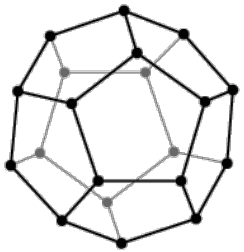


ELAVIO 2017

Escuela Latinoamericana de Verano en Investigación Operativa
(Operations Research Summer School for Young Latin American Scholars)

Buenos Aires and Miramar, Argentina
February 24th to March 4th

Applications of Operations Research and Statistics to Sports Analytics

Sports Analytics (Statistics + OR) for predicting soccer tournament outcomes

M. Guajardo (NHH) & D. Sauré (UChile)

Why predicting outcomes in sports?

- Current use: “enhancing the attractiveness of football broadcasting”
  - **TVN**, TV2 Norway, ESPN
- Use in designing betting strategies
  - Pre game vs. **In-play betting**
  - **Huge market!!**
- Fun thing to do
  - Hard to argue against data
  - Betting against friends…

First step: what is a prediction?

- A fist look at the betting markets

- Fair bets: $Odd(H) \cdot P(H) = 1$
- Bookmakers set unfair odds to make a profit

$$\frac{1}{Odd(H)} + \frac{1}{Odd(D)} + \frac{1}{Odd(A)} := Z > 1$$

- Corrected inverse-probabilities: $P(H) = \frac{1}{Odd(H) \cdot Z}$

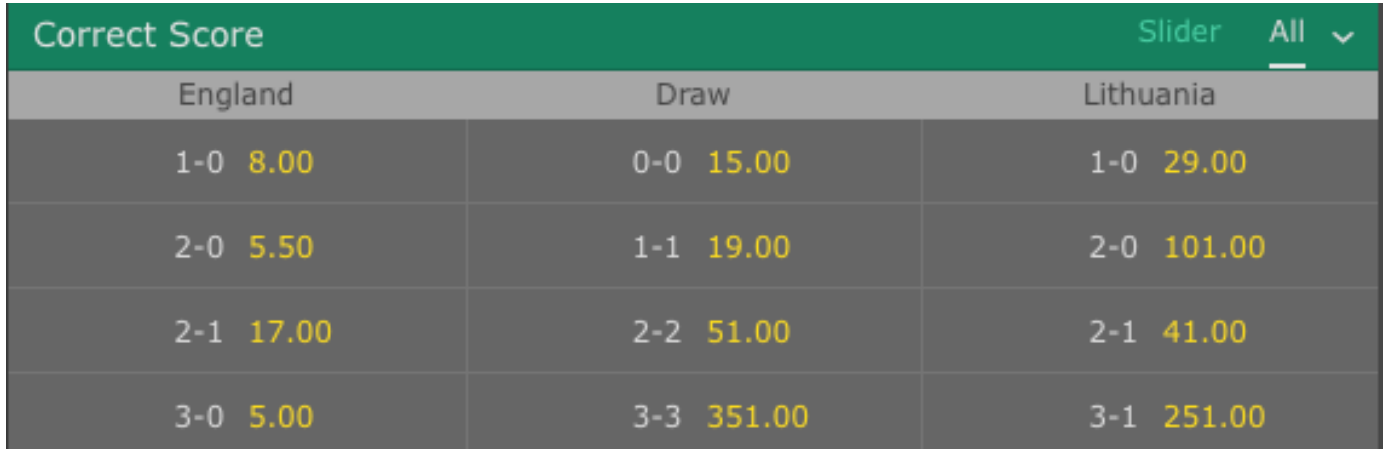
Predictions on a broader context

- Broader spectrum of possible bets
  - e.g. Asian Handicap, First team to score, Exact goals, Corners, Cards, etc.

To Win Outright Change Market	
Each-way 1/2 1-2	Book Closes 22 Mar 20:00
Germany 5.50	Argentina 9.00
France 9.00	Brazil 9.00
Spain 10.00	Belgium 15.00
Italy 17.00	England 17.00
Portugal 29.00	Colombia 34.00
Russia 34.00	Holland 34.00
Uruguay 41.00	Croatia 41.00
Chile 41.00	Mexico 67.00

Predictions on a broader context

- Prediction as a distribution over possible outcomes
  - Goals per team, cards, injuries, etc



Correct Score		Slider	All
England	Draw	Lithuania	
1-0 8.00	0-0 15.00	1-0 29.00	
2-0 5.50	1-1 19.00	2-0 101.00	
2-1 17.00	2-2 51.00	2-1 41.00	
3-0 5.00	3-3 351.00	3-1 251.00	

  - Find a **joint distribution** over outcomes (goals)
  - Extrapolate to hypothetical games (tournaments)

Notation

- Game as joint stochastic processes

$$N_i(s,t) = \#\text{ goals scored by team } i \text{ between time } s \text{ and } t \quad i \in \{h,a\}$$

  - We make the distinction between home and away teams, understanding that all teams may play at home against any other team
  - Need the joint distribution of $(N_h(s,t), N_a(s,t))$ for any pair of teams, and conditional on $(N_h(0,u), N_a(0,u), u \leq s)$

  - For now, set time so *t*=1 corresponds to 90', and define

$$P(g_h, g_a) := P(N_h(0,1) = g_h, N_a(0,1) = g_a)$$

MODELS FOR PRE-GAME PREDICTION

Quick Background Check

- Poisson distribution
  - $X \sim Poisson(\lambda)$

$$P(X = k) = \frac{1}{k!}\lambda^k e^{-\lambda}$$

- Binomial approximation
  - $X_n \sim Binom(n, p_n) \quad ; \quad n \cdot p_n = \lambda$

$$X_n \Rightarrow Poisson(\lambda)$$

Models for Pre-game Prediction

- Frequentist approach
  - Estimate the joint distribution of the goals scored by two teams

$$P(g_h, g_a) = \frac{\#\text{games between H and A that ended with score } (g_h, g_a)}{\#\text{games between H and A}}$$

  - Main problem: not enough data

LAST FIVE GAMES					DATE	COMPETITION
Argentina		2-1		Chile	Jun 6, 2016	Copa America
Chile		1-2		Argentina	Mar 24, 2016	World Cup Qualifying - CONMEBOL
Chile		0-0		Argentina	Jul 4, 2015	Copa America
Chile		1-2		Argentina	Oct 16, 2012	World Cup Qualifying - CONMEBOL
Argentina		4-1		Chile	Oct 7, 2011	World Cup Qualifying - CONMEBOL

Models for Pre-game Prediction

- Solution: assume some structural (parametric, very, very simple) model
  - reduce number of parameters to estimate
  - leverage data from other matches

$$P(g_h, g_a) = f_\alpha(g_h, g_a)$$

model parameters

  - Calibrate parameters using observed history

Models for Pre-game Prediction

- Moroney (56): number of goals follow a Negative Binomial Distribution

$$P(g_h, g_a) = \binom{g_h + r_h - 1}{g_h} (1 - p_h)^{r_h} p_h^{g_h} \cdot \binom{g_a + r_a - 1}{g_a} (1 - p_a)^{r_a} p_a^{g_a}$$

  - Two parameters (r,p)
  - Interpretation: teams attempt to score until failing r times
  - (Alternative explanation) Poisson goals with random (gamma) rate
  - Weakness: offensive capability does not depend on rival nor on home/ away status

