Luck is Hard to Beat: The difficulty of sports prediction

Luck is Hard to Beat: The difficulty of sports prediction

Raquel Aoki Renato Assuncao Pedro Vaz de Melo

Halifax, August 13, 2017

Department of Computer Science,
Universidade Federal de Minas Gerais, Brazil

1.

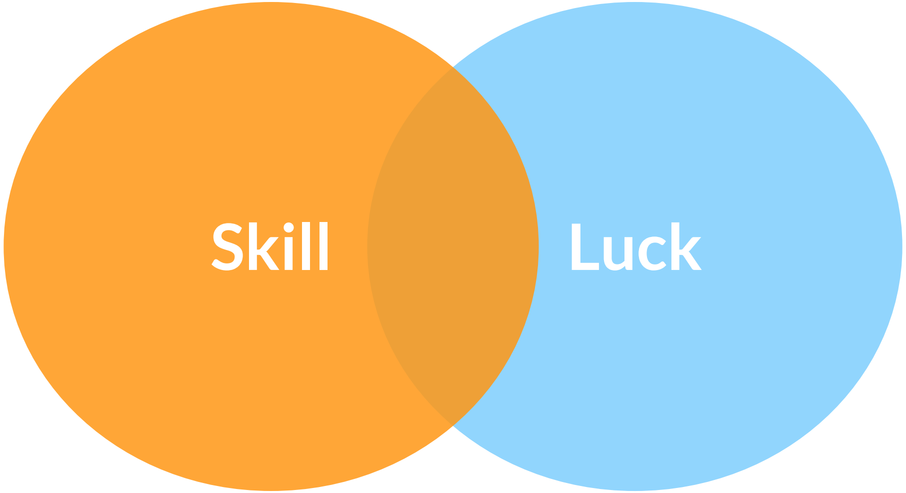
Introduction

Sports are **extremely** surprising

- Unpredictability cannot be avoided

Sports are **extremely** surprising

- Unpredictability cannot be avoided



Attractive area

- Relatively isolated systems
- Datasets available
- Popularity

Φ Coefficient: Contributions

- Skill influence

- Significance test

- Which teams should be removed to make a league random

Probabilistic Graphical Model: Contributions

- Teams' skill in a season/league

- Players and teams characteristics more influents

2.

Related Work

Sports Results Forecasting

Which team will score next or who will win the game?

Miljkovic et al. [2010]	Naive Bayes
Gabel & Redner[2012],	Random Walk
Merritt & Clauset[2014]	Random Walk
Chen & Joachims [2016]	Probabilistic Graphical Models

Sport Leagues Characterization

Factors which has influence on sport results

Vaz de Melo et al. [2012]	Network between basketball players
Wang et al. [2015]	Best soccer tactics
Van Haaren et al. [2016]	Spatial and time patterns in volleyball games
Ribeiro et al. [2016]	Advantage in home games

Sport Leagues Characterization

Teams' Skill

Tarlow et al. [2014]	Football teams' skill estimated by probabilistic graphical models
Chetrite et al. [2015]	Number of potential winners
Pelechrinis et al. [2016]	NBA teams' ranking using PageRank
Spiegelhalter [2007]	Luck in soccer leagues

3.

Disentangling luck and skill

A luck and skill coefficient

- Random leagues

A luck and skill coefficient

- Baseline leagues
  - Teams with the same number of games
  - Same number of games as home teams
  - The result is the points accumulated sum

A luck and skill coefficient

- Baseline leagues examples

SUPERLIGA LFH BRASILEIRÃO NBA

A luck and skill coefficient

- X_h: random variable representing the points earned by a team when it plays in a match at home
- X_a: random variable representing the points earned by a team when it plays away

X_a and X_h are specific for each sport.

A luck and skill coefficient

- It depends on the possible results in the sport

Basketball	Points	
	Home	Away
P_h	1	0
P_a	0	1

A luck and skill coefficient

- It depends on the possible results in the sport

Soccer	Points	
	Home	Away
P_h	3	0
P_t	1	1
P_a	0	3

A luck and skill coefficient

- It depends on the possible results in the sport
- Probabilities non-negative
- Socce example: $P_h + P_t + P_a = 1$
- Those probabilities are specifics for each sport/league/season.

A luck and skill coefficient

- X_a and X_h have multinomial distribution

$$\mu_{X_h} = P_h \times 3 + P_t \times 1 + P_a \times 0$$

$$\sigma^2_{X_h} = P_h \times 3^2 + P_t \times 1^2 + P_a \times 0^2 - \mu^2_{X_h}$$

A luck and skill coefficient

- X_a and X_h have multinomial distribution

$$\mu_{X_h} = P_h \times 3 + P_t \times 1 + P_a \times 0$$

$$\sigma^2_{X_h} = P_h \times 3^2 + P_t \times 1^2 + P_a \times 0^2 - \mu^2_{X_h}$$

Similarly,

$$\mu_{X_a} = P_a \times 3 + P_t \times 1 + P_h \times 0$$

$$\sigma^2_{X_a} = P_a \times 3^2 + P_t \times 1^2 + P_h \times 0^2 - \mu^2_{X_a}$$

A luck and skill coefficient

- Y_{2k} is a random model

$$Y_{2k} = \sum_i (X_{hi} + X_{ai})$$

It represents the cumulative points after 2k games in a league

A luck and skill coefficient

For *k* large and independent games:

- ▷ $Y_{2k} \sim N(\mu_{2k}, \sigma^2_{2k})$ (Unique for each season)
- ▷ $\mu_{2k} = k(\mu_{X_h} + \mu_{X_a})$
- ▷ $\sigma^2_{2k} = k(\sigma^2_{X_h} + \sigma^2_{X_a})$ (Theoretical variance)

A luck and skill coefficient

Comparison between the theoretical variance and the sample variance s^2

- For one season/league: Valid
- For more than one season or different leagues and sports: Not valid

