Everton	10	2	2	1	7	5	1	2	2	3	3	2	13
9 Blackpool	10	1	1	2	7	8	3	0	3	8	13	-6	13
10 Fulham	10	2	2	1	7	5	0	4	1	5	6	1	12
11 Bolton	10	1	3	1	6	6	1	3	1	7	8	-1	12
12 Sunderland	10	2	3	0	5	3	0	3	2	4	9	-3	12
13 Liverpool	10	2	2	1	7	6	1	1	3	3	8	-4	12
14 Aston Villa	10	2	3	0	5	1	1	0	4	4	12	-4	12
15 Birmingham	10	2	2	1	4	3	0	3	2	6	9	-2	11
16 Stoke	10	2	1	2	6	6	1	0	4	4	8	-4	10
17 Wigan	10	1	2	3	4	14	1	2	1	3	4	-11	10
18 Blackburn	10	1	2	2	4	5	1	1	3	5	7	-3	9
19 Wolverhampton	10	2	2	1	7	6	0	1	4	3	10	-6	9
20 West Ham	10	1	1	3	5	9	0	2	3	2	9	-11	6

## 280 matches to go

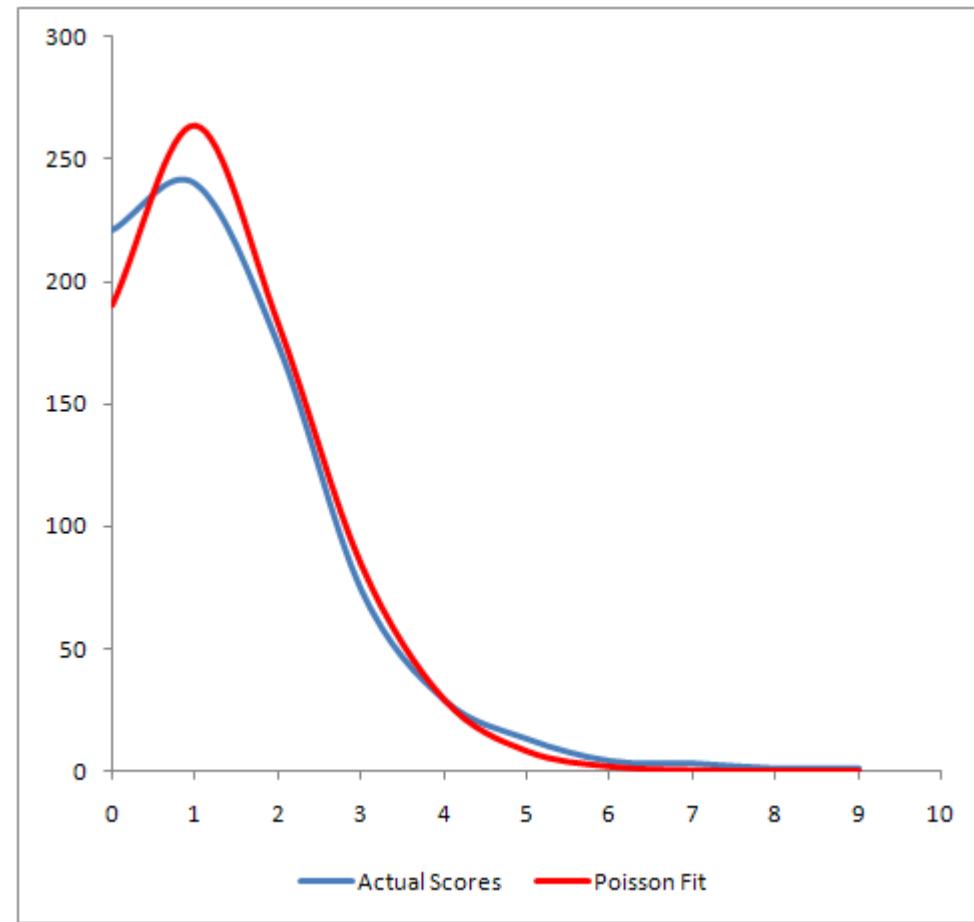
- We know where we're going

Tue	09/11/10	Stoke City	19:45	Birmingham City
		Tottenham Hotspur	20:00	Sunderland
Wed	10/11/10	West Ham United	19:45	West Bromwich ...
		Wigan Athletic	19:45	Liverpool
		Wolverhampton ...	19:45	Arsenal
		Aston Villa	19:45	Blackpool
		Chelsea	19:45	Fulham
		Newcastle United	19:45	Blackburn Rovers
		Everton	20:00	Bolton Wanderers
		Manchester City	20:00	Manchester United
Sat	13/11/10	Aston Villa	12:45	Manchester United
		Manchester City	15:00	Birmingham City
		Newcastle United	15:00	Fulham
		Tottenham Hotspur	15:00	Blackburn Rovers
		West Ham United	15:00	Blackpool
		Wigan Athletic	15:00	West Bromwich ...
		Wolverhampton ...	15:00	Bolton Wanderers
		Stoke City	17:30	Liverpool
Sun	14/11/10	Everton	14:00	Arsenal
		Chelsea	16:10	Sunderland