Models for Pre-game Prediction

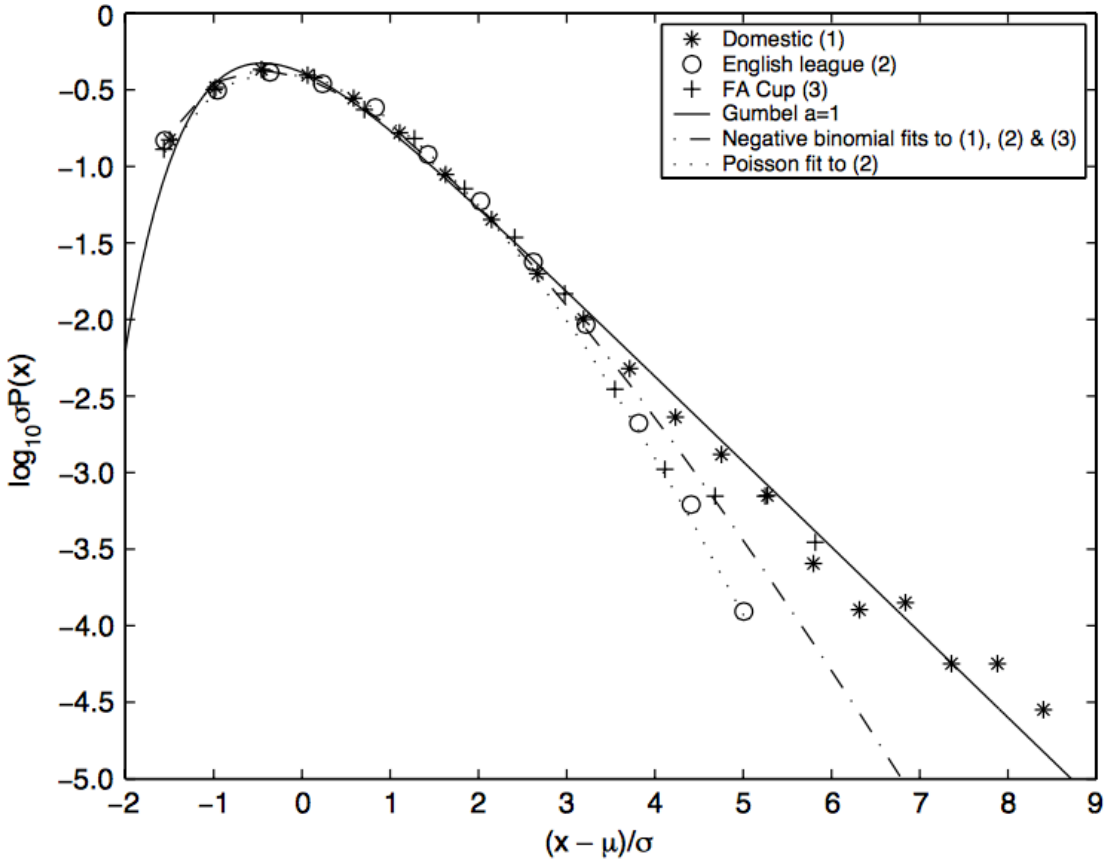
- Greenhough et al (01): number of goals follow a GEV

$$P(g_h, g_a) = \frac{1}{\sigma_h} e^{-e^{-\frac{g_h-\mu_h}{\sigma_h}} - \frac{g_h-\mu_h}{\sigma_h}} \cdot \frac{1}{\sigma_a} e^{-e^{-\frac{g_a-\mu_a}{\sigma_a}} - \frac{g_a-\mu_a}{\sigma_a}}$$

  - Does not account for teams abilities
  - Focus on fitting overall domestic goal distribution
  - Conclude Poisson model provides good enough fit to English leagues

Models for Pre-game Prediction

- Greenhough et al (01): number of goals follow a GEV



Models for Pre-game Prediction

- Maher (82): number of goals follow a Poisson Distribution

$$P(g_h, g_a) = \frac{1}{g_h!}\,\lambda_{h,a}^{g_h}\,e^{-\lambda_{h,a}} \cdot \frac{1}{g_a!}\,\lambda_{a,h}^{g_a}\,e^{-\lambda_{a,h}}$$

  - Interpretation: teams have many possessions during a game, independent of everything, each ends with a goal with probability *p*
  - Improvement:

$$\lambda_{h,a} = \alpha_h \cdot \beta_a \quad ; \quad \lambda_{a,h} = \alpha'_a \cdot \beta'_h$$

Models for Pre-game Prediction

- Dixon and Coles (97)'s modeling guidelines
  - Should account the different abilities of both teams
  - Should account for the *home effect*
  - A team's ability should be base on recent performance
  - A team's ability is likely to be best summarized in separate measures of their abilities to attack and to defend
  - When assessing past performance, one should account for the rivals' abilities
- Study of independent Poisson assumption
  - Poisson (ok…)
  - Independence (not so much for low scores…)

Models for Pre-game Prediction

- Dixon and Coles (97)'s model: *almost* independent Poisson goals

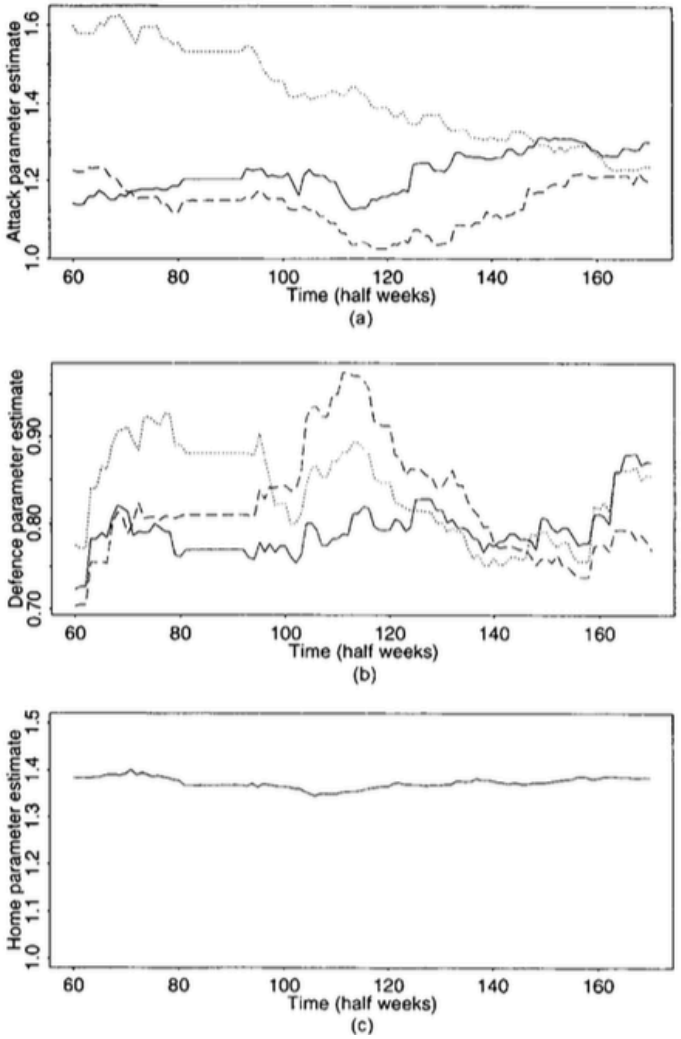
$$P(g_h, g_a) = \tau_{\lambda_{h,a},\lambda_{a,h}}(g_h, g_a) \cdot \frac{1}{g_h!}\lambda_{h,a}^{g_h} e^{-\lambda_{h,a}} \cdot \frac{1}{g_a!}\lambda_{a,h}^{g_a} e^{-\lambda_{a,h}}$$

$$\lambda_{h,a} = \alpha_h \cdot \beta_a \, \gamma \quad ; \quad \lambda_{a,h} = \alpha_a \cdot \beta_h$$

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1 - \lambda\mu\rho & \text{if } x = 0; y = 0 \\ 1 + \lambda\rho & \text{if } x = 0; y = 1 \\ 1 + \mu\rho & \text{if } x = 1; y = 0 \\ 1 - \rho & \text{if } x = 1; y = 1 \\ 1 & \text{otherwise} \end{cases}$$

Models for Pre-game Prediction

- Dixon and Coles (97)'s model's enhancement
  - Index parameters by time
  - Because there is not enough data, assume locally constant parameters (maximize likelihood locally by weighting-down history)



Models for Pre-game Prediction

- Dyte and Clarke (00): FIFA rating-based independent Poisson model

$$P(g_h, g_a) = \frac{1}{g_h!}\,\lambda_{h,a}^{g_h}\,e^{-\lambda_{h,a}} \cdot \frac{1}{g_a!}\,\lambda_{a,h}^{g_a}\,e^{-\lambda_{a,h}}$$

$$\lambda_{h,a} = exp(a + b\,R_h + c\,R_a + v_h 1_{\text{home}})$$
$$\lambda_{a,h} = exp(a + b\,R_a + c\,R_h + v_a 1_{\text{away}})$$

