Predicting the World Cup

Dr Christopher Watts
Centre for Research in Social Simulation
University of Surrey

Possible Techniques

- Tactics / Formation (4-4-2, 3-5-1 etc.)
  - Space, movement and constraints
  - Data on passes attempted and received
  - Agent-based simulation? Robo soccer? Computer games?
- Picking a team
  - Data on who was playing whenever Rooney scored
  - Combinatorial optimisation
- Statistical modelling of matches
  - Data on goals scored in each match
  - Poisson model, Markov Chain Monte Carlo (MCMC)
  - Data on win/draw/lose
  - Probit model
- Prediction distinct from Explanation

Why *MCMC*?

- Data readily available
  - – BBC Sport website, FIFA website, etc.
- Answers interesting questions
  - – Who is likely to win this match?
  - – What odds of it ending 5-1?
- Answers these questions on a large scale
  - – Dozens of matches from one model

Procedure

- Get dataset
- Fit mathematical model (training)
- Don’t overfit model (validation)
- Predict outcomes or estimate odds (test)
- Go to William Hill, Ladbrokes etc.

Some Reading

- Dixon & Coles (1997)
- Karlis (2003)
- Graham & Stott (2008)
- Spiegelhalter & Ng (2009)
- Greenhough et al. (2002)
- Denis Campbell, The Observer, Sunday 28 May 2006

The model

- Let # goals scored by i against j be Poisson-distributed with parameter

  lambda = (A_i / D_j)

  where

  A_i is Attacking strength of i

  D_j is Defensive strength of j

Premier League

- 20 teams in division so

  20 attack + 20 defence = 40 unknowns

- But every team will play every other home and away

  20 x 19 = 380 matches per season

  – Use some of this as training data, some as validation and predict the rest

- Network of known results constrains the unknown parameters

Questionable assumptions (1)

- Poisson distribution
  - – Scoring one goal is no more likely after scoring three than after scoring none
    - No confidence / morale effects, no learning
  - – 9:0 shouldn't appear every other season (nor every other century?)
- Alternatives
  - – Weibull function (Discretised)
    - Two parameters (alpha, beta) in place of lambda
  - – Negative Binomial

Questionable assumptions (2)

- Same parameters all season?
  - New teams members in August and January
  - Rain-soaked pitches lead to defensive mistakes (esp. in November)
  - Fatigue (African Cup of Nations, Europe)
  - Injuries
  - Managerial “tinkering”, “rotation”
- Extra parameters for seasonality?

Can we gamble?

- Bookmakers' odds reflect:
  - their need to make a profit
    - so implied probabilities will not sum up to 1
  - their need to hedge bets
    - 1 million patriots bet on England
  - more information than just past results
    - e.g. Rio Ferdinand is out! (8 to 1, from 7 to 1)
- Identify undervalued outcomes
  - E.g. bet against the favourite
- Operate on a large scale (Expensive!)

MCMC Simulation

- Each combination of 20x2 parameters represents a possible system state
- During simulation system jumps from state to (more likely) state
- Over time system tends to something close to the most likely state (hopefully)
  - The parameter values that best fit the data

Max Likelihood

- Likelihood Ratio

$$\frac{P(\text{ Results data } | \text{ Theory1 })}{P(\text{ Results data } | \text{ Theory2 })}$$

- $P(X=x) = \text{lambda}^{x} * e^{-\text{lambda}} / x!$
- Algorithm options:
  - – Always adopt the larger (Ascent)
  - – Random choice stratified using odds ratio (Gibbs sampling)