Usual actuarial checklist here: data capture

# The number of goals scored/conceded by a team can be fit using a Poisson distribution

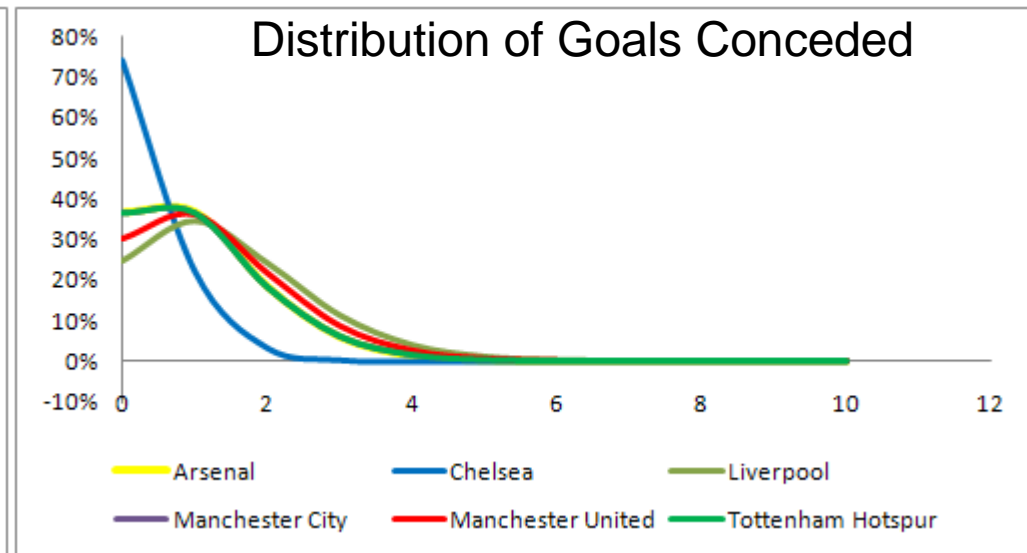
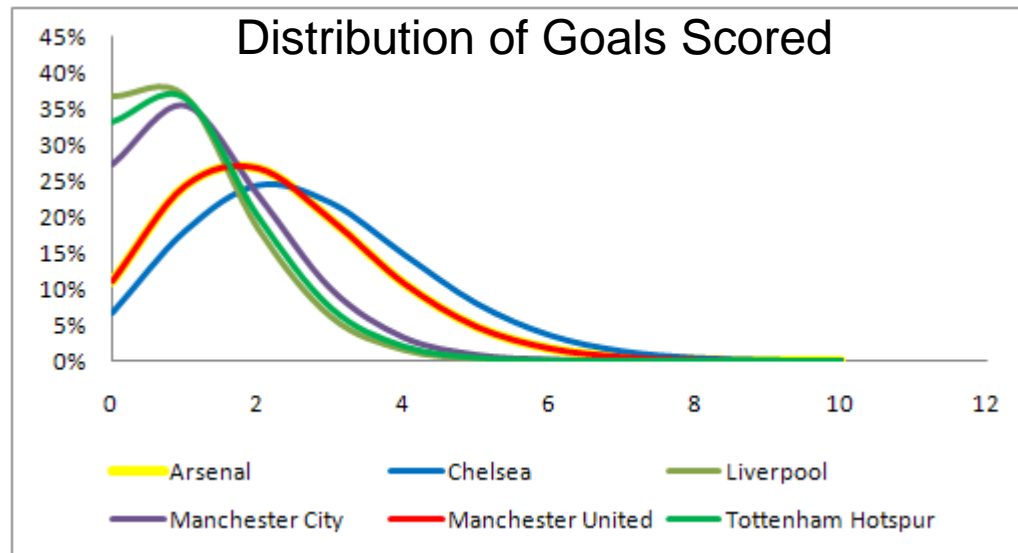
- In the 2009/10 season, there were 20 teams, and with each playing everyone else twice, there were 380 matches
- In each match, two teams 'scored goals' – making 760 data points, illustrated here
- A Poisson fits this distribution very well



# The goals scored and goals conceded results have been used to estimate a Poisson parameter for each team

- We are using a Poisson distribution assuming that the chance of scoring / conceding in a match some time in the future can be estimated using results from earlier in the season

$$f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$



# We are then able to model the outcome of a match using these goal “scoring” and “conceding” estimates

- A match coming up soon between Sunderland and Tottenham Hotspur can be modeled as follows
- Generate the goal scoring and goal conceding probabilities for each team based on it's record (in this case their involvement in the first 100 matches of the season)

Goals in first 10 matches	Team	Probability of scoring or conceding					
		0	1	2	3	4	5
9	Sunderland score	41%	37%	16%	5%	1%	0%
11	Tottenham Hotspur score	33%	37%	20%	7%	2%	0%
12	Sunderland concede	30%	36%	22%	9%	3%	1%
10	Tottenham Hotspur concede	37%	37%	18%	6%	2%	0%

- Use this to develop an estimate of the number of goals scored by Sunderland & conceded by Tottenham
- Use this to develop an estimate of the number of goals conceded by Sunderland & scored by Tottenham
- Randomly simulate the match and calculate the result

# We are then able to model the outcome of a match using these goal “scoring” and “conceding” estimates

- A match coming up soon between Sunderland and Tottenham Hotspur can be modeled as follows
- Generate the goal scoring and goal conceding probabilities for each team based on it's record (in this case their involvement in the first 100 matches of the season)
- Use this to develop an estimate of the number of goals scored by Sunderland & conceded by Tottenham
- Use this to develop an estimate of the number of goals conceded by Sunderland & scored by Tottenham

	0	1	2	3	4	5
Sunderland score and Tottenham Hotspur concede	47%	42%	10%	1%	0%	0%
Tottenham Hotspur score and Sunderland concede	35%	47%	15%	2%	0%	0%

- Randomly simulate the match and calculate the result i.e.
  - 3 points for Sunderland and 0 for Tottenham
  - 1 each, or
  - 3 for Tottenham and 0 for Sunderland

# Once you have a parameter set, what next? Runs and runs and runs and runs and runs and runs...

- Now that we have agreed on the parameter set, we can 'randomly simulate the results of the rest of the season'
- We have done this 10,000 times

*As an aside, the actual likelihood for our best set of parameters (which is the best estimate for the model we have developed, or the maximum likelihood estimate of the parameters) can be used to show that the score lines so far this season have about a 1 in  $10^{-225}$  probability of having occurred*

This is a small number, but since there are an infinite number of possibilities...



# Testing the model – sometimes the model answer isn't what you expect!



**Just because you have a good looking model, doesn't mean you have the answer!**

# Predictions are difficult things to make... it's easy to make a blooper...



Man City are title rivals - Ancelotti



## Jose tipping City as a big title threat

By Kevin Aitken

JOSE MOURINHO believes Manchester City are one of only three teams which can win the Premier League this season and admits the big-spending Eastlands outfit are 'dominant' in the transfer market.

The Real Madrid boss had wanted to sign Aleksander Kolarov this summer but was outbid by City, who paid £16million for the full-back, and also splashed out over £100m on David Silva, James Milner, Mario Balotelli, Yaya Toure and Jerome Boateng.

'I think it will again be Man United, Chelsea and of course Man City [to win the title], because they have a great squad,' said Mourinho, who also denied he is preparing a late move for City striker Emmanuel Adebayor.

'It is very difficult for Roy [Hodgson] to make Liverpool champions. I think he needs time and it's not easy because I don't think the club went in the right direction and don't think Arsenal [can win it].'

And of City, who impressed in Monday's 3-0 win over Liverpool, he added: 'They are dominant in the market. The player they say 'this is the player I want' is the player they get. I was very

### Micah still hoping for call from Fabio

MICAH RICHARDS is hoping his improved form can catapult him back into the England fold – although the Manchester City defender will miss the European Championship qualifying double-header next month as he has been named in the Under-21 squad instead. 'Hopefully there is a spot there for me,' Richards said. Newcastle striker Andy Carroll has also been put on Under-21 duty for the matches against Portugal and Lithuania.

interested in Kolarov when I came here but I couldn't compete with them – they went to values that you cannot go.'

Mourinho also believes his former player Balotelli can be a City success following his £26m arrival from Inter.

'I had some problem because he is a kid and because a coach always wants to educate a kid and always wants a kid to go in the right direction,' Mourinho told Sky Sports News HD.

'Mario has incredible potential. He has every football quality to adapt.'

BBC

Home

News

Sport

Weather

iPlayer

TV

Radio

More...

Search

## SPORT FOOTBALL

Watch Sport news bulletin

Sport Homepage

Football

Premier League

Live Scores

Results

Fixtures

Table

Predictor

Top Scorers

A-Z of Sports

Related BBC sites

News

Weather

Sport Relief

Page last updated at 15:57 GMT, Saturday, 25 September 2010 16:57 UK

E-mail this to a friend

Printable version

### Chelsea will still win Premier League - Roberto Mancini



Manchester City manager Roberto Mancini still insists that Premier League pacesetters Chelsea will retain their title despite losing 1-0 to his team.