  - Main issue: FIFA ratings are the **worse***

Models for Pre-game Prediction

- Janke et al (09): limits of discrete-time model
  - N time periods, teams score with probabilities p
  - Scoring probabilities $(p_h(N_h(0,t)), p_a(N_a(0,t)))$

$$p(n) = p(n-1) + \kappa \quad ; \quad p(n) = p(n-1)\kappa$$

  - Additive self-affirmation converges to NBD
  - GEV arises as a limit in which the probability of scoring on a period depends on the tail of a known distribution

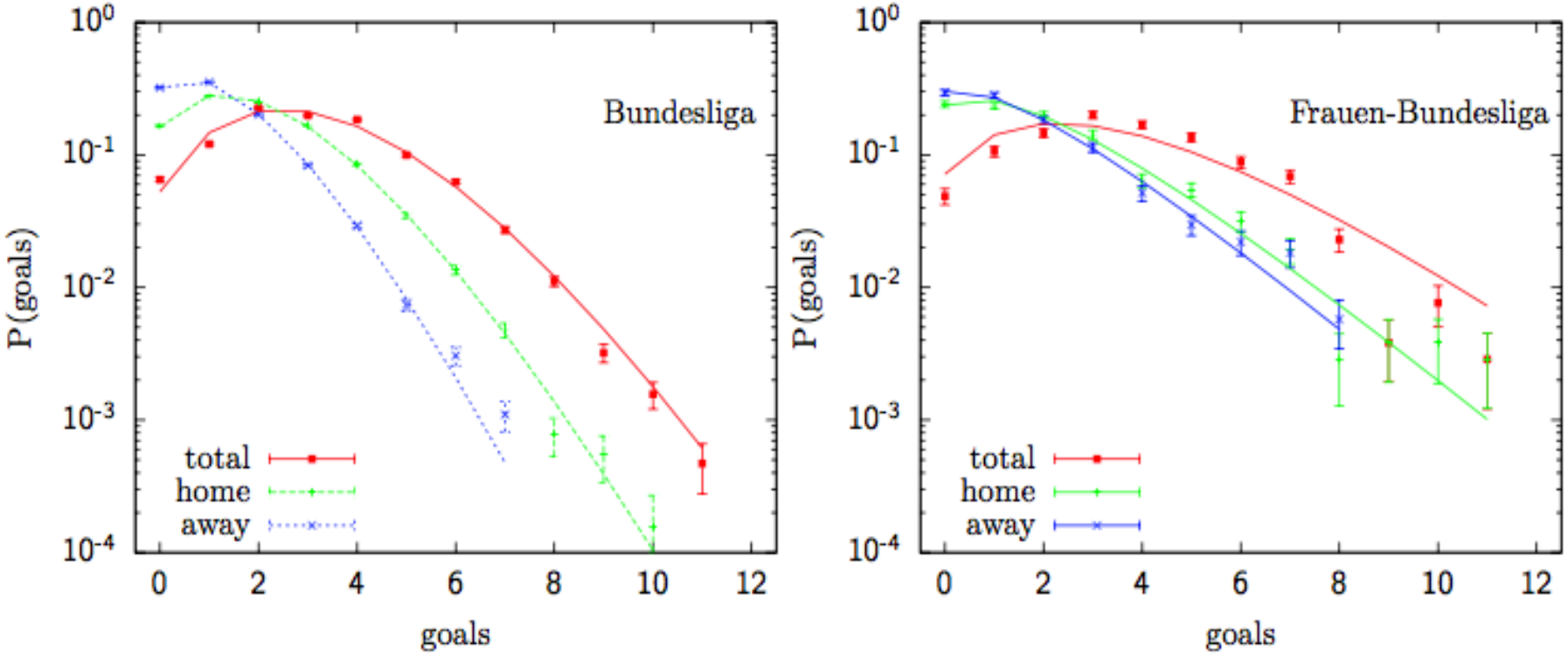
Models for Pre-game Prediction

- Janke et al (09): Bundesliga

		Bundesliga		Frauen-Bundesliga	
		Home	Away	Home	Away
Poisson	λ	1.91 ± 0.01	1.16 ± 0.01	1.78 ± 0.04	1.36 ± 0.04
	χ^2/d.o.f.	9.21	9.13	14.6	14.4
NBD	p	0.11 ± 0.01	0.09 ± 0.01	0.45 ± 0.03	0.46 ± 0.03
	r	16.24 ± 1.82	12.08 ± 1.69	2.38 ± 0.24	1.97 ± 0.22
	p_0	0.0202	0.0125	0.0160	0.0133
	κ	0.0012	0.0010	0.0067	0.0068
	χ^2/d.o.f.	1.08	2.22	2.32	1.37
GEV	ξ	-0.10 ± 0.01	-0.02 ± 0.01	0.04 ± 0.04	0.25 ± 0.07
	μ	1.17 ± 0.02	0.57 ± 0.01	0.83 ± 0.08	0.77 ± 0.07
	σ	1.33 ± 0.01	0.96 ± 0.01	1.49 ± 0.06	1.18 ± 0.05
	χ^2/d.o.f.	3.43	7.95	3.40	1.55
Gumbel	μ	1.18 ± 0.01	0.58 ± 0.01	0.81 ± 0.08	0.58 ± 0.07
	σ	1.21 ± 0.01	0.94 ± 0.01	1.53 ± 0.05	1.31 ± 0.05
	χ^2/d.o.f.	24.5	7.26	3.17	4.09

Models for Pre-game Prediction

- Janke et al (09): Bundesliga



Background in Statistical Inference I

- Parametric model of uncertainty, use data to estimate model parameters
- Data:

$$\{(g_h^t, g_a^t), (a,h) \in G,\, t = 1, \ldots, T\}$$

- Likelihood function:

$$L(\text{data}, \alpha) = \prod_{(h,a)\in G,\, t \leq T} f_\alpha(g_h^t, g_a^t)$$

- Maximum Likelihood estimator (MLE): use as parameter estimates those that maximize the likelihood of observing the data realized

$$MLE = \arg\max_{\alpha \in \Lambda} L(\text{data}, \alpha)$$

Background in Statistical Inference I

- Example: independent Poisson model (Maher 82)

$$P(g_h, g_a) = \frac{1}{g_h!}\,\lambda_{h,a}^{g_h}\,e^{-\lambda_{h,a}} \;\cdot\; \frac{1}{g_a!}\,\lambda_{a,h}^{g_a}\,e^{-\lambda_{a,h}}$$

$$\lambda_{h,a} = \alpha_h \cdot \beta_a \quad ; \quad \lambda_{a,h} = \alpha'_a \cdot \beta'_h$$

- First step: identification of the model

$$\sum_i(\alpha_i - \beta_i) = 0, \qquad \sum_i(\alpha'_i - \beta'_i) = 0$$

- Second step: write (log) likelihood function, and solve FOC (closed-form expression)

Models for Pre-game Prediction

- Rue & Salvesen (00) :Bayesian dynamic linear model
  - Poisson goals on a given match
  - Rates dependent on attacking and defending skills
  - Skills themselves are not constant throughout a season
  - Also, strong teams underestimate weaker teams (and the other way around)

$$P(g_h^t, g_a^t | \text{history}) = \frac{1}{g_h^t!}(\lambda_h^t)^{g_h^t} e^{-\lambda_h^t} \cdot \frac{1}{g_a^t!}(\lambda_a^t)^{g_a^t} e^{-\lambda_a^t} \ \rho(g_h^t, g_a^t)$$

$$log\ \lambda_h^t = c_h + x_t \left(a_h^t - b_a^t - \frac{\gamma}{2}(a_h^t + d_h^t - a_a^t - d_a^t) \right)$$

$$log\ \lambda_a^t = c_a + x_t \left(a_a^t - b_h^t + \frac{\gamma}{2}(a_h^t + d_h^t - a_a^t - d_a^t) \right)$$