Log Likelihood

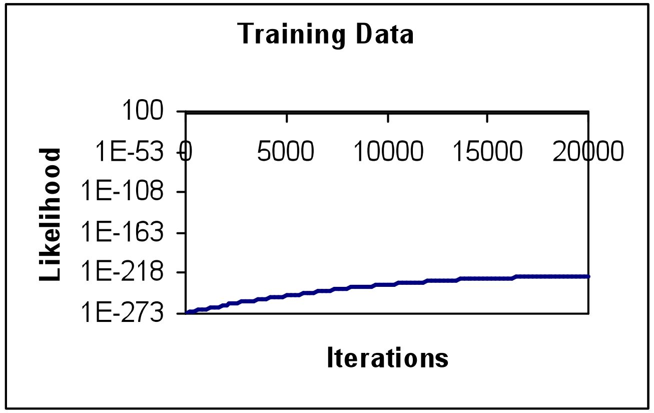
- Likelihood of the theory parameters:

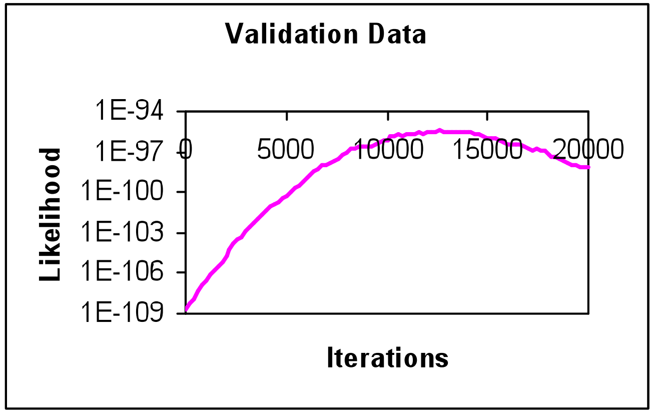
  P (Goals scored X_{ij} = x | X_{ij} ~ Pois(A_i / D_j))

- Multiply corresponding probability for each goal score (home, away) for each match in data set
  - Equivalently: Sum the log likelihoods
- Assumptions!
  - Every match result is independent of every other
  - Goals scored is independent of goals conceded

Validation data

- Use separate validation data to demonstrate when model is over-fit to training data
- Likelihood given *validation data* peaks
  - – Around 13000 iterations in this example





Premiership 2009-10

Team	HP	AP	GH	GCH	GA	GCA	H_Att	H_Def	A_Att	A_Def
Man Utd	17	16	2.65	0.65	2.00	1.00	1.55	1.48	1.40	1.00
Chelsea	16	17	3.25	0.88	1.88	0.94	1.45	0.65	1.43	1.32
Everton	17	16	1.88	1.18	1.25	1.50	1.43	1.02	1.37	0.80
Liverpool	16	17	2.50	0.81	0.82	1.18	1.41	1.34	0.52	1.45
Arsenal	17	16	2.59	0.88	1.94	1.19	1.40	0.96	1.33	1.26
Man City	15	17	2.20	1.07	1.82	1.41	1.38	0.98	1.38	0.67
Hull	15	17	1.40	1.47	0.59	2.59	1.25	0.82	0.58	0.54
Aston Villa	16	16	1.63	0.81	1.13	1.19	1.16	1.09	0.93	0.93
West Ham	16	17	1.56	1.63	0.88	1.82	1.08	0.59	0.61	0.59
Fulham	16	16	1.50	0.75	0.69	1.56	1.08	1.38	0.57	0.68
Stoke	17	15	1.35	1.12	0.60	1.07	1.07	0.74	0.57	1.17
Tottenham	16	16	2.19	0.63	1.44	1.38	1.05	1.42	1.26	0.95
Birmingham	17	16	1.00	0.71	1.06	1.63	1.02	1.29	0.91	0.65
Sunderland	17	16	1.76	1.00	0.88	2.13	1.02	1.13	0.83	0.57
Bolton	17	16	1.29	1.65	0.88	2.06	1.01	0.72	0.67	0.54
Blackburn	16	17	1.50	0.88	0.65	2.12	0.99	1.18	0.49	0.65
Portsmouth	17	16	1.18	1.71	0.50	1.94	0.96	0.59	0.52	0.61
Burnley	17	16	1.24	1.41	0.69	2.94	0.75	0.61	0.59	0.63
Wigan	16	17	0.88	1.25	0.94	2.59	0.61	0.75	0.85	0.56
Wolverhampton	16	17	0.63	1.25	1.06	1.82	0.60	1.04	1.02	0.78

- 4th April, 2-3 matches to go

Prediction reliability?