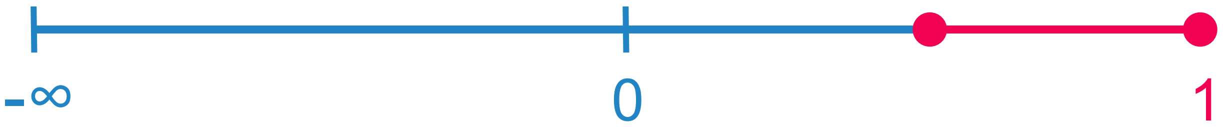
A luck and skill coefficient

Solution: φ Coefficient

$$\phi = \frac{S^2 - \sigma_{2k}^2}{S^2}$$

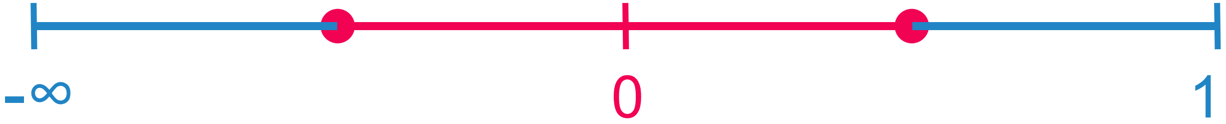
- Defined in (-∞,1]
- Distance between the random model variance and the observed variance

A luck and skill coefficient



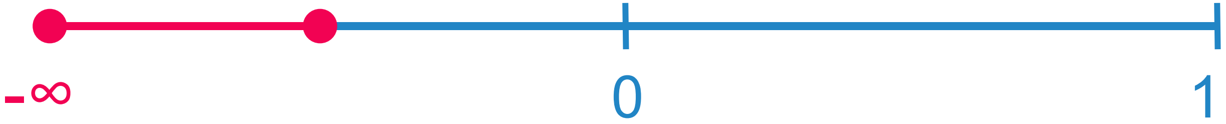
- Close to 1 : skill factor has more influence

A luck and skill coefficient



- Close to 0: skill factor has a small influence in the results

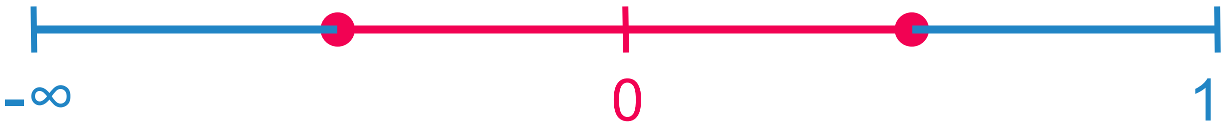
A luck and skill coefficient



- Negative values: observed variance is less than the expected value in the random model

A luck and skill coefficient

- Confidence Interval(95%) around 0;



- Monte Carlo simulation considering true the random model hypothesis;

- Unique confidence interval

A luck and skill coefficient

If φ is different of 0, how many teams should be removed from a league in order to turn it random?

- Simulation

Skill

If the skill component has influence in a league, it is possible to estimate the teams' skill?

Skill estimation

- Probabilistic Graphical Model
- Seasons/Leagues with φ significantly different of 0 and positive
- Skill component decomposed into explanatory features

Skill estimation

Skill of *n* teams

- $\alpha_1, \alpha_2, \ldots, \alpha_n$

Probability of i beat j

- $\pi_{ij} = \frac{\alpha_i}{\alpha_i + \alpha_j}$

The α values are calculated as:

- $\alpha_i = exp(\mathbf{w}^T x_i)$

Skill estimation

- x_i : vector of *d* dimension composed with i team features
    - Explains the skill associate to α_i
- The model learns the relevance/weight (**w**) of each feature to find the skill

Skill estimation

Likelihood function:

- Games final score

- *K* games and *n* teams
  - N_k: points sum of each team
  - S_k : point of the home team

- N_k is a sequence of success and failures
- S_k~ Binomial Distribution conditioned in N_k

Skill estimation

Likelihood function:

- N_k could be very large (basketball)
  - Computationally intractable

- Binomial distribution -> Poisson distribution

$$S_k \sim \text{Poisson}\left(N_k \times \frac{\alpha_{h(k)}}{\alpha_{h(k)} + \alpha_{a(k)}} + \varepsilon_k\right)$$

- Random effect ε_k in each game

Skill estimation

Likelihood function:

- Log-linear model $\log(\alpha_i) = \mathbf{w}^T \mathbf{x}_i$
  - **w**: weight features
  - **x**: set of features of team i:

$$\mathcal{D} = \{\mathbf{x}, S_k, N_k, \forall k = 1, ..., K\}$$

$$L(\mathcal{D}|\mathbf{w}, \varepsilon_k) \approx \prod_{k=1}^{K} \left(N_k \frac{\alpha_{h(k)}}{\alpha_{h(k)} + \alpha_{a(k)}} + \varepsilon_k \right)^{S_k} \times \exp\left(-N_k \frac{\alpha_{h(k)}}{\alpha_{h(k)} + \alpha_{a(k)}} + \varepsilon_k \right)$$