$x_t \sim Bernoulli(p)$

random variables

Models for Pre-game Prediction

- Rue & Salvesen (00) :Bayesian dynamic linear model

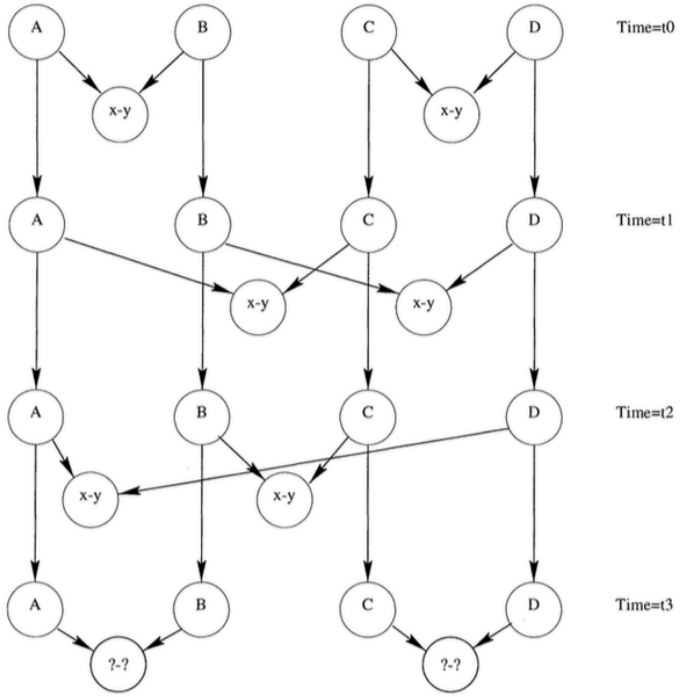
$$a_i^t | a_i^s \sim N(a_i^s, (t-s)\sigma_{a,i}^2), \qquad d_i^t | d_i^s \sim N(d_i^s, (t-s)\sigma_{d,i}^2)$$

  - Dynamics inspired by Brownian motions
  - Choose p = 0.2
  - Time scale chosen so that

$$E[\log \lambda_h^t | \text{history up to } s] = \lambda_h^s$$
$$Var[\log \lambda_h^t | \text{history up to } s] = 2\sigma_{a,h}^2 (t-s)$$

Models for Pre-game Prediction

- Rue & Salvesen (00) : Inference
  - 4 teams example
  - Likelihood function written using conditional distribution



$$\pi(\theta) = \pi(a_A^{t_0}, d_A^{t_0})\,\pi(a_B^{t_0}, d_B^{t_0})\,\pi(a_C^{t_0}, d_C^{t_0})\,\pi(a_D^{t_0}, d_D^{t_0})$$
$$\times\, \pi(g_A^{t_0}, g_B^{t_0} | a_A^{t_0}, d_A^{t_0}, a_B^{t_0}, d_B^{t_0})\,\pi(g_C^{t_0}, g_D^{t_0} | a_C^{t_0}, d_C^{t_0}, a_D^{t_0}, d_D^{t_0})$$
$$\times\, \pi(a_1^{t_1}, d_1^{t_1} | a_1^{t_0}, d_1^{t_0})\,\pi(a_2^{t_1}, d_2^{t_1} | a_2^{t_0}, d_2^{t_0})\,\pi(a_3^{t_1}, d_3^{t_1} | a_3^{t_0}, d_3^{t_0})\,\pi(a_4^{t_1}, d_4^{t_1} | a_4^{t_0}, d_4^{t_0})$$
$$\times\, \pi(g_A^{t_1}, g_C^{t_1} | a_A^{t_1}, d_A^{t_1}, a_C^{t_1}, d_C^{t_1})\,\pi(g_B^{t_1}, g_D^{t_1} | a_B^{t_1}, d_B^{t_1}, a_D^{t_0}, d_D^{t_0})$$
$$\times \ldots$$

Background in Statistical Inference II

- In Bayesian models, underlying parameters are random, whose distribution can be computed using Bayes rule

- Markov Chain Monte Carlo (MCMC)
  - No closed-form for the posterior
  - Approximate numerically using Monte Carlo
  - Construct a Markov Chain whose stationary distribution matches the posterior distribution, simulate, let it reach steady state, and sample from it

Background in Statistical Inference II

- Suppose we want to sample from

$$p_i = a_i / (\sum_{j=1}^{N} a_j), \quad i = 1, \ldots, N$$

- Consider MC with N states and symmetric transition matrix Q
- Modify transition so that on each period there is a transition from *i* to *j* with probability

$$P_{i,j} = q_{i,j} \min\{1, a_j / a_i\}, \quad i \neq j$$

  - Metropolis, Metropolis-Hastings,…
- Computationally expensive, but always* available!

MODELS FOR IN-PLAY PREDICTION

Quick Background Check continued

- Time-homogeneous Poisson Process
  - N(0) = 0 with probability one
  - Counting process with independent and stationary increments
  - Number of events on interval of length *t* distributed Poisson(λt)

$$P(N(t) = k) = \frac{1}{k!}(\lambda t)^k e^{-\lambda t}$$

- Alternative characterization
  - N(0)=0 with probability one
  - Counting process with independent increments
  - $P(N(t+s) - N(t) = 1) = \lambda t + o(t)$

    $P(N(t+s) - N(t) > 1) = o(t)$
- Yet another characterization
  - Exponentially distributed inter-arrival times

Quick Background Check continued

- Non-homogeneous Poisson Process
  - N(0) = 0 with probability one
  - Counting process with independent increments
  - Number of events on between time s and *t* distributed

$$P(N(s,t) = k) = \frac{1}{k!} m(s,t)^k e^{-m(s,t)}$$

$$m(s,t) = \int_s^t \lambda(u) du$$

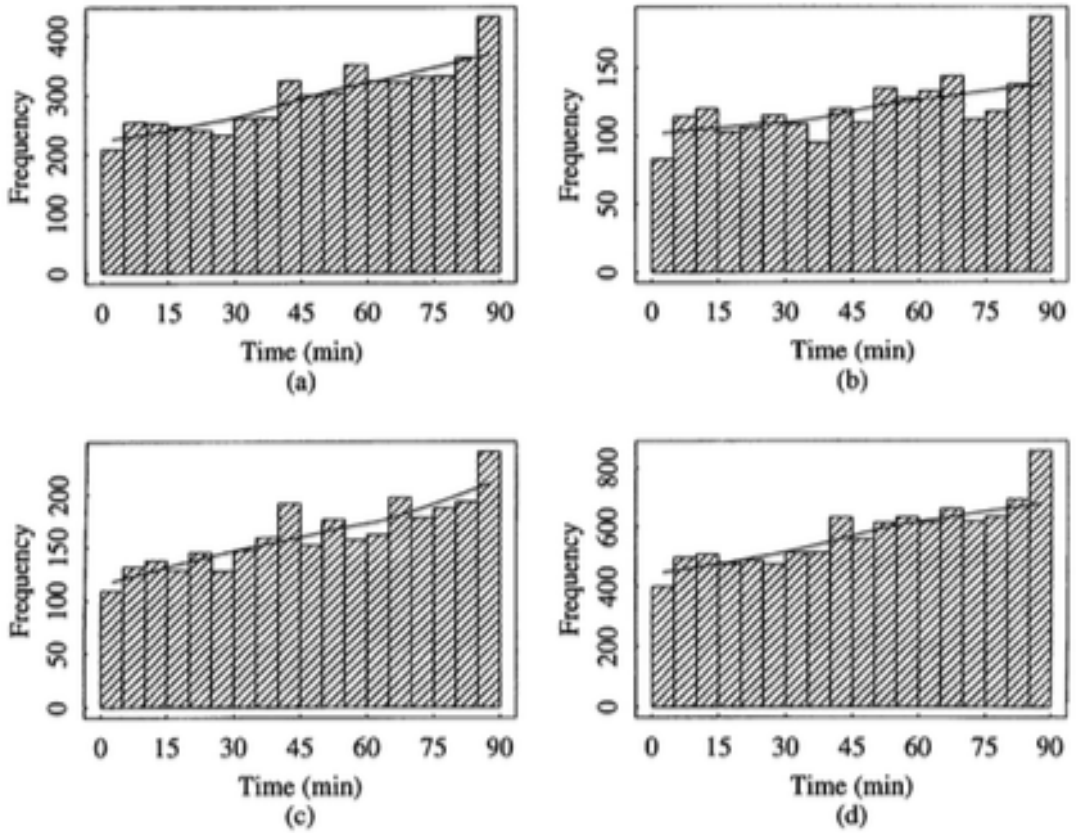
- Alternative characterization
  - N(0)=0 with probability one
  - Counting process with independent increments
  - $P(N(s, s+t) = 1) = \lambda(s)t + o(t)$
  - $P(N(s, s+t) > 1) = o(t)$