#### MANCHESTER CITY

- ▶ Your say - 606
- ▶ Weather
- ▶ BBC Manchester sport
- ▶ Official club website

#### CHELSEA

- ▶ Your say - 606
- ▶ Weather
- ▶ BBC London sport
- ▶ Official club website

#### SEE ALSO

- ▶ Manchester City 1-0 Chelsea  
25 Sep 10 | Premier League

# Half Time Entertainment: Clip of some great goals from the premier league season so far

---



# Lessons Learnt from the 2010 World Cup Prediction Model: retrospective data, prospective gamblers

Model Predictive Ability

## World Cup fever

Greg Becker and Amirade Kaith who is likely to win the 2010 FIFA World Cup in South Africa? Can you get an answer from a model?



## Model

Baseline (Actual Results)

UBS

JP Morgan

Goldman Sachs

Simon Kuper and Stefan Szymanski (Wired Mag)

## Predictive Ability

100%

66%

59%

53%

41%



Betfair was better than Frontier Economic, Goldman or JP Morgan!

# There is only downside potential... unless you cover your bets...

10 | **The Capitalist** EDITED BY VICTORIA BATES GOT A STORY? EMAIL: thecapitalist@cityam.com

## CITY LOSES TO OCTOPUS IN FOOTY FORECASTS

OHL, HOW they chortled in the City yesterday at comparisons between the world's most famous cephalopod and those investment banks who dared to put out predictions on the outcome of the World Cup.

Paul the psychic octopus, they sniggered, had managed to get EVERY SINGLE ONE of his predictions correct, beating statistical odds of 1/256.

Kaggle, the data prediction platform, jumped in on the action, yelling that in its own World Cup predicting competition, JP Morgan, Goldman Sachs, UBS and Deutsche Bank had all fared between 28th and 64th out of 65. (JP Morgan picked England to win it - don't laugh, it did seem at least remotely plausible at the time - while the other three plumped for Brazil.)

All of the banks kept a dignified silence on their thrashing yesterday except for Evolution, whose fixed income specialist Gary Jenkins had also released a tongue-in-cheek forecast for the tournament (Brazil, again) and was happy to give his tuppence-worth on the results.

"Clearly against Paul, we've all done appallingly," Jenkins roared enthusiastically. "But he has got a huge advantage over me - one, he doesn't have to work and can sit in his tank all day watching footy, and two, he's got more brains. Everything's working in his favour..."

Good to see at least one of the red-faced analysts taking the defeat on the chin.

"At least one of the City's red-faced analysts took their World Cup forecast defeat squarely on the chin"



LUDLOW & TENBURY WELLS  
**Advertiser**

News | Sport | Leisure | Info | Your Say | Announcements | Events | Family | Jobs | Home

Ludlow | Features | **Tenbury Wells** | Bishop's Castle | Church Stretton | Cleobury Mortimer | Craven Arms | Farm

Ludlow Advertiser > News > Tenbury Wells >

### TENBURY WELLS

No suspicious circumstances in death of Paul the Octopus

11:45am Wednesday 27th October 2010

Print | Email | Share | Comments(0)

ONE of the more fishy international stars on the books of a Tenbury theatrical agent has died.

Chris Davis has confirmed that Paul the Octopus who made global headlines because of his talent for predicting the results of



Paul the Octopus.

Paul has shown that some things are still certain though...

# Analysis by David Forrest and Robert Simmons into tipsters, betting odds and statistical models shows...

- Statistical models do better than tipsters
- Combining tipster estimates can be better than using an individual tipster
- Betting odds are the best – maybe because the bookies pay the biggest salaries?

But they conclude that the betting exchanges involve people working with more complex models and more data – so don't bet against them!

feature

Read all about it?

If statistics are of little help in predicting individual football results, are the professional followers of the sport any better? **David Forrest and Robert Simmons** analyse the performance of newspaper tipsters in helping the gambler to beat the bookies.

**Can sports tipsters help? Evidence for English football**

We live in a world where much complex information relevant to forecasting the future is available from many sources, one of which is the print media. Newspapers give specialist advice on buying horses, cars and furniture, and on finance. They also give much information about football. Here a prime motive for readers wanting probabilistic forecasts of future events is to give them an edge in the betting on match results. We look at the effectiveness of tipster advice published in daily newspapers and compare it with the performance of forecasts derived from formal statistical modelling, and from bookmakers' odds themselves. We show that newspaper advice is rather ineffective and that statistical modelling (of the kind outlined in the previous article) is a superior source of guidance. But it is a bookmaker betting odds that are the most efficient predictors of match results. This is perhaps unsurprising: newspaper tipster advice, for example, is free, whereas bookmakers would lose substantial sums of money if their odds were poor predictors of match results.

**Tipsters**

Until recently many daily newspapers offered a column in their sports sections, usually on Monday or Tuesday, where a tipster would list a set of predictions of results for fixtures on the upcoming weekend. In the *Daily Mirror* the tipster would offer a set of accompanying statistical tables as form guides, rather like the guides published to aid horse race bettors. He might also offer some analytical commentary. The service was intended to offer free assistance to punters who took part in the football pools. Presumably editors felt that without

this service fewer readers would buy their particular newspapers.

This service has been replaced more recently by more specific advice – particularly in those papers which we will refer to as broadsheets – where expert tipsters nominate 'best bets' over a variety of sports events, which might be as diverse as golf championship or American football games.

**Evaluation of tipster performance**

We thought it would be useful to establish statistically whether free football tipster services were effective. Although the scope of this inquiry was narrow, the methodology could easily be applied to other settings where specialist advice is sought and is available at low cost.

In a detailed analysis published in the *International Journal of Forecasting* (2006)<sup>1</sup> we looked at a set of 1694 English professional league football matches played between December 1996 and April 1998. We inspected the match predictions of tipsters of three English newspapers: the *Daily Mail*, the *Daily Mirror* and *The Times*. These were selected to give a representative cross-section of British newspaper types and appeal over the population. The proportions of matches called successfully by the tipsters were:

<i>Daily Mail</i>	42.6%
<i>Daily Mirror</i>	41.1%
<i>The Times</i>	42.9%

The frequencies of actual match outcomes were:

Home wins	47.5%
Draws	28.1%
Away wins	24.4%

It immediately follows that always predicting a home win would have achieved a higher success

rate than following any of the tipsters. This is a consequence of the importance of 'home advantage' that is universal across team sports. But the implied criticism here is unfair to the tipsters. Recall that their columns were designed to help bettors predict draws in the football pools game. The tipsters felt obliged to call at least some matches as draws in their predictions. Also, a game within the football pools enabled punters to identify a number of away wins so these also had to feature in tipster selections.

**Ordered logit modelling**

A more rigorous assessment of tipster performance is ordered logit regression analysis. In this method we think of teams as ordered (ranked) by relative strengths. We do not observe 'team strength' as this includes such intangible factors as leadership and 'team spirit'. These relative strengths may run much outcomes through the logistic distribution. The dependent variable is then constructed as categorical, for example a home win = 0, a draw = 1 and an away win = 2. Tipster forecasts can be treated as dummy variables, for example corresponding to draw and away predictions with home win prediction as an omitted base category.