Skill estimation

Priori distribution of **w** weight

- $p(\mathbf{w}) \sim \mathrm{Normal}(\mathbf{w};\mathbf{0},2\mathbf{I})$

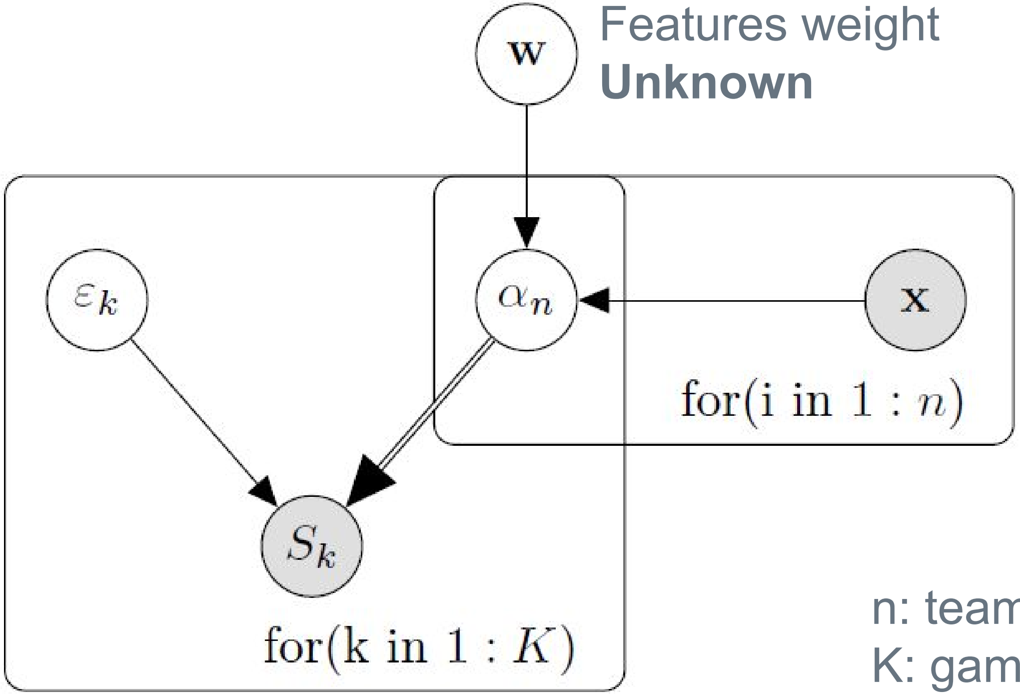
Priori distribution of random effect

- $p(\boldsymbol{\varepsilon}) \sim \mathrm{Normal}(\boldsymbol{\varepsilon};\mathbf{0},3\mathbf{I})$

Skill estimation

Features weight
Unknown

Random Effect
Luck
Unknown



Teams
Features
Known

Score Home Team

n: teams amount
K: games amount

Skill estimation

Posteriori Distribution

- Proportional to product between the Likelihood function and the *Prior* distribution

$$
\begin{aligned}
p(\mathbf{w}, \varepsilon | \mathcal{D}) \propto \textstyle\prod_{k=1}^{K} & \left(N_k \frac{\alpha_{h(k)}}{\alpha_{h(k)} + \alpha_{a(k)}} + \varepsilon_k \right)^{S_k} \\
& \times \exp\left(-N_k \frac{\alpha_{h(k)}}{\alpha_{h(k)} + \alpha_{a(k)}} + \varepsilon_k \right) \exp\left(-\frac{1}{2\sigma_w^2} \mathbf{w}^T \mathbf{w} \right) \\
& \times \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \varepsilon^T \varepsilon \right)
\end{aligned}
$$

Skill estimation

Posteriori Distribution

- Metropolis-Hastings Algorithm (MCMC)
    - *Posteriori* distribution Random samples
    - Inference about the parameters

Skill estimation

Models selection

- Features combination
- *Deviance Information Criterion* (DIC)
  - Model complexity
  - Data fit