- 2009-10 saw a tight contest at top and bottom!
- Even with 3 games to go prediction was inaccurate

	16-Mar-10	21-Mar-10	04-Apr-10
Man Utd	1	1	3
Arsenal	2	2	2
Chelsea	3	3	1
Tottenham	4	4	5
Aston Villa	5	6	7
Man City	6	5	4
Liverpool	7	7	6
Everton	8	8	8
Hull	17	17	17
West Ham	18	18	18
Portsmouth	19	20	19
Burnley	20	19	20

The World Cup

- 32 nations, selected from 207, 6 continents
- Fit FIFA data for last 5 years
  - World & Continental competitions
  - Qualifiers (Home + Away)
  - Finals (Usually only one Home team)
  - Friendlies (Home or Away)
- Few inter-continental matches
- Longer time scale
  - 2-3 matches, then long breaks
  - Finals: 7 matches in 5 weeks

Monte Carlo Simulation

- Given model of teams simulate the tournament
- Sample scores for each match
- Calculate points, winners
- Repeat 10000 times

- Estimate odds for:
  - Particular teams reaching the Last 16, Quarter Finals etc. and Winning the competition

Beat the bookies

- Estimate odds
- If bookmakers offer longer odds…

- England (rows) vs. USA (columns)
  - None of these are tempting

	0	1	2	3	4
0	12.1	14.2	34.2	121.1	564.5
1	6.7	7.9	19.6	70.4	329.6
2	8.0	9.4	23.0	82.5	385.6
3	14.7	17.2	41.2	145.5	677.2
4	35.8	41.6	97.6	341.5	1585.0

Parameters fit and estimated chances

Group	Team	Att	Def	Rank_Att	Rank_Def	Failed	GR	GW	QF	SF	F	W
G	Brazil	4.32	2.79	1	2	6.3%	5.0%	12.7%	17.5%	15.4%	10.8%	32.3%
D	Germany	4.12	2.08	2	15	16.6%	7.2%	10.6%	19.6%	14.1%	16.7%	15.4%
H	Spain	3.23	2.79	4	2	13.9%	12.6%	19.7%	15.9%	18.9%	7.6%	11.5%
E	Netherlands	2.93	3.23	5	1	19.0%	6.5%	9.6%	32.8%	14.4%	7.2%	10.6%
B	Argentina	3.39	2.18	3	10	20.7%	8.0%	15.8%	23.0%	12.7%	11.6%	8.2%
C	England	2.79	2.41	7	6	10.3%	13.0%	28.8%	18.6%	14.9%	9.2%	5.3%
A	France	2.53	2.65	11	4	24.7%	13.1%	15.7%	23.2%	12.1%	7.2%	4.0%
F	Italy	2.53	2.53	11	5	15.7%	14.6%	28.2%	23.8%	11.4%	3.5%	2.8%
D	Serbia	2.93	1.89	5	26	43.0%	14.0%	6.4%	17.0%	11.4%	5.7%	2.5%
E	Denmark	2.65	2.18	8	10	41.9%	12.1%	7.2%	23.6%	10.2%	2.8%	2.2%
G	Portugal	2.41	2.29	16	9	39.2%	26.3%	7.1%	15.3%	9.0%	2.0%	1.2%
A	Uruguay	2.53	1.89	11	26	41.1%	17.9%	11.4%	18.1%	7.3%	3.3%	1.0%
B	Greece	1.63	2.18	56	10	61.0%	17.6%	9.2%	8.6%	2.7%	0.7%	0.2%
C	USA	2.08	1.63	27	40	38.3%	30.6%	14.5%	11.5%	4.0%	1.1%	0.1%
H	Chile	1.98	1.71	31	35	67.3%	19.4%	6.7%	4.7%	1.6%	0.3%	0.0%

Any tips?