We can use the ordered logit regression model to answer three questions:

1. Do tipsters' selections significantly improve upon random selections?
2. Do tipsters' selections offer something extra (unspecified) to easily available public information?
3. Does a consensus forecast taken from the three tipsters outperform any single tipster?

To answer the first question we regress match outcomes on tipster dummy variables using the

formation variables were significant predictors of match results. Using the likelihood ratio test described above, our restricted model had public information variables only whereas the unrestricted model added tipster predictions. We found that there was only one sport, the *Daily Mail*, for whom the addition of tipster predictions added significant explanatory power to the model. The other two tipsters failed this particular test. Actually in a further test, we found that tipsters did not even appear to make full use of the public information available as summarised in our measures.

forecasting in predicting future events. Perhaps experts typically give incorrect weighting to non-statistical information.

But in the article by Boulier and Stokler it is the betting market that is found to yield the most accurate set of forecasts for US football. The terms of bets offered already display enough frisson that it is hard to beat them by consulting experts. This is unsurprising. Bookmakers lose considerable sums of money if they set 'incorrect' odds or spreads and so they may be expected themselves to recruit the best experts.

**References**

1. Forrest, D. and Simmons, R. (2006) Forecasting sport: the behaviour and performance of football tipsters. *International Journal of Forecasting*, 16, 317–331.
2. Boulier, B. and Stokler, H. (2005) Predicting the outcomes of National Football League games. *International Journal of Forecasting*, 19, 297–329.
3. Forrest, D., Simmons, R. and Goldsack, J. (2007) Odds versus forecasting: the case of football tipsters. *International Journal of Forecasting*, 21, 951–964.

Dr David Forrest is Reader in Economics at the University of Suffolk, specialising in sport economics and the economics of gambling, and with Dr Robert Simmons has co-authored numerous papers on betting in sport. Dr Simmons, Senior Lecturer at Lancaster University, has an international reputation as a sports economist. He is also a football referee.

20 [significance](#)

© 2010 The Actuarial Profession • [www.actuaries.org.uk](http://www.actuaries.org.uk)

Forrest D, Simmons R 2006 "Read all about it" RSS  
Significance p151-153 Volume 3 Issue 1 March 2006

[contents](#) [significance](#) 21

# John Goddard has looked at ‘streaks’, ‘firing a manager’, ‘fighting for survival’, the effect of ‘playing in Europe’...

- Winning teams keep on winning
  - Confidence?
- Firing a manager doesn’t seem to help a team if one controls for ‘mean reversion’
- Home advantage has an effect, and while it has been decreasing over the most recent 35 seasons, the home ground advantage is larger when the away team has to travel further
- Relegation threatened teams “fighting for survival” are more likely to beat their mid-table rivals at the end of the season than before
- The “playing in Europe” effect has not been shown to be a hindrance, although it has been for some top teams

Goddard J 2006 “Who wins the football” RSS  
Significance p16-19 Volume 3 Issue 1 March 2006

feature

## Who wins the football?

Portsmouth have dumped their manager—will it change their results? Chelsea are on a winning streak—does it mean anything? Media pundits and sports fans—and punters—devote time and energy to speculation about who will win Saturday’s beautiful game. The discussions—in pubs, newspapers and dressing rooms—are usually based on subjective opinion, or evidence that is at best partial or impressionistic. Can statisticians do better? John Goddard offers a large-scale analysis of patterns in English league results from the past 35 football seasons—new managers, winning and losing streaks, home advantages and all.

The relationship between the odds available for betting on the outcome of sporting contests and the true probabilities associated with those outcomes is an important one in the statistical and economic modelling of professional team sports. In this article I shall discuss various methods and findings from a statistical analysis of a football dataset covering 7514 matches played in the four divisional tiers of the English Premier League and Football League, over the 35 most recent seasons from 1990–1991 to 2004–2005.

Specifically I examine evidence for the existence of ‘streaks’ (mean sequences of good or bad match results) caused by temporary variations in team confidence or form. I examine the short-term impact of managerial change on match results. I look briefly at several other influences on match results, including home advantage and whether or not the leading teams are participating in European competition. Finally, what practical issues arise when using statistical analysis to try to formulate a profitable betting strategy?

Streaks

Streaks might influence a team’s match result probabilities in either direction:

- sequences of good results might increase confidence, increasing the probability of another win, whereas a sequence of poor results might sap confidence, reducing the probability of a win;
- alternatively a sequence of good results might create complacency, reducing the probability of another win, whereas a sequence of poor results might inspire a team to raise its efforts—or force the manager to change personnel or tactics—increasing the probability of a win.

Observed match result probabilities, conditional on the direction of certain sequences of match results, can be calculated using the 35-season match results dataset. For example, the unconditional probability of a home win is 0.485 (the overall proportion of home wins in this dataset).

Figure 1. Mean reversion in home performance: teams that kept their manager, teams that changed their manager

Figure 2. Mean reversion in away performance: teams that kept their manager, teams that changed their manager

Table 1. European involvement and league performance 1990–2005

Match results	No. of league matches played in seasons when involved in Europe	Average league points per match	No. of league matches played following a European match	Average league points per match
All teams	3642	1.054	649	1.075
Arsenal	408	1.921	100	1.800
Chelsea	345	1.882	72	1.480
Liverpool	426	1.071	63	1.066
Manchester United	585	2.044	133	1.962

Table 4 shows the expected proceeds—and the size of the 95% confidence interval for the proceeds—as  $N$  increases. Suppose  $\mu = 0.05$  and bets are placed at odds of 10 to 1. With  $N = 100$  bets, the expected proceeds are £1, but the 95% confidence interval extends from –£4.5 to +£6.5. As many as  $N = 1553$  bets are required in order to make the lower limit of this confidence interval zero, or to yield a 0.95 probability of a profit.

John Goddard is Professor of Financial Economics at the University of Wales, Bangor. His research interests include the economics of professional football, statistics in sport, statistical econometrics and the economics of the financial sector.

**David Spiegelhalter and Yin-Lam Ng have shown that statistical models can outperform sports commentators**

- They modelled a single round of matches
- Their model had similarities
  - They also used a Poisson approach to estimate ‘number of goals scored/conceded’
  - They also looked at a result as being a combination of scoring/conceding
  - They developed a way for teams to interact

Spiegelhalter D, Ng Y-L 2009 "One match to go" - RSS  
Significance p151-153 Volume 6 Issue 4 December 2009

## Finally, we get to the model predictions!

# feature

# One match to go!

Can statistics really predict the results of Premier League football matches? David Spiegelhalter and Yin-Lam Ng put their skills to the test, and their reputations on the line—and scored!

On May 23rd, 2009, the 20 teams in the English Premier League each had one match left to play. West Bromwich Albion (West Brians) were at the bottom of the league with 31 points and Manchester United (Man U) were at the top with 87. The bottom three teams would be relegated. West Brians were certain to be one of them, but there were four other teams trying to avoid the other two places. Man U 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. I-hull, though, were one of the teams up for relegation and so had everything to play for.