4.1
Results

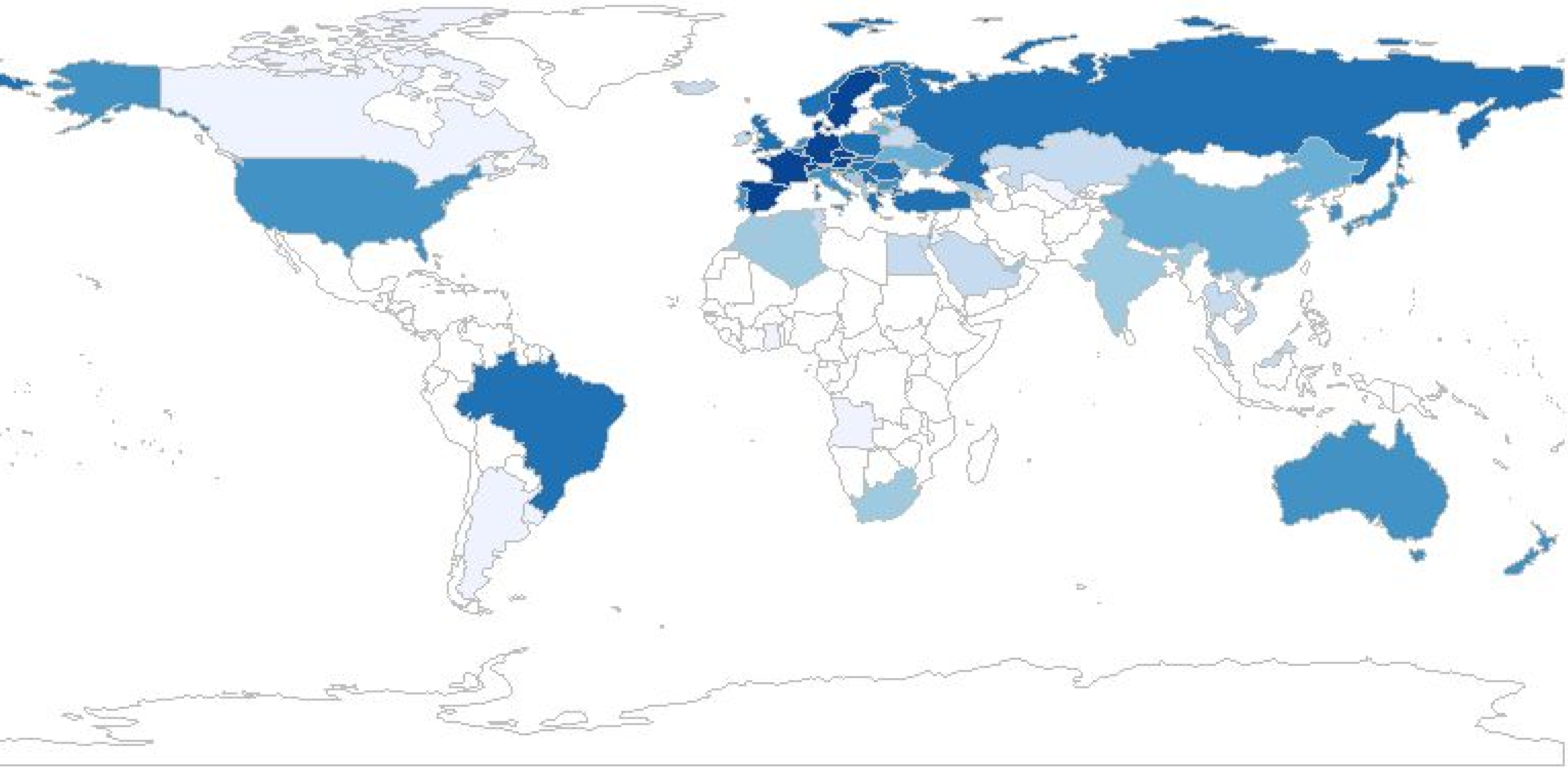
Φ Coefficient

270713 games

1503 seasons

84 countries

4 sports

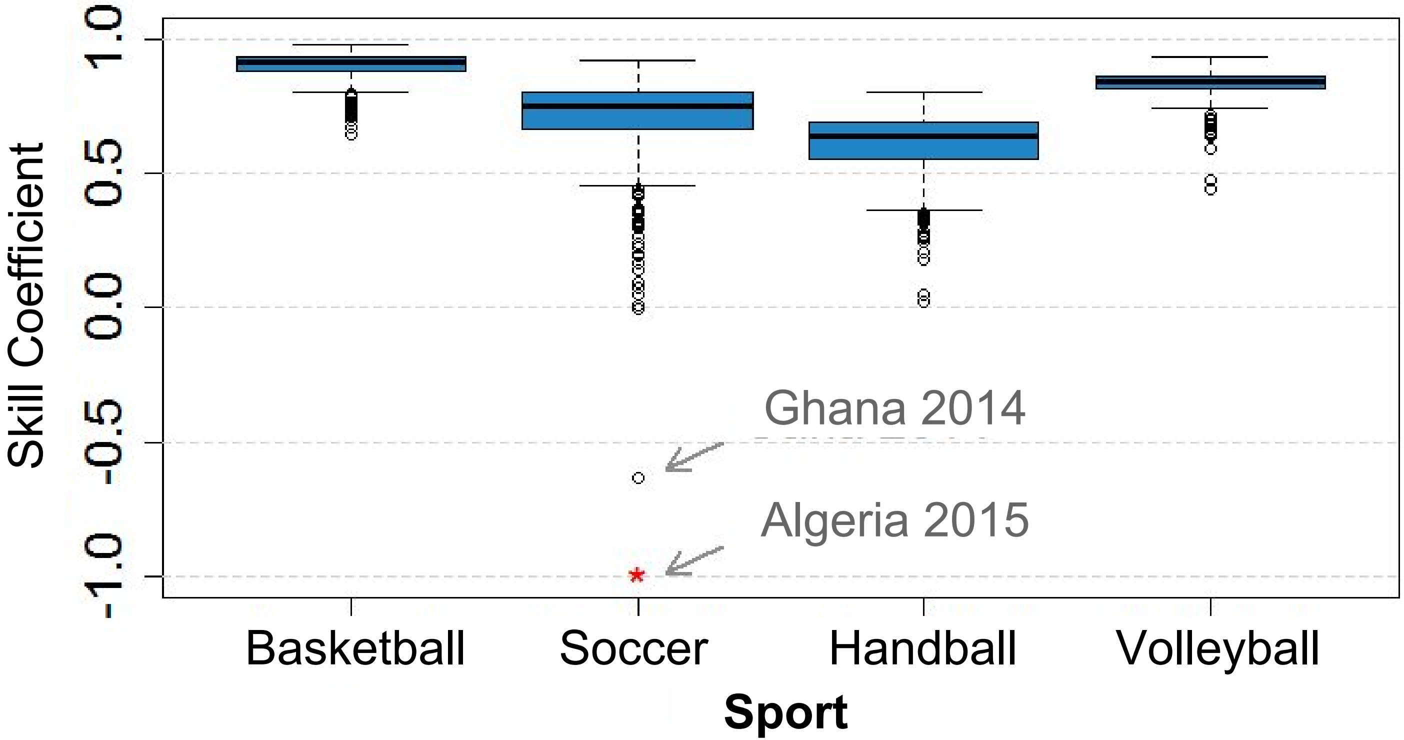


Datasets

Leagues between 2007 and 2016

Sport	Leagues		Seasons	
	Count	%	Count	%
Handball	25	12.63%	234	15.57%
Basketball	42	21.21%	310	20.63%
Volleyball	51	25.76%	328	21.82%
Soccer	80	40.40%	631	41.98%
Sum	198	100%	1503	100%

Results



Skill Coefficient
1.0
0.5
0.0
-0.5
-1.0
Ghana 2014
Algeria 2015
Basketball
Soccer
Handball
Volleyball
Sport