Quick Background Check continued

- Birth and death processes
  - Stochastic process on integers
  - Time on a given state is exponentially distributed
  - Exit from a state triggered either by a birth or a death
  - $((\lambda_n, \mu_n), n \in \mathbf{Z})$
  - Closed-form expression for stationary probabilities
  - Special case of a Markov Chain
  - Pure-birth model: state-dependent Poisson Processes

Models for In-play Prediction

- Dixon and Robinson (98): histogram of goal times



Models for In-play Prediction

- Dixon and Robinson (98): score and time dependent non-homogeneous **Poisson Process**
  - A gradual increase in scoring rates (due to tiredness)
  - Variation due to dependence on the current score

Table 1. Estimates and standard errors of the rate of the time to the next goal, in a match picked at random, when the score is (0, 0), (1, 0), (0, 1), (1, 1), (2, 0), (0, 2) (2, 1), (1, 2) and (2, 2)†

Rate	*Estimate*	*Standard error*
ν_{00}	0.0250	0.0004
ν_{10}	0.0289	0.0007
ν_{01}	0.0293	0.0009
ν_{11}	0.0302	0.0010
ν_{20}	0.0353	0.0014
ν_{02}	0.0315	0.0018
ν_{21}	0.0327	0.0019
ν_{12}	0.0369	0.0026
ν_{22}	0.0372	0.0029

†For example, $\nu_{00} = 0.025$ corresponds to a scoring rate of one goal every 40 minutes on average while the score is (0, 0).

Models for In-play Prediction

- Dixon and Robinson (98): counting processes

$$N_i(s,t) = \#\text{ goals scored by } i \text{ between time } s \text{ and } t \quad i \in \{h,a\}, s \leq t$$

  – Non-homogeneous Poisson characterization

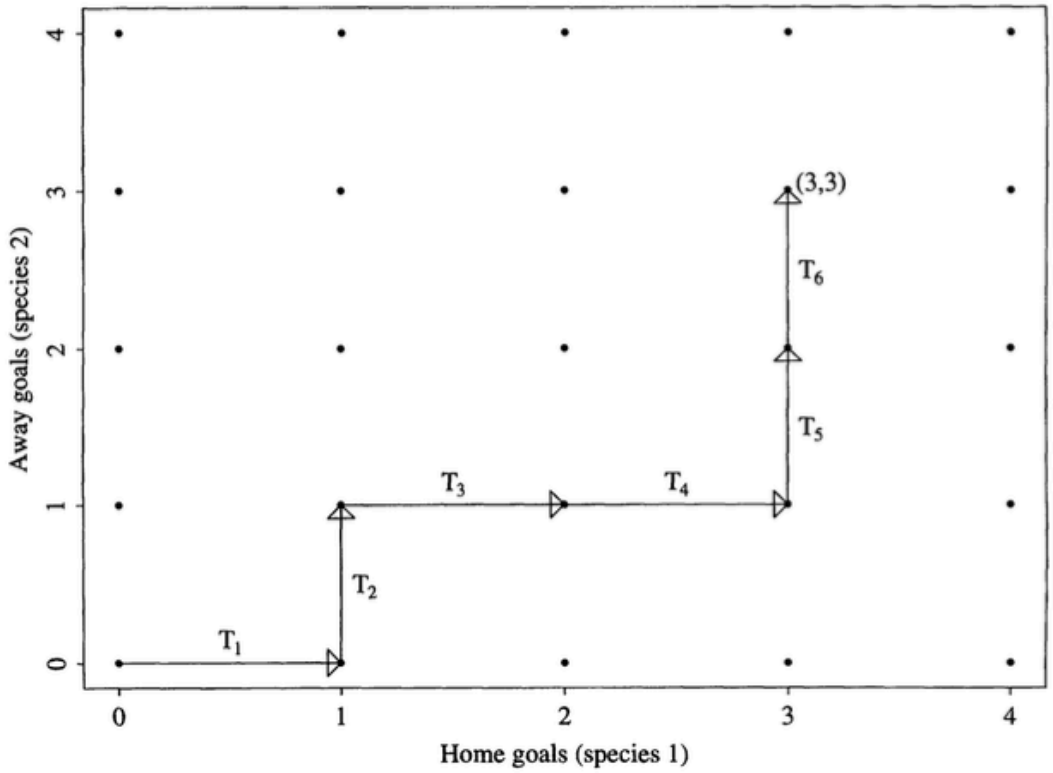
$$P(N_h(s,s+t) = 1 | N_h(s) = x, N_a(s) = y) = (\alpha_h\, \beta_a\, \gamma\, \rho(x,y) + \xi_h)t + o(t)$$
$$P(N_a(s,s+t) = 1 | N_h(s) = x, N_a(s) = y) = (\alpha_a \beta_h \rho'(y,x) + \xi_a)t + o(t)$$

$$P(N_h(s,s+t) > 1 | N_h(s) = x, N_a(s) = y) = o(t)$$
$$P(N_a(s,s+t) > 1 | N_h(s) = x, N_a(s) = y) = o(t)$$

+ independent increments property

Models for In-play Prediction

- Dixon and Robinson (98): two-dimensional birth process



Models for In-play Prediction

- Dixon and Robinson (98)
  - Consider score-depending function only with seven possible values
  - British leagues 93-96

$$\rho(x,y) = \begin{cases} 1 & (x,y) = (0,0) \\ \rho_{1,0} & (x,y) = (1,0) \\ \rho_{0,1} & (x,y) = (0,1) \\ \rho_{1,1} & (x,y) = (1,1) \\ \rho_{2,2} & x = y, x \geq 2 \\ \rho_{2,1} & x - y \geq 1, x \geq 2 \\ \rho_{1,2} & y - x \geq 1, x \geq 2 \end{cases}$$

Match state	*Home team parameters*	*Away team parameters*
(1, 0)	$\hat{\lambda}_{10} = 0.86\ (0.05)$	$\hat{\mu}_{10} = 1.33\ (0.09)$
(0, 1)	$\hat{\lambda}_{01} = 1.10\ (0.08)$	$\hat{\mu}_{01} = 1.07\ (0.08)$
(x, y), $x + y > 1$ and $x - y \geqslant 1$	$\hat{\lambda}_{21} = 1.01\ (0.06)$	$\hat{\mu}_{21} = 1.53\ (0.11)$
(x, y), $x + y > 1$ and $x - y \leqslant -1$	$\hat{\lambda}_{12} = 1.13\ (0.10)$	$\hat{\mu}_{12} = 1.16\ (0.11)$
Time variation	$\hat{\xi}_1 = 0.67\ (0.08)$	$\hat{\xi}_2 = 0.47\ (0.07)$

Models for In-play Prediction

- Dixon and Robinson (98): Full-game simulation

Score	*Model VI probability*	*Approximate confidence interval*	*Observed*
(0, 0)	0.09	(0.08, 0.11)	0.09
(1, 0)	0.12	(0.11, 0.14)	0.13
(0, 1)	0.08	(0.07, 0.09)	0.07
(1, 1)	0.13	(0.12, 0.14)	0.13
(x, 0)	0.13	(0.11, 0.16)	0.16
(x, 1)	0.16	(0.13, 0.17)	0.15
(0, y)	0.06	(0.05, 0.07)	0.07
(1, y)	0.09	(0.08, 0.10)	0.09
(x, y)	0.14	(0.11, 0.16)	0.11