BBC Radio 4's *More or Less* is more or less the only series on British national broadcasting that deals with numbers in a serious way. Its producers had heard of the work that we had been doing on modelling European football results, and they asked us to produce predictions for those final ten matches using a statistical method that could be explained, so non-sporting listeners, on the radio. This was quite a tricky challenge. Prediction itself is easy. Predicting accurately is the hard bit. In this case we knew that our predictions would be announced on the radio before the matches and then afterwards compared with what really happened and how well either pundit did. Our equations would be on the line; and, to an extent, would be the reputation of statistics.

Complex statistical models are used extensively in the sports betting industry, but we wanted a fairly straightforward model that could be explained using familiar concepts derived from the raw data. We found that using some basic theory we could give early precise a reasonable probability for all the possible results of a game, although then we used a slightly more sophisticated analysis for our actual probabilistic predictions.

We can start by looking at the state of the league on May 22nd, 2009, with goals for and goals against (see Table 1).

The average number of goals scored by the teams in the season so far, and therefore also the average number of goals conceded, was 46. If we divide the number of goals scored by 46, we get a measure of "attack strength"; Arsenal 1.39 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", so Stoke

City's 1.31 shows they have 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, up till now and the average number scored by an away team. Home teams had scored, on average, 1.40 goals per match; for away teams the figure was 1.08. Home teams clearly do better.

Now suppose we want to predict the result of Hull versus Man U. We start by estimating how many goals Hull will score. They are playing at home, so if they were an average team in an average match we expect them to score 1.08. But this is not an average team: over the season they have scored only 89% of the average number of goals, and so their "attack strength" is 0.85. Multiplying up we get  $1.08 \times 0.85 = 0.92$  as the number of goals we might expect them to score against an average team. But their opposition is not average either: Man U's defence weakness is 0.52, since they have conceded only 52% of the average. So we get a total of  $1.08 \times 0.85 \times 0.52 = 0.62$  expected goals by Hull, which does not look too good for their supporters.

For Man U, their headline is 1.08, the average number of goals scored by an away team. But by

The crucial match: Hull City vs Manchester United. Photo courtesy Hull City AFC 2009. Not to be reproduced without permission.

modelled | [significance](#) | 151

all major leagues  
5 years, and the  
the best model  
number that al-  
tendency to  
which we might

as very simple  
performance  
in future results  
to mean factors.  
ing us avoid rel-  
their strength  
e and so it could  
a much better  
R, we had gone  
thought is, as  
were more like 2  
I-hull winning,  
most likely re-  
the statistical  
shown to hold

bility is 20%,  
ity around 40%  
more likely "spe-  
r making com-  
a prediction—  
place (what we  
tion. If we add  
some combina-  
line we get the  
Some of these  
is the 72% prob-  
Arsenal-Serie  
B is considered  
album. Given  
I with the three  
likely likely.  
we read out on  
day 22nd, 2009,  
variation, were  
as without any  
were also given  
/www.bbc.c  
a /www.cip-  
in time with  
takenly said the  
0-0 prediction  
we should have  
ambush's succe-  
probability 10%

ally would  
for penalty  
prediction  
since these  
drawn and  
this would  
our model  
ity by 40%  
us. We can  
over penalty  
we expected  
the "true"  
statistics for  
R, signifi-  
less than  
ed it to be,  
were lucky,  
e our being  
probabilities

are refined  
technique and  
they experi-  
of the latest  
it is a radica-  
change con-  
al evidence  
state. And,  
one exactly  
on—sim-  
ilarly to my  
ty bookmen,  
use them to  
sense people  
ent, and we  
do waiting  
but we are  
e—it could

ing accu-  
fication, 36,  
(1997) Mod-  
performance  
Statistics, 46,

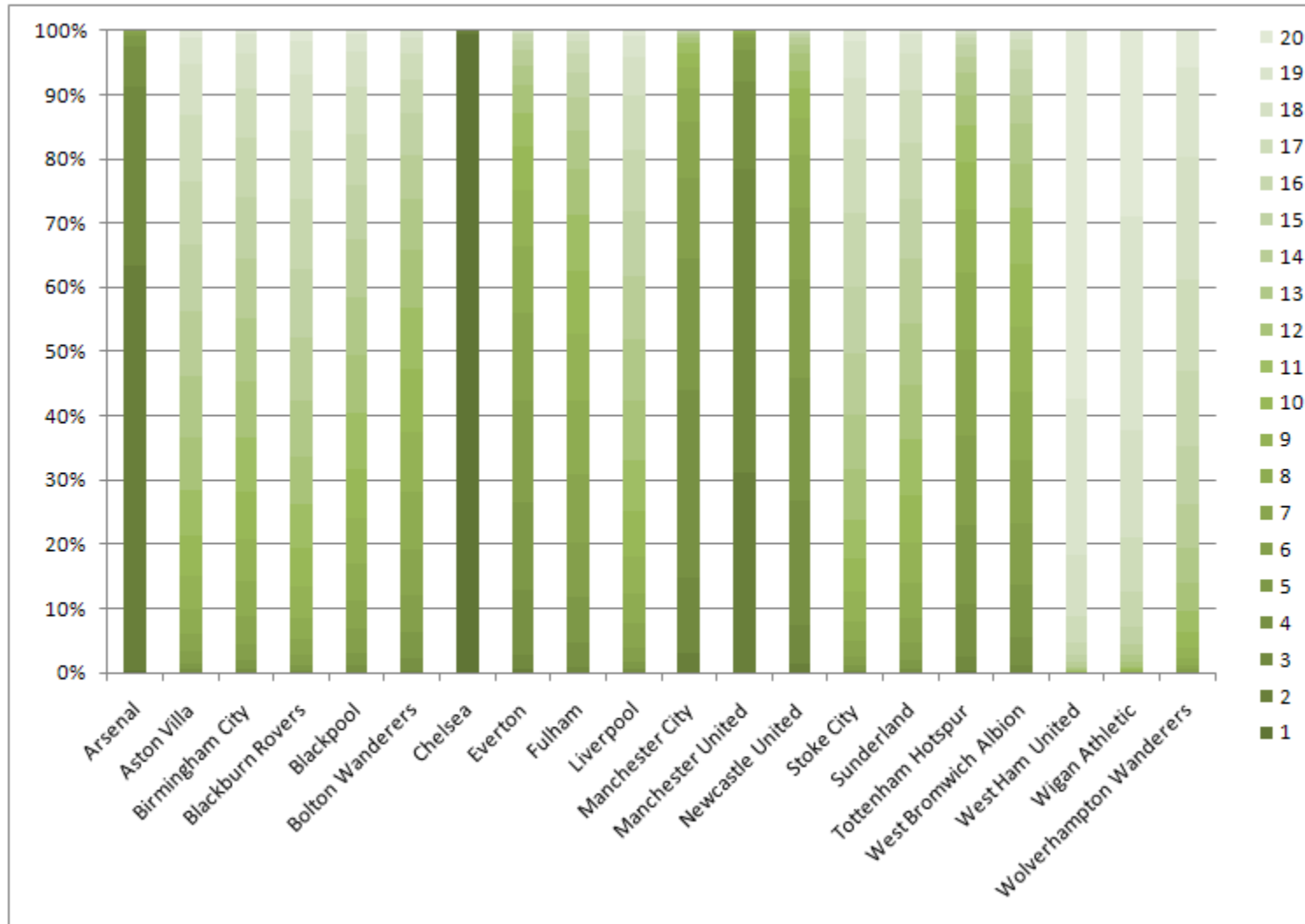
003) Analy-  
tical models.  
re day of the  
but when the  
me very reduced  
predictions", we  
ne of win, draw  
This was par-