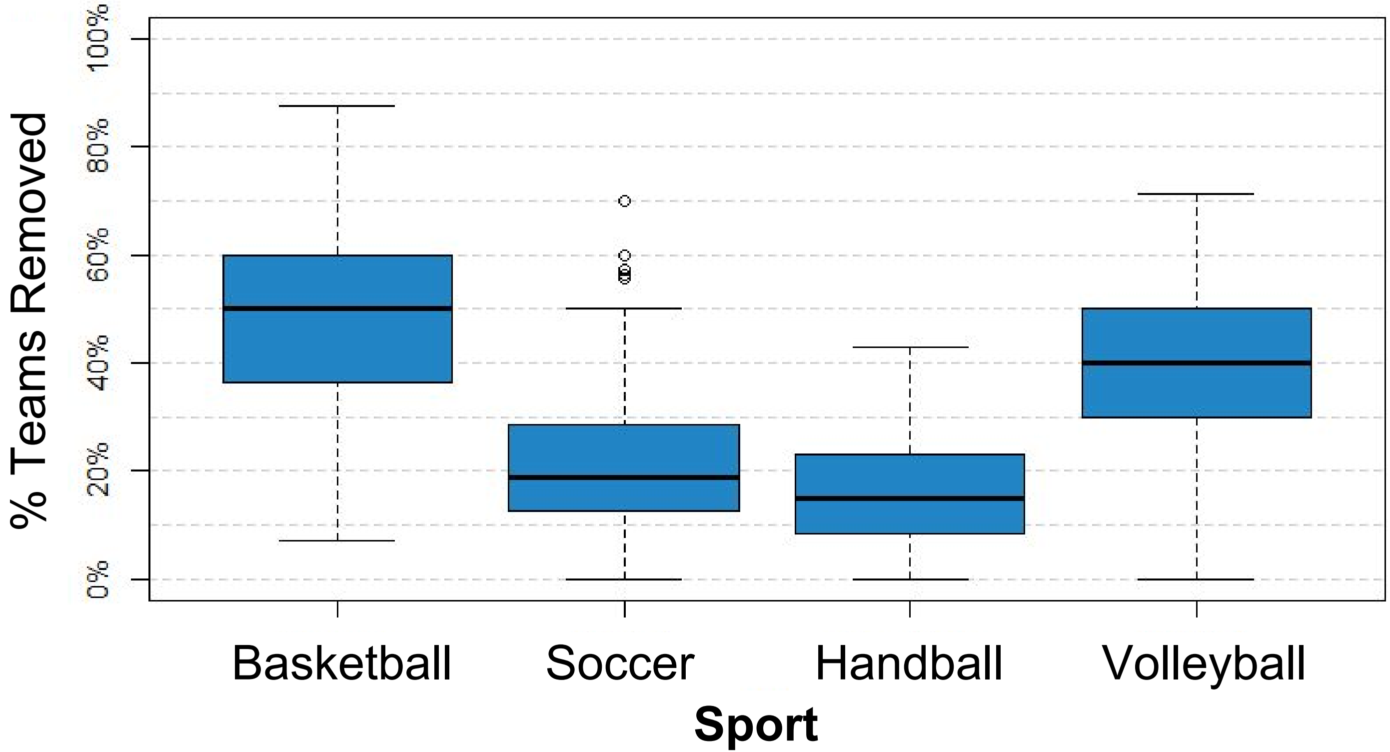
Results

- Basketball and volleyball
  - Many points in a single match
  - It is harder to a less skilled team win just by luck
- Soccer and handball
  - Few relevant events in a match
  - A soccer game has an average of 2.62 goals

Results

Sport	Skill		Luck		Sum
	Count	%	Count	%	
Basketball	310	100%	0	0%	310
Soccer	586	92.87%	45	7.13%	631
Handball	192	82.05%	42	17.75%	234
Volleyball	326	99.39%	2	0.61%	328
Sum	1414	94.08%	89	5.92%	1503

Results



% Teams Removed
0%
20%
40%
60%
80%
100%
Basketball
Soccer
Handball
Volleyball
Sport

Results

- *Primera División*, Spain
  - 3.2(16%) teams removed by season
  - Real Madri (10) and Barcelona (9)

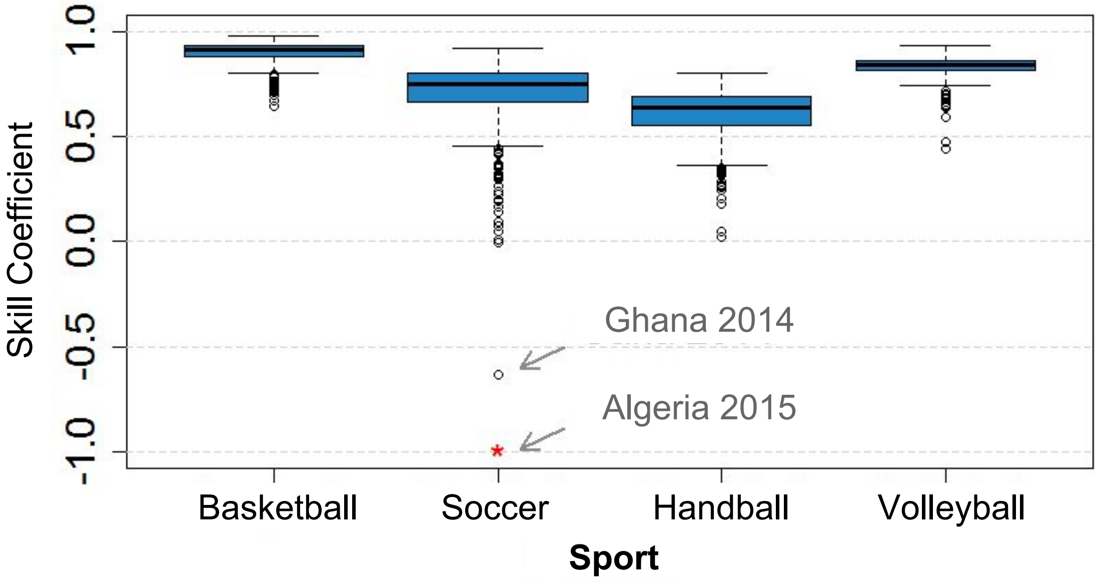
- *Premier League*, England
  - 4.9(25%) teams removed by season
  - Manchester United(9)