IMPROVING THE FIFA RANKING USING PRE-GAME PREDICTION

FIFA Ranking Overview

- Attempts to rank the teams from best to worst
- Used for the determination of the group seeds for the World Cup final draw

Ranking pre-Brasil 2014 World Cup

Rnk	Team	Oct 2013 Pts	+/-	Pos
1	Spain	1513	0	
2	Germany	1311	1	
3	Argentina	1266	-1	
4	Colombia	1178	1	
5	Belgium	1175	1	
6	Uruguay	1164	1	
7	Switzerland	1138	7	
8	Netherlands	1136	1	
8	Italy	1136	-4	
10	England	1080	7	

These 7 teams (plus the host team Brazil) were the group seeds

FIFA Ranking Overview

Last Updated 14 Aug 2014 Next Release 18 Sep 2014

Rnk	Team	Aug 2014 Pts	+/-	Pos
1	Germany	1736	0	
2	Argentina	1604	0	
3	Netherlands	1507	0	
4	Colombia	1495	0	
5	Belgium	1407	0	
6	Uruguay	1316	0	
7	Spain	1241	1	▲
7	Brazil	1241	0	
9	Switzerland	1218	0	
10	France	1212	0	
11	Portugal	1152	0	

info

ARG

Points 1604.26

Prev.Pts 1605.53

POINTS LAST MONTH

M * I * T * C = P

POINTS OUTSIDE RANKING CALCULATION

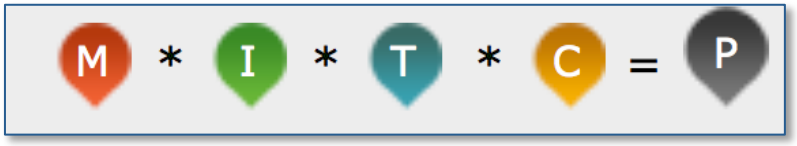
M * I * T * C = P

IRL - ARG	0:1	3	1	164	1	492

POINTS IN THE PAST 4 YEARS

	AVERAGE	WEIGHT	AVG. WEIGHT
2011	383.75	20%	76.75
2012	577.13	30%	173.14
2013	650.94	50%	325.47
2014	1028.9	100%	1028.9
			TOTAL 1604

FIFA Ranking



- **M: Match Result**
  - Win: 3
  - Tie: 1
  - Loss: 0
  - Win/loss by penalties: 2/1
- **I: Importance of Match**
  - Friendly match: 1
  - World Cup qualifier: 2.5
  - Confederation competition: 3
  - World Cup final competition: 4
- **T: Strength of Opposing Team**
  - (200 – ranking of opp.) $_{\vee}$50
- **C: Strength of Confederation:**

  (Computed using inter-confederation matches in last three World Cups).
  - Conmebol and Uefa: 1
  - Concacaf: 0.88
  - OFC: 0.85
  - AFC and CAF: 0.86

FIFA Ranking – Example #1

	ITALY	VS	ARGENTINA
Date		14-08-2013	
Type of match		Friendly	
Position in ranking	6		4
M: Match Result	0		3
I: Importance of Match	1		1
T: Strength of Opposing Team	196		194
C: Strength of Confederation	1		1
P=M I T C	0		582

FIFA Ranking – Example #2

	ARGENTINA	VS	IRAN
Date		21-06-2014	
Type of match		World Cup	
Position in ranking	5		43
M: Match Result	3		0
I: Importance of Match	4		4
T: Strength of Opposing Team	157		195
C: Strength of Confederation	0.86		1
P=M I T C	1620		0

FIFA Ranking Total Points Formula

- Computed as a weighted sum of single-match point averages obtained in each of the last 4 years

$$P_{\text{Total}} = \bar{P}_t + 0.5\bar{P}_{t-1} + 0.3\bar{P}_{t-2} + 0.2\bar{P}_{t-3}$$

- Average single-match points during year *t*
- Must play at least 5 games

- Total ranking points: order teams from highest to lowest
- Seven top teams seeded at the World Cup final draw

Flaws of the FIFA Ranking Formula

- D1) Friendly Matches
  - Friendly matches have low **Importance**, thus teams playing more international friendly matches are at a disadvantage
  - During 2013…

Colombia played **2** friendly matches
Belgium played **3** friendly matches

Netherlands played **4** friendly matches
Italy played **4** friendly matches
England played **4** friendly matches

Rnk	Team	Oct 2013 Pts	+/-	Pos
1	Spain	1513	0	
2	Germany	1311	1	
3	Argentina	1266	-1	
4	Colombia	1178	1	
5	Belgium	1175	1	
6	Uruguay	1164	1	
7	Switzerland	1138	7	
8	Netherlands	1136	1	
8	Italy	1136	-4	
10	England	1080	7	

A Few Examples…

- What if England played no friendlies in 2013...

Spain	Spain	Germany
Germany	Germany	Argentina
Argentina	Argentina	**Chile**
Colombia	Colombia	Colombia
Belgium	Belgium	Belgium
Uruguay	Uruguay	Uruguay
Switzerland	**England**	Germany

- If Chile had played no friendlies in 2013…

Other Disadvantages

- D2) Strong/Weak teams
  - A draw against the top-ranked team (Germany) earns fewer points than a win over the 100th-ranked team (Latvia)

- D3) Home/Away
  - Home and away wins earn the same number of points, even though a team's chances of winning at home are clearly better

- D4) Confederation Schedules
  - Different confederations have different schedules and frequencies for their conference level tournaments
    - Timing matters, numbers matters

Proposed Methods

- Ranking procedure as a simple modification of current one

FOLCKLORE vs. DATA

- Key idea: ideal ranking should reflect the average positions from double-round robin tournaments
- Implementation: pre-game prediction model + Monte Carlo

Reference Ranking

- Maher (82): independent Poisson, attack- defense, home/away effect
- Data: top 100 teams, 2009 October 2013

Orden	País	Prob. Campeón
1	Brasil	24,45%
2	España	20,49%
3	Argentina	15,33%
4	Holanda	12,49%
5	Alemania	5,3%
6	Inglaterra	5,26%
7	Francia	3,27%
8	Uruguay	3,0%
9	Ecuador	2,94%
10	Colombia	2,0%
11	Croacia	1,18%
12	Rusia	1,07%
13	Italia	1,04%
14	México	0,95%
15	Chile	0,19%

Proposal: exploratory study

- Logistic regression framework – outcome probabilties as functions of relevant factors
- Vector X summarizes team features $V_i = \beta^T X_i + e_i$

$$p_1 = P(Y = 1|X_1, X_{-1}) = \frac{exp(\beta^T X_1)}{1 + exp(\alpha + \beta^T X_1 + \beta^T X_{-1})}$$

$$p_0 = P(Y = 1|X_1, X_{-1}) = \frac{1}{1 + exp(\alpha + \beta^T X_1 + \beta^T X_{-1})}$$

$$p_{-1} = P(Y = 1|X_1, X_{-1}) = \frac{exp(\beta^T X_{-1})}{1 + exp(\alpha + \beta^T X_1 + \beta^T X_{-1})}$$

Proposal: exploratory study

- Finding relevant factors:

$$V_i = \beta_0 + \beta_{difstrong} * difstrong_i + \beta_{localia} * localia_i + \beta_{visita} * visita_i + \beta_{AFC} * AFC_i + \beta_{OFC} * OFC_i + \beta_{CAF} * CAF_i + \beta_{CONMEBOL} * CONMEBOL_i + \beta_{UEFA} * UEFA_i$$

  - Diference in ranking (FIFA)
  - Home/away factor
  - Home/away (continent) factor
  - Conference factor

Proposal: exploratory study

- Results

Tabla 5: Resultados modelos de regresión.