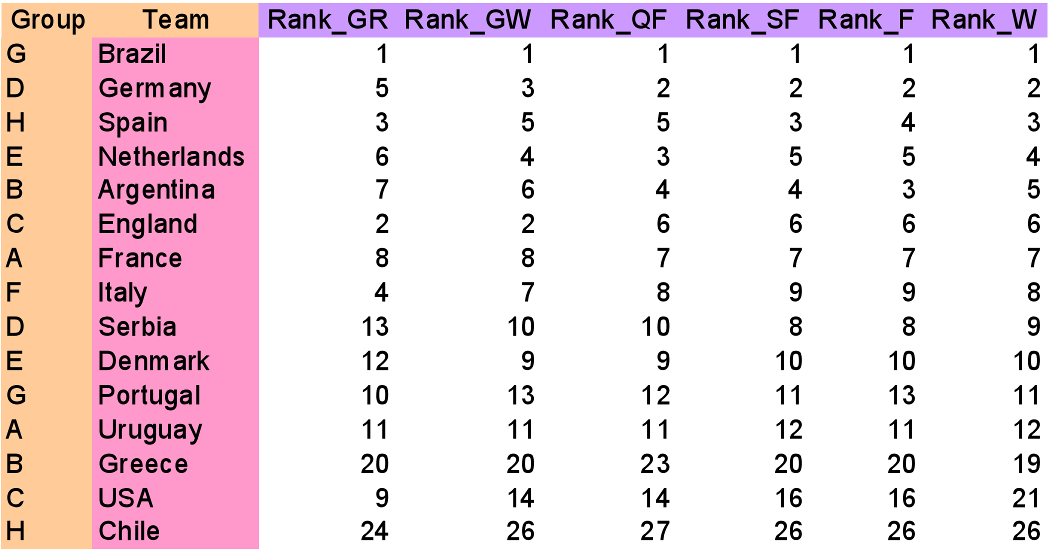
- Model says Brazil have odds of 2.1 to 1
  - William Hill offer 9 to 2 (=4.5:1)
- England bad bet at 18 to 1 (WH: 8 to 1)
- Germany best bet:
  - Model says 11 to 2 (WH: 14 to 1!)
  - Denmark, Serbia also undervalued
- Forget Italy, Portugal
  - It’s not going to be USA, Chile or Greece either…

Surprised?

- Germany again?!?
  - Had Home advantage 4 years ago
  - Ballack is out this time
  - *Bundesliga* uses balls from *Adidas*
- Why are Spain not higher?

Easy group?

- Ranked by Chance of getting at least this far



Group	Team	Rank_GR	Rank_GW	Rank_QF	Rank_SF	Rank_F	Rank_W
G	Brazil	1	1	1	1	1	1
D	Germany	5	3	2	2	2	2
H	Spain	3	5	5	3	4	3
E	Netherlands	6	4	3	5	5	4
B	Argentina	7	6	4	4	3	5
C	England	2	2	6	6	6	6
A	France	8	8	7	7	7	7
F	Italy	4	7	8	9	9	8
D	Serbia	13	10	10	8	8	9
E	Denmark	12	9	9	10	10	10
G	Portugal	10	13	12	11	13	11
A	Uruguay	11	11	11	12	11	12
B	Greece	20	20	23	20	20	19
C	USA	9	14	14	16	16	21
H	Chile	24	26	27	26	26	26

- Spain could face Brazil, Portugal or Ivory Coast in the Last 16
- Things get tougher for England after the Group stage

Extensions

- Reweighted data by age
  - – Let importance of result decay exponentially over time
- Focus on last 12 months
  - – Spain now become favourite
  - – England still only 5% chance!

Any lessons?

- We model (adaptive!) human social behaviour
  - Use MCMC to fit network data
    - As in Siena / stocnet (ERGM)
  - Energy models (my PhD topic)
    - Individuals energise/de-energise each other when they interact
    - This affects future interactions
      - interaction ritual chains theory (Collins)
  - Stratification: success breeds success (as in science)
  - Learning models (Learning to beat x? To fear x?)