- NBA, United States of America
  - 17-25(+50%) teams removed by season

Results

Algeria, *Division 1*

- - 1.93
  season 2015
- Albert Bodjongo incident



Results

Mail Online

Home | News | U.S. | Sport | TV&Showbiz | Australia | Femail | Health | Science | Money | Vi

Football | Transfer News | Premier League | Champions League | Boxing | UFC | F1 | Tennis | Rugby | Cricket |

Algerian League is so tight all 16 teams can mathematically still win the title with four rounds of matches to go

- 11 points split leaders Setif and bottom-placed Hussein Dey with four rounds of the Algerian League to play
- All 16 teams could mathematically still win the championship
- Top two qualify for CAF Champions League while three are relegated
- Incredibly tense final day is expected on June 12

By ADAM SHERGOLD

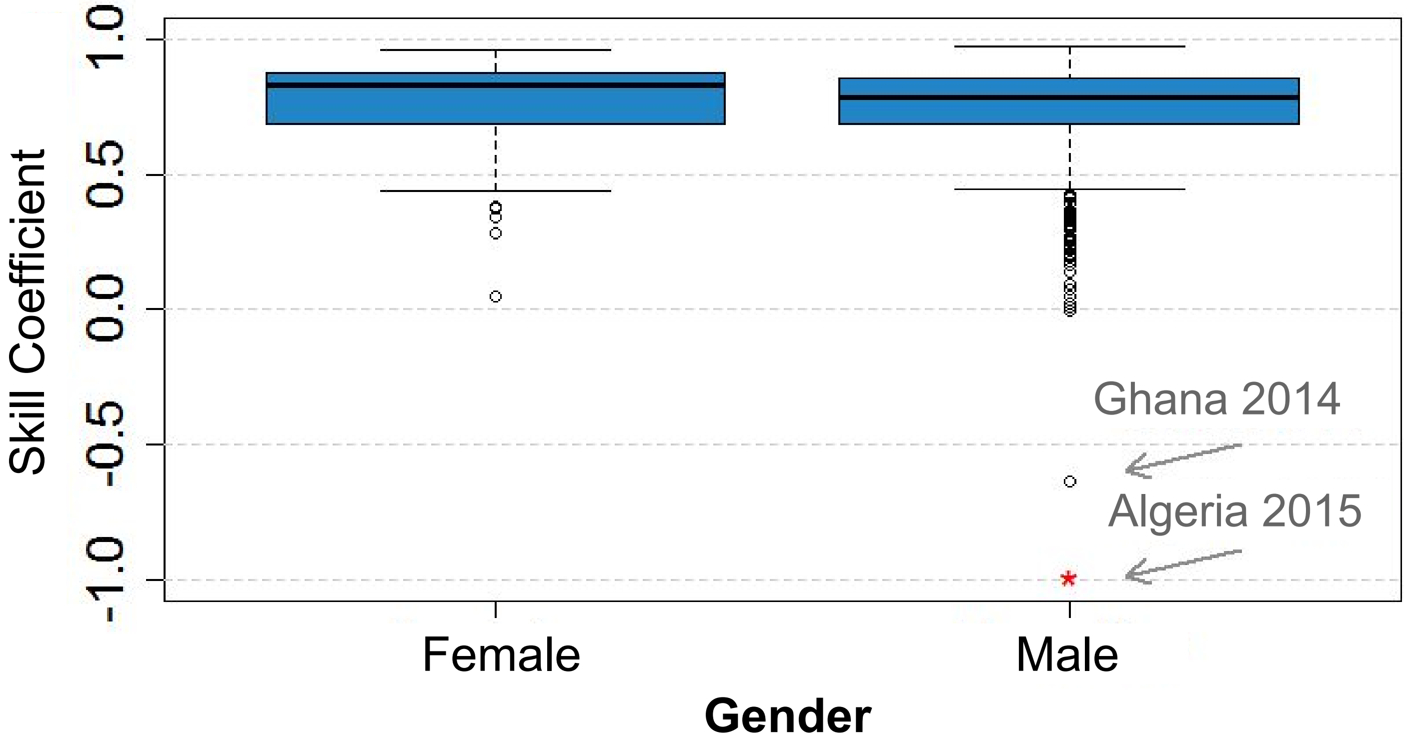
PUBLISHED: 10:44 GMT, 27 April 2015 | **UPDATED:** 10:44 GMT, 27 April 2015

Share

211 shares

21 View comments

Results



Skill Coefficient
1.0
0.5
0.0
-0.5
-1.0
Ghana 2014
Algeria 2015
Female
Male
Gender

4.2
Results

Skill

Dados

League

- ▷ *National Basketball Association* (NBA)
  - ○ Datasets available
  - ○ Skill component large(0.97)
- ▷ Theory can be extend to other leagues and sports

Dados

- A model for each season (2012-2016)