152 | [significance](#) | November 2009

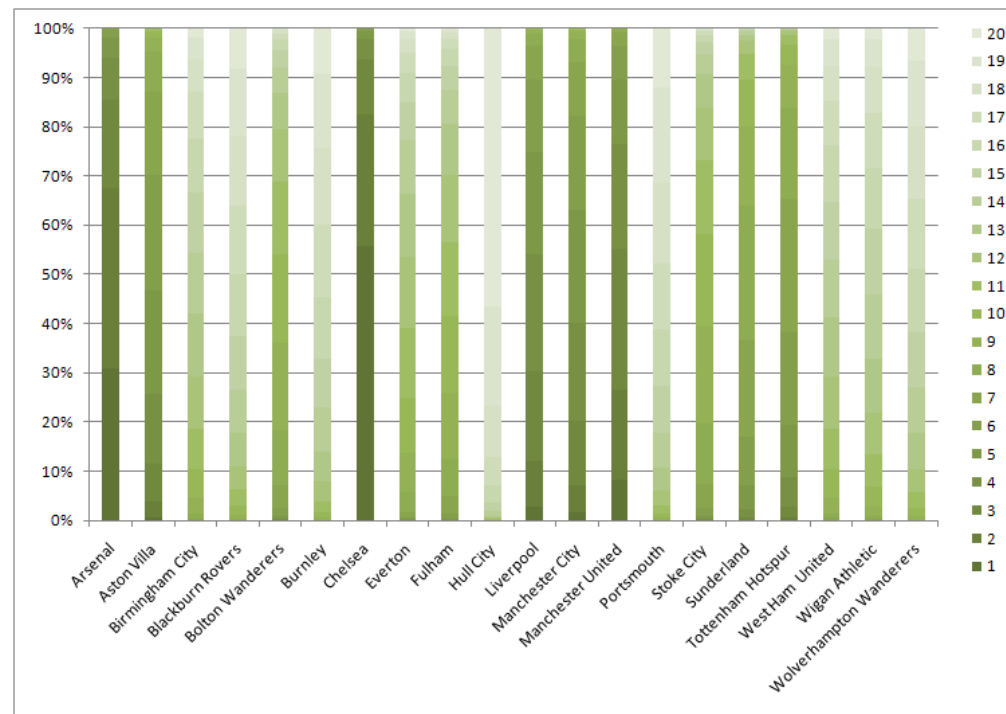
November 2009 | [significance](#)

30

# The slide you have been waiting for: The darker, the more likely the team will win (2010-11)

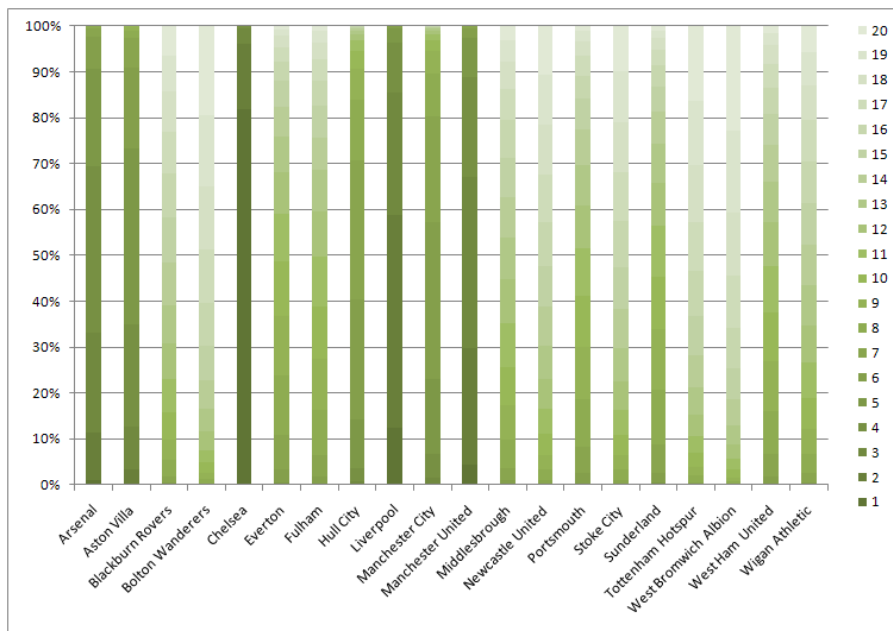


# The 2009-10 predictions after 100 games predicted that Chelsea was the most likely team to win



The model predicts that Chelsea is the most likely team to win – and they did!

# The predictions at this stage of the 2008-2009 season had Chelsea as likely winners



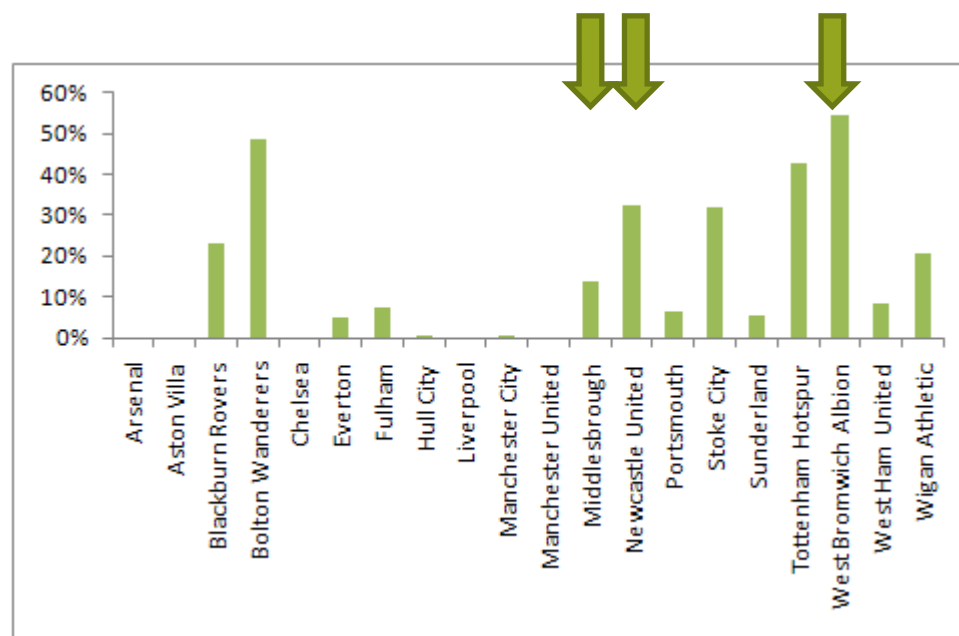
Both Chelsea and Liverpool were in a better position. Manchester United was forecast to win with a 4% probability, and finish in the top 3 more than 2 out of 3 times

# This season, the premier league winner is getting almost as much press as the relegation zone

How good is the model at predicting the teams to be relegated at this stage of the season?