	MR		MRLn		MRLnLc		MRLnFc	
Variable	Coeficiente	Error	Coeficiente	Error	Coeficiente	Error	Coeficiente	Error
β_0	0,3071	0,0274	0,3019	0,0519	0,1478	0,0895	0,1478	0,0895
$difstrong$	0,0131	0,0003	0,014	0,0003	0,014	0,0003	0,014	0,0003
$Localia$			0,3463	0,0614	0,5005	0,0954	0,3601	0,0954
$Visita$			-0,5058	0,0645	-0,3517	0,0974	-0,467	0,0974
$Continente$					0,2049	0,0975		
AFC							0,0611	0,06
OFC							0,477	0,2012
$CONMEBOL$							0,3136	0,1021
$UEFA$							-0,1454	0,0631
CAF							0,1705	0,0633

Proposal: exploratory study

- Results:

$$\beta_{localia} - \beta_{visita} = 2 * x * \beta_{diffstrong}$$

$$0,3464 + 0,5058 = 2 * x * 0,014$$

$$x = 30,4$$

Tabla 6: Resumen Resultados Modelos propuestos.

Modelo	Log Verosimilitud	N° Param.	AIC	BIC
MR	-7.704	2	15.412	15.426
MRLn	-7.409	4	14.828	14.856
MRLnLc	-7.407	5	14.825	14.860
MRLnFc	-7.391	9	14.800	14.862

Proposed Methods: Outcome

Tabla 7: Rankings obtenidos mediante las cuatro metodologías de puntaje.

	Metodología 1	Metodología 2	Metodología 3	Metodología 4
1	España	Brasil	Brasil	Brasil
2	Argentina	España	España	España
3	Alemania	Argentina	Argentina	Argentina
4	Uruguay	Alemania	Alemania	Alemania
5	Colombia	Estados Unidos	Chile	Uruguay
6	Costa de Marfil	Chile	Italia	Holanda
7	Italia	Holanda	Uruguay	Chile
8	Estados Unidos	Uruguay	Estados Unidos	Costa de Marfil
9	Brasil	Italia	Holanda	Italia
10	Bélgica	Costa de Marfil	Colombia	Japón
11	Holanda	Colombia	Costa de Marfil	Estados Unidos
12	Ghana	Bélgica	Bélgica	Inglaterra
13	Chile	Ghana	Inglaterra	Colombia
14	México	Japón	Ghana	Corea del Sur
15	Suiza	México	Francia	Grecia
16	Inglaterra	Francia	Japón	México

Application to FIFA WC Draw

Pot 1	
1	Brazil
2	Argentina
3	Colombia
4	Uruguay
5	Belgium
6	Germany
7	Spain
8	Switzerland

Pot 2	
1	Algeria
2	Cameroon
3	Côte d'Ivoire
4	Ghana
5	Nigeria
6	Chile
7	Ecuador

Pot 3	
1	Australia
2	Iran
3	Japan
4	Korea Republic
5	Costa Rica
6	Honduras
7	Mexico
8	USA

Pot 4	
1	Bosnia-Herzeg.
2	Croatia
3	England
4	France
5	Greece
6	Italy
7	Netherlands
8	Portugal
9	Russia

GROUP A	GROUP B	GROUP C	GROUP D
Brazil	Spain	Colombia	Uruguay
Croatia	Netherlands	Greece	Costa Rica
Mexico	Chile	Côte d'Ivoire	England
Cameroon	Australia	Japan	Italy

GROUP E	GROUP F	GROUP G	GROUP H
Switzerland	Argentina	Germany	Belgium
Ecuador	Bosnia-Herzegovina	Portugal	Algeria
France	Iran	Ghana	Russia
Honduras	Nigeria	USA	Korea Republic

2014 FIFA World Cup Final Draw

	Pot 1		
11	1	Brazil	
3	2	Argentina	
4	3	Colombia	
6	4	Uruguay	
5	5	Belgium	
2	6	Germany	
1	7	Spain	
7	8	Switzerland	

	Pot 2		
32	1	Algeria	
59	2	Cameroon	
17	3	Côte d'Ivoire	
23	4	Ghana	
33	5	Nigeria	
12	6	Chile	
22	7	Ecuador	

	Pot 3		
57	1	Australia	
49	2	Iran	
44	3	Japan	
56	4	Korea Republic	
31	5	Costa Rica	
34	6	Honduras	
24	7	Mexico	
13	8	USA	

	Pot 4		
16	1	Bosnia-Herzeg.	
18	2	Croatia	
10	3	England	
21	4	France	
15	5	Greece	
8	6	Italy	
8	7	Netherlands	
14	8	Portugal	
19	9	Russia	

GROUP A	GROUP B	GROUP C	GROUP D	GROUP E	GROUP F	GROUP G	GROUP H
Brazil	Spain	Colombia	Uruguay	Switzerland	Argentina	Germany	Belgium
Croatia	Netherlands	Greece	Costa Rica	Ecuador	Bosnia-Herzegovina	Portugal	Algeria
Mexico	Chile	Côte d'Ivoire	England	France	Iran	Ghana	Russia
Cameroon	Australia	Japan	Italy	Honduras	Nigeria	USA	Korea Republic
112	78	80	56	84	101	52	112

2014 FIFA World Cup Final Draw

- Draw procedure questionable
  - Based almost exclusively on georafical considerations (which by the way might have been violated w.p. 12%)
  - Currently and historically strong teams are relegated to difficult groups
  - Consider FIFA ranking only for seeding purposes, not for setting pot composition
- This realization: unbalanced groups
  - One group contained 3 past World Cup Champions
  - Another group contained the 2 finalist of the 2010 World Cup

Proposed Draw Allocation

Step 1

- Takes a given ranking as a base. For example, consider one based on the FIFA ranking (50%), participation in previous World Cups (37.5%) and confederation cups (12.5%)
- Eight seeded teams: Brazil, Germany, Argentina, Spain, Italy, Uruguay, Netherlands and England
- Other teams assigned to pots according to their score/ranking

Step 2

- Assign teams to groups so as to minimize some notion of fairness

Step 1: Pots' Composition

Pot 1	Pot 2	Pot 3	Pot 4
Brazil (0.6367)	United States (0.5235)	Russia (0.4895)	Japan (0.4511)
Germany (0.6373)	France (0.5214)	Croatia (0.4884)	Australia (0.4427)
Argentina (0.5899)	Belgium (0.5104)	Ghana (0.4771)	Ecuador (0.4335)
Spain (0.5869)	Mexico (0.5043)	Bosnia (0.4768)	Iran (0.4377)
Italy (0.5833)	Colombia (0.5007)	Nigeria (0.4699)	Costa Rica (0.4335)
Uruguay (0.5593)	Portugal (0.5007)	Ivory Coast (0.4683)	Algeria (0.4315)
Netherlands (0.5470)	Chile (0.4980)	Greece (0.4633)	Cameroon (0.4292)
England (0.5250)	Switzerland (0.4906)	South Korea (0.4575)	Honduras (0.4264)

Step 2: MIP Formulation

- Only constraint is that no group can have more than one team from the same confederation, except Europe, from which there can be 2 teams
- Let p_i denote the score associated with teams *i*
- Mixed Integer Programming formulation for deciding final draw

$$x_{i,g} = \begin{cases} 1 & \text{if team } i \text{ assigned to group } g \\ 0 & \sim \end{cases}$$

$$y_g = \text{score of group } g$$

Step 2: MIP Formulation

$$\min z_{\max} - z_{\min}$$ (Minimize score gap)

$$s.t. \quad \sum_g x_{i,g} = 1 \qquad \forall\, i \in \text{Teams}$$ (All teams assigned)

$$\sum_g p_i\, x_{i,g} = y_g \qquad \forall\, g \in \text{Groups}$$ (Compute score)

$$z_{\min} \le y_g \qquad \forall\, g \in \text{Groups}$$ (Compute min score)

$$z_{\max} \ge y_g \qquad \forall\, g \in \text{Groups}$$ (Compute max score)

$$\sum_{i \in R} x_{i,g} \le n_R \qquad \forall g \in \text{Groups},\ R \in \text{GeoCond}$$