1. **Regular Season: round-robin system (≈ 1230 matches)**

2. Playoffs: 16 teams

Datasets

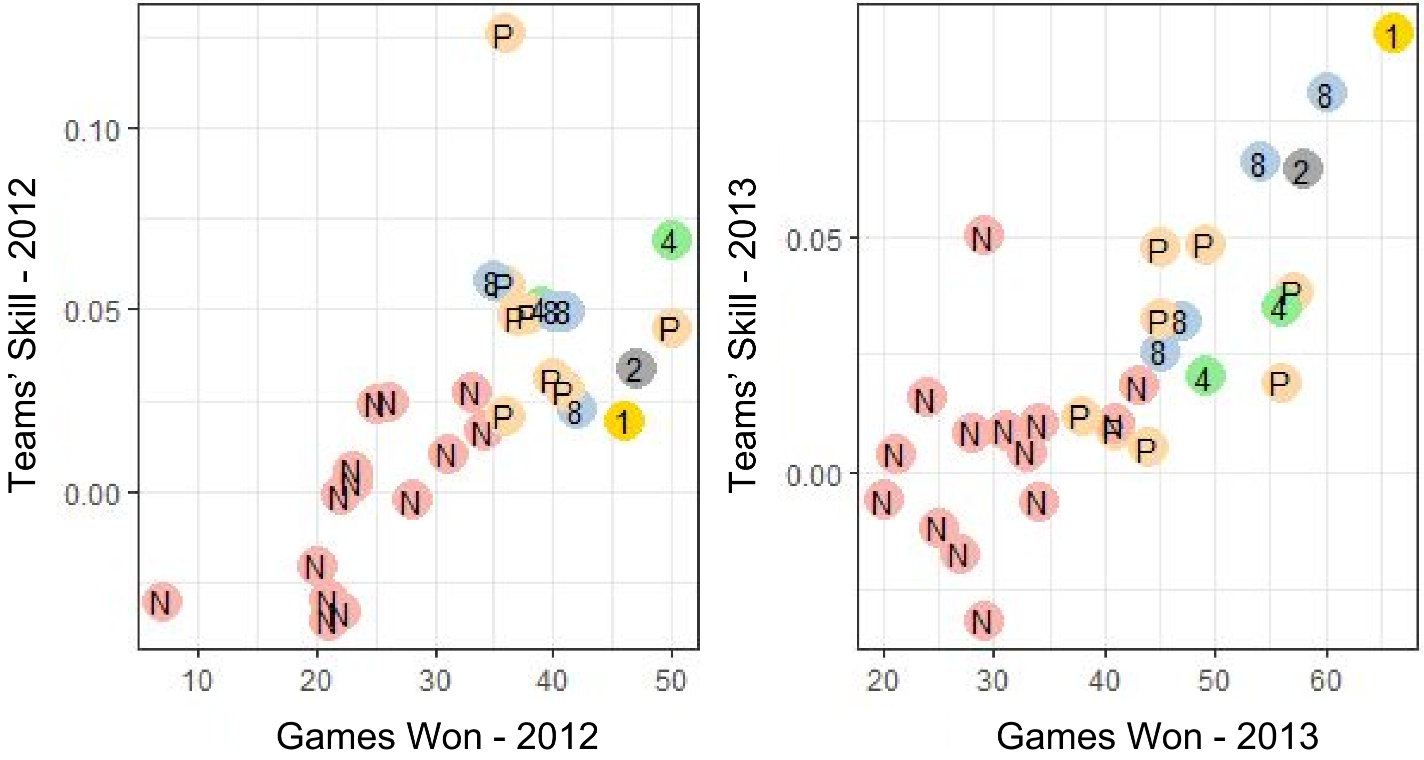
CO	East conference or West conference
A5	Average salary of the top 5 players salaries
A6	Average salary of the 6 to 10 top salaries
SD	Standard deviation for the players salaries
AP	Average Player Efficiency Rating(PER)
VL	Team Volatility, measuring how much a team change its players between seasons
RV	Roster Aggregate Volatility, measuring how much the players have been transferred from other teams in the past years
CC	Team Inexperience, the graph clustering coefficient.
RC	Roster Aggregate Coherence, measuring the strength of the relationship among the players
SI	Roster Size, the number of players

Results

Model selection

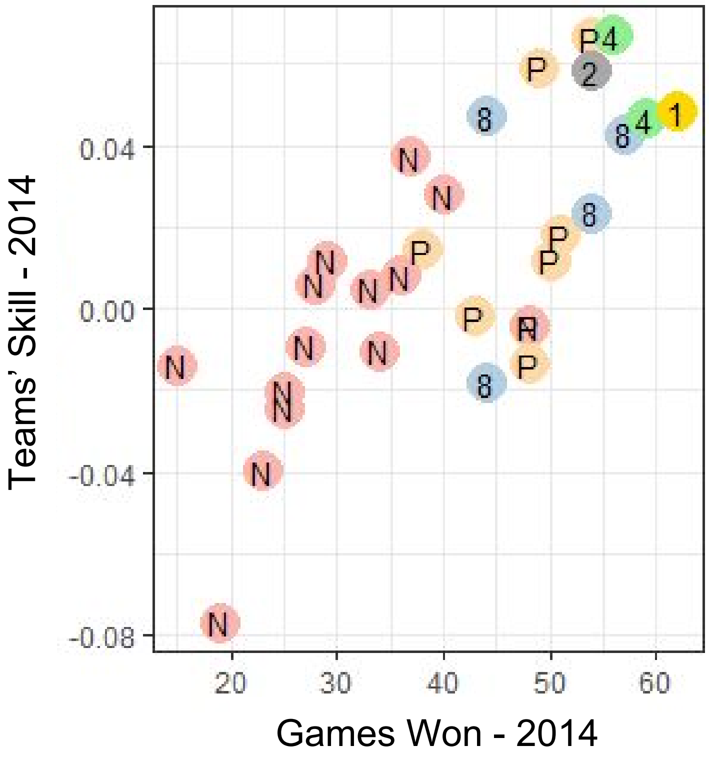
Model	Average DIC
1 - CO+A5+AP+VL+RC+SI	**-851727,84**
2 - CO+A5+A6+AP+SD+VL+RC+SI	-848333,74
3 - CO+A5+AP+SD+VL+RC+SI	-851158,02
4 - CO+A5+A6+AP+VL+RV+CC+RC+SI	-841589,70