**2008/2009**

- Who was relegated? (Estimated probability of going down at this stage of the season in brackets)
  - Middlesbrough (14%)
  - Newcastle United (32%)
  - West Bromwich Albion (54%)
- Who was predicted to be relegated after 100 games?
  - The model predicted that many teams had a good chance of going down

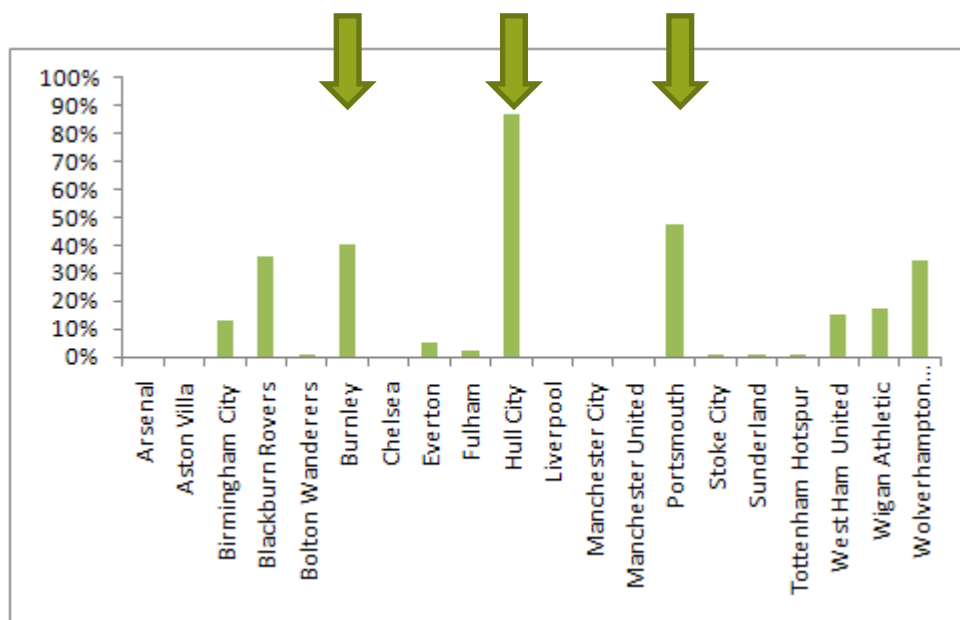


- The three that did, were all in the 8 most likely to go down

# This season, the premier league winner is getting almost as much press as the relegation zone

How good is the model at predicting the teams to be relegated at this stage of the season?

**2009/2010**



- Who was relegated? (Estimated probability of going down at this stage of the season in brackets)
  - Burnley (40%)
  - Hull (87%)
  - Portsmouth (48%)
- Who was predicted to be relegated after 100 games?
  - The three forecast to have the greatest probability of going down, went down!

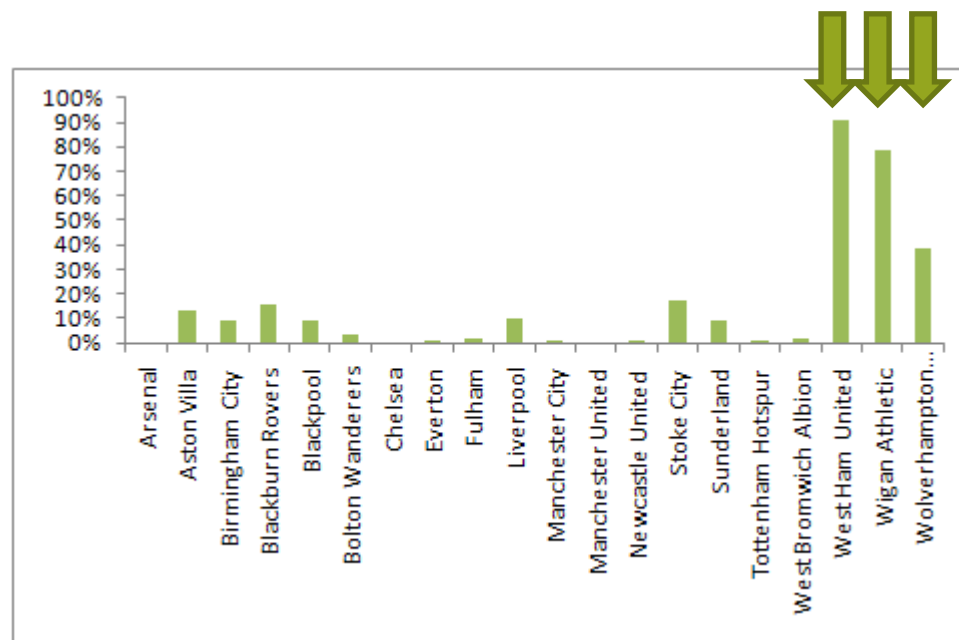
**So the model worked very well for the last season, but not so well for the one before.  
How about 2010/11?**

# Who is forecast to get relegated? Something beginning with...

# W

## 2010-11

- Who is being predicted to be relegated?
  - West Ham United (91%)
  - Wigan Athletic (79%)
  - Wolverhampton Wanderers (39%)



# Interestingly, the betting stats agree – the three W's are in trouble!

Key	statto odds	Ladbrokes	William HILL	sportingbet	BETFRED	Paddy Power	BLUESQ	bet365	StanJames.com	Boylesports	888sport	CENTREBET	betdaq	GoalW'n	betfair	BETDAQ	WBX	SPORTING INDEX
2.14 odds																		
2.14 best odds																		
2.12 odds shortening																		
2.16 odds lengthening (click on odds to bet)																		
Blackpool	2.68	1.5	1.67	1.57	1.57	1.67	1.57	1.73	1.73	1.62	1.57		1.57		1.78	1.78	1.77	
West Ham United	1.69	1.91	1.91	1.85	1.91	1.83	1.91	1.91	1.91	1.91	1.91		1.91		1.95	1.93	1.92	
Wolverhampton Wndrs	2.44	1.91	1.91	1.85	1.91	1.83	1.8	1.91	1.91	1.83	1.8		1.8		2.02	2	1.99	
Wigan Athletic	1.95	1.8	1.91	2	2	1.83	1.91	2	1.91	1.91	1.91		1.91		2.02	2.02	1.99	
West Bromwich Albion	15.5	6	6.5	6	6.5	6	6	6	6	5.5	6		6		8.2	8	7.8	
Birmingham City	4.9	7	5.5	6.5	6.5	5.5	6	5	5.5	6	6		6		6.8	7	6.4	
Stoke City	5.5	7	7.5	7.5	7.5	7.5	8	7	7	8	8		7.5		8.6	8.4	8.2	
Newcastle United	19.5	7	7	7	7.5	8	7	7	6	7.5	7		7		9	4.8	8.8	
Blackburn Rovers	7	7	8.5	7.5	8	8	7	8	7	7.5	7		8		9	9.2	8.6	
Bolton Wanderers	6	10	9	7.5	8	8	8	9	7	8	8		8		9.6	9.4	9.2	
Fulham	7.6	10	11	9	11	8.5	10	8.5	9	9	10		9		12.5	12.5	11.5	
Sunderland	9.2	10	12	11	11	11	12	10	9	11	12		11		12	10	10.5	
Aston Villa	30	26	29	29	26	26	26	26	21	29	26		26		30	31	28	
Liverpool	85	34	26	34	34	31	34	34	34	34	34		34		34	32		
Everton	210	34	67	51	67	67	51	67	51	67	51		67		80	74	70	
Tottenham Hotspur	260	151				251				251					400	395		
Manchester City	500					751				301					300	295	260	
Arsenal	-					751				751					440	435		
Manchester United	-					2001				2001					300	295	270	
Chelsea	-					2001				2001					720	715		