(Geo considerations)

$$x_{i,g} \in \{0,1\},\ y_g \ge 0,\ z_{\min} \ge 0,\ z_{\max} \ge 0$$

Step 2: MIP Formulation

- Geographic considerations

R = {Conmebol, Concacaf, AFC, CAF} with n_R=1

R = {UEFA} with n_R=2

R ={Pot1,Pot2,Pot3,Pot4} with n_R=1

- Model with 10 continuous variables, 256 binary variables
- Optimal objective function 0.0324
- Multiple solutions (even after fixing the groups of the seeded teams)
- Solution time <1 seconds using CPLEX 12.6

Resulting Draws

	Alternative 1A							
	G1	G2	G3	G4	G5	G6	G7	G8
Teams	Brazil	Germany	Argentina	Spain	Italy	Uruguay	Netherlands	England
	Switzerland	Chile	Portugal	Mexico	Colombia	Belgium	France	United States
	Greece	South Korea	Nigeria	Ivory Coast	Bosnia	Croatia	Ghana	Russia
	Cameroon	Honduras	Iran	Ecuador	Costa Rica	Algeria	Australia	Japan
Group Score	2.0206	2.0191	1.9982	1.9989	1.9963	1.9896	1.9882	1.9891

	Alternative 1B							
	G1	G2	G3	G4	G5	G6	G7	G8
Teams	Brazil	Germany	Argentina	Spain	Italy	Uruguay	Netherlands	England
	Switzerland	Chile	Belgium	Colombia	Portugal	Mexico	France	United States
	Greece	South Korea	Bosnia	Ivory Coast	Nigeria	Croatia	Ghana	Russia
	Cameroon	Honduras	Algeria	Costa Rica	Ecuador	Iran	Australia	Japan
Group Score	2.0206	2.0191	2.0086	1.9894	1.9953	1.9897	1.9882	1.9891

	Alternative 2							
	G1	G2	G3	G4	G5	G6	G7	G8
Teams	Brazil	Germany	Argentina	Spain	Italy	Uruguay	Netherlands	England
	Switzerland	Russia	Bosnia	United States	Colombia	France	Belgium	Portugal
	Australia	Ecuador	Nigeria	South Korea	Croatia	Ivory Coast	Mexico	Chile
	Cameroon	Costa Rica	Greece	Algeria	Honduras	Japan	Iran	Ghana
Group Score	2.0000	1.9998	1.9999	1.9993	2.0007	2.0001	1.9994	2.0007

HAVING FUN DURING COPA AMERICA 2015-2016, 2018 WCQ AND EURO 2016

Copa América 2015

- Maher (82) – Independent Poisson Goals
- Inference – Last five years of history (MLE)
- 1e7 Simulated tournaments

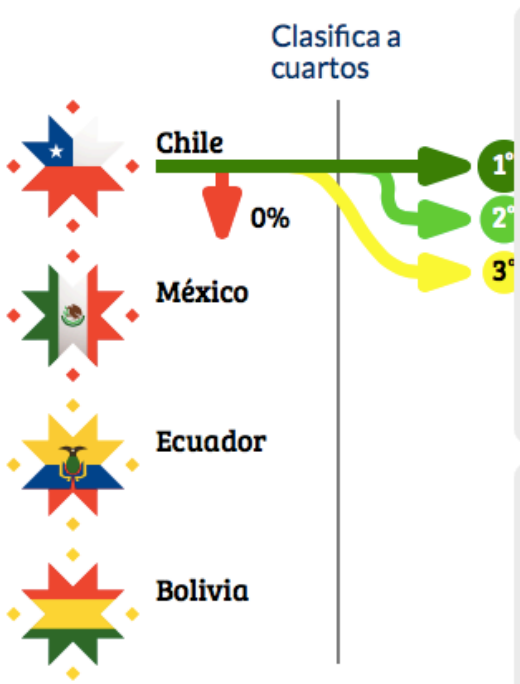
- Results announced and updated on CEINE webpage
- Multiple appearances in the press

Copa América 2015

Etapa Grupal

Posiciona tu mouse sobre la bandera de un país para conocer las probabilidades asociadas a cada uno de sus partidos en la fase de grupos, además de sus chances de clasificar a la segunda fase como primero de su grupo, segundo, o mejor tercero.

Grupo A



CHILE:

Calendario de Juegos

Oponente	Gana	Empata	Pierde	Resultado
Ecuador	✔			2-0
México		✔		3-3
Bolivia	✔			5-0

Clasificación a cuartos más probable:

1°: Chile (100%)

2°: Bolivia (100%)

Copa América 2015

Equipo	Grupo	Cuartos	Semifinal	Final	Campeón
Chile	A	100%	62%	71%	51%
Argentina	B	100%	48%	74%	49%
Bolivia	A	100%	29%	0%	0%
Brasil	C	100%	79%	0%	0%
Colombia	C	100%	52%	0%	0%
Ecuador	A	0%	0%	0%	0%
Jamaica	B	0%	0%	0%	0%
México	A	0%	0%	0%	0%
Paraguay	B	100%	21%	26%	0%
Perú	C	100%	71%	29%	0%
Uruguay	B	100%	38%	0%	0%
Venezuela	C	0%	0%	0%	0%

2018 WC Qualifiers: web page

Post 6ª fecha | Post 4ª fecha | Post 3ª fecha | Post 2ª fecha | Situación Inicial

Probabilidades después de la 6ª fecha.

País	Primero	Segundo	Tercero	Cuarto	Directo	Repechaje
Chile	37%	25%	17%	11%	90%	7%
Argentina	32%	25%	18%	13%	88%	8%
Brasil	11%	17%	20%	20%	68%	18%
Uruguay	11%	15%	19%	21%	66%	19%
Colombia	7%	13%	18%	21%	59%	22%
Ecuador	2%	4%	8%	13%	26%	21%
Paraguay	< 0.5%	< 0.5%	< 1%	2%	3%	5%
Perú	< 0.5%	< 0.5%	< 0.5%	< 0.5%	< 0.5%	< 1%
Bolivia	0%	< 0.5%	< 0.5%	< 0.5%	< 0.5%	< 0.5%
Venezuela	0%	0%	< 0.5%	< 0.5%	< 0.5%	< 0.5%

2018 WC Qualifiers today

Equipo	Primero	Segundo	Tercero	Cuarto	Quinto	DIRECTO
Argentina	2.6143	24.0092	23.4285	22.4264	20.2892	72.4784
Bolivia	0	0	0	0	0	0
Brazil	83.8061	10.6809	3.591	1.4167	0.4624	99.4947
Chile	3.5436	19.4509	23.7382	23.499	21.6748	70.2317
Colombia	1.4992	12.156	21.4012	26.7861	26.7355	61.8425
Ecuador	0.1067	1.1514	3.1497	6.9357	16.1123	11.3435
Praguay	0.0002	0.0053	0.0376	0.1763	0.7622	0.2194
Peru	0	0.0057	0.0374	0.1792	0.8509	0.2223
Uruguay	8.4299	32.5406	24.6164	18.5806	13.1127	84.1675
Venezuela	0	0	0	0	0	0

The Big Data Challenge – EURO 2016

- Scoring formula:

$$\text{Score} = \sum_{i=1}^{n} Score_i$$

$$Score_i = (1 - Prob_l)^2 \cdot I_l + (1 - Prob_e)^2 \cdot I_e + (1 - Prob_v)^2 \cdot I_v$$

- +1000 contestants, US$5000 prize

Posición	Nickname	Puntaje
1	dsml16	19.47246085
2	NachoCorreaF	19.76623484
3	kikox4	20.33860241
4	pablocxra	20.45953740
5	bizz8	20.50604150
6	David_D	20.57502213
7	ignaciacaamano	20.76623484

A Refined Prediction Model

- Nate Silver – Soccer Power Index (SPI), ESPN
- Attack/defense factor as sum of contribution from players
- Adjusted goals

A Refined Prediction Model

- Nate Silver – Soccer Power Index (SPI), ESPN
- Attack/defense factor as sum of contribution from players
- Adjusted goals