# Betting odds have Chelsea as the 2010/11 favourite – with the odd's reflecting a 63-69% chance of winning the league

## » English Premier League : Winner

ALL	BEST	EXCHANGE	EXCHANGE+	--- Find It Quick! ---													
Key 2.14 odds 2.14 best odds 2.12 odds shortening 2.16 odds lengthening (click on odds to bet)	statto odds	ladbrokes	William Hill	betfred	Paddy Power	BLUESQ	bet365	bet James.com	Boylesports	888sport	ENTREBET	betdaq	GoalWn	betstar	BETDAQ	WEX	SPORTING INDEX
Chelsea	1.6	1.44	1.57	1.5	1.53	1.57	1.53	1.53	1.53	1.53	1.5	1.6	1.6	1.59	1.59	49-52	
Manchester United	3.6	5	5	5	5	5	5	5	5	5		5	5	5.3	5.3	35-38	
Arsenal	13	6	6	6	6	6	6	6	6	6	6	6	6	7.2	7.2	33-36	
Manchester City	160	15	15	15	13	15	15	17	15	17	15	15	10	22	22	20	21-24
Tottenham Hotspur	320	81	81	126	101	101	101	101	126	126	101	101	75	180	200	170	7-9
Liverpool	500	151	126	151	201	151	151	126	151	151	151	151	150	160	170	150	4.5-6
Everton	190	201	301	251	201	251	251	251	351	401	251	251	150	500	530	480	3.5-4.5
Aston Villa	999	501	751	751	751	601	751	751	1001	1001	751	751	150	1000	1000	930	0.5-1.5
Newcastle United	999	751	1001		1001	601	1501			1501	1501		500	1000	1000	950	0.25-1
Birmingham City	-	2001	2001		1001	1001	2501			2001	2501		1000	1000	1000	940	0.1-0.5
Blackburn Rovers	-	2001	1001		2001	1501	2001			2501	2001		2000	1000	1000	940	0.1-0.5
Sunderland	999	2001	1001		1251	1001	1501			2501	1501		500	1000	1000	950	0.1-0.5
Bolton Wanderers	-	2501	2001		1251	2001	2501			3501	2501		2000	1000	1000	950	
Fulham	-	2501	1001		1001	1001	2001			4001	2001		500	1000	1000	960	0.1-0.5
Stoke City	-	3501	2001		2501	1501	4001			3001	4001		1000	1000	1000	950	
West Bromwich Albion	999	2501	2001		501	1001	2501			1501	2501		5000	1000	1000	950	
West Ham United	-		2001		7501	2001				6001			1000	1000	1000	950	0.05-0.2
Wigan Athletic	-		2001		7501	2501				7501			5000	1000	1000	940	
Wolverhampton Wndrs	-		2001		2501	3001				7501			5000	1000	1000	950	
Blackpool	-		2001		7501	2501				12501			5100	1000	1000		

- Manchester United is second most likely to win, and Arsenal third
- Our model has the same top 3, but reflects a belief that Arsenal has a better chance of finishing top
- The odds reflected by the betting stats are far less supportive of a Chelsea victory than our model

Are those betting taking into account more factors? Are those betting less rational?

Betting on: **Chelsea**

Total matched on this event: **£4,749,760**

Betting summary - Volume: **£2,204,269**

Last price matched: **1.64**

### Price/Volume over time



Betting on: **Man Utd**

Total matched on this event: **£4,749,760**

Betting summary - Volume: **£1,181,261**

Last price matched: **5.30**

### Price/Volume over time



Betting on: **Arsenal**

Total matched on this event: **£4,749,760**

Betting summary - Volume: **£596,281**

Last price matched: **7.00**

### Price/Volume over time



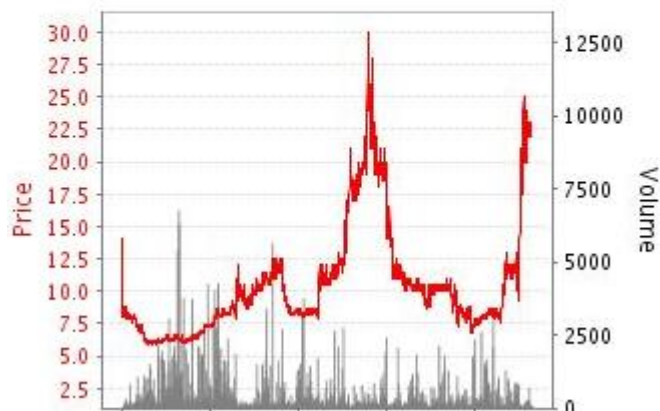
Betting on: **Man City**

Total matched on this event: **£4,749,760**

Betting summary - Volume: **£472,297**

Last price matched: **22.00**

### Price/Volume over time



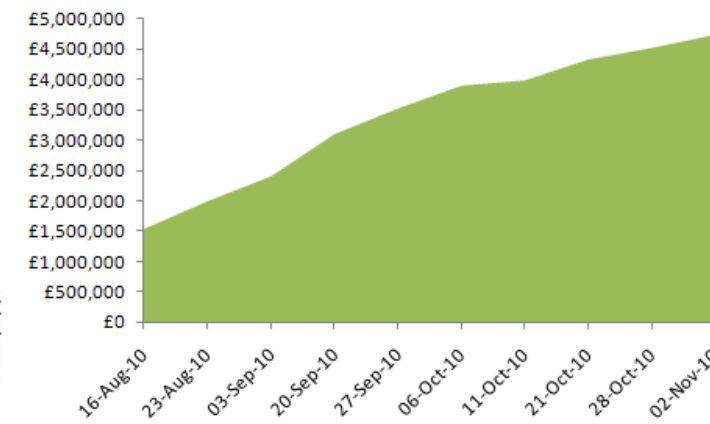
Betting on: **Liverpool**

Total matched on this event: **£4,749,760**

Betting summary - Volume: **£195,952**

Last price matched: **160.00**

### Price/Volume over time



**The odds are  
continuously changing**

# Bringing it back to our day jobs...

- Be aware of your model limitations
- Just because it works for the past – doesn't mean it'll work for the future
- The known unknowns, and the unknown unknowns
  - Otherwise, someone could get injured?
- Model predictions should be interpreted using actuarial judgement



**Is the efficient market hypothesis strong or weak?**

# Questions or comments? (No curve balls please...)

Expressions of individual views by members of The Actuarial Profession and its staff are encouraged.

The views expressed in this presentation are those of the presenter.