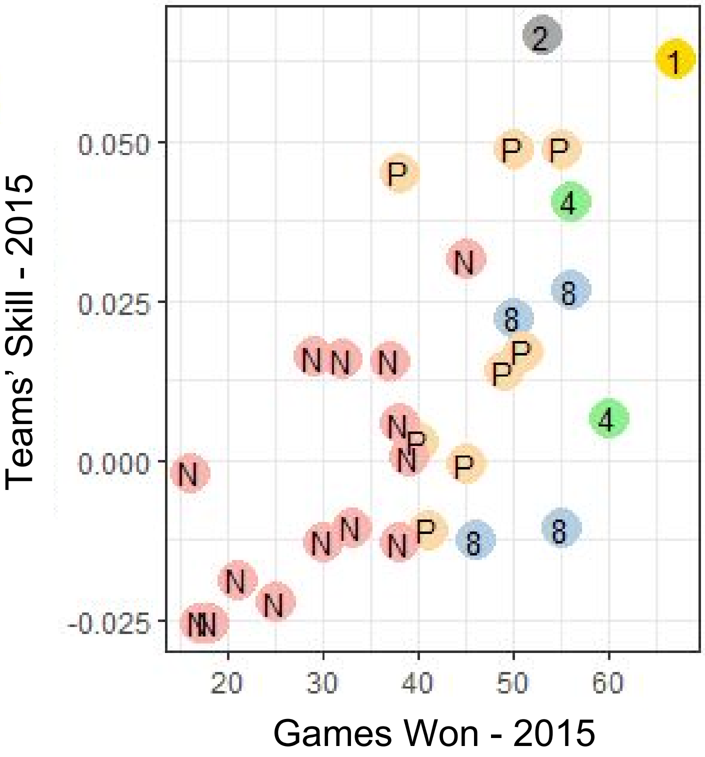
Results



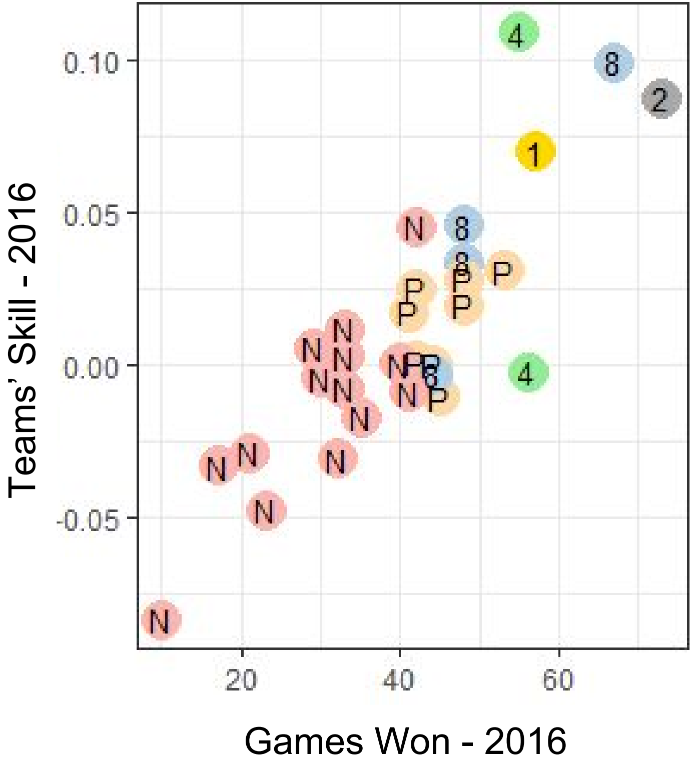
Teams' Skill - 2012
0.10
0.05
0.00
10
20
30
40
50
Games Won - 2012
Teams' Skill - 2013
0.05
0.00
20
30
40
50
60
Games Won - 2013

Results

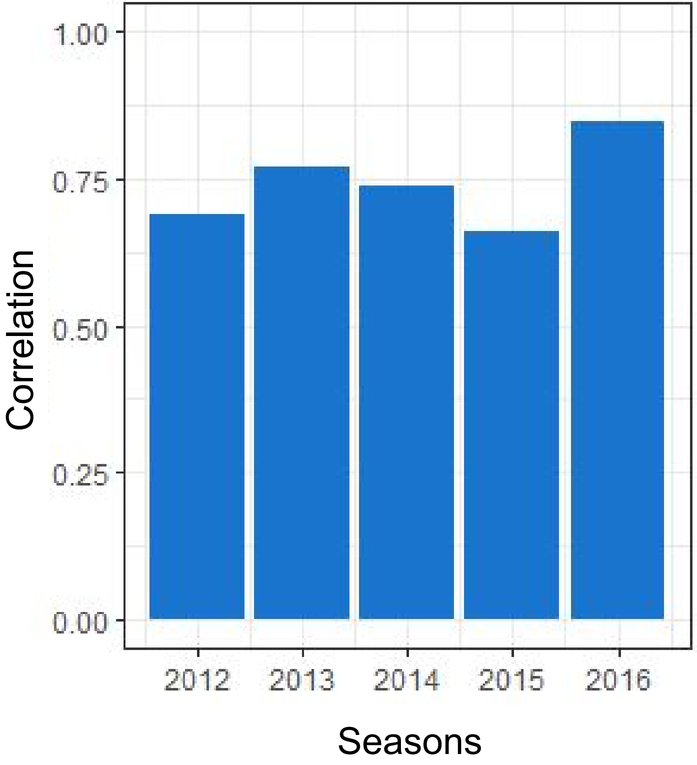

Teams' Skill - 2014
Games Won - 2014
0.04
0.00
-0.04
-0.08
20
30
40
50
60


Teams' Skill - 2015
Games Won - 2015
0.050
0.025
0.000
-0.025
20
30
40
50
60

Results



Teams' Skill - 2016
0.10
0.05
0.00
-0.05
20
40
60
Games Won - 2016



Correlation
1.00
0.75
0.50
0.25
0.00
2012
2013
2014
2015
2016
Seasons

Results

Season	P(U)	P(U|A)	P(U|H)	P(U|A,R+)	P(U|H,R+)
2012	0.35	0.27	0.44	0.25	0.24
2013	0.36	0.25	0.47	0.16	0.13
2014	0.37	0.29	0.45	0.15	0.14
2015	0.37	0.30	0.44	0.22	0.20
2016	0.34	0.26	0.43	0.19	0.13
Average	0.36	0.27	0.45	0.19	0.17

Results

Season	P(U)	P(U|A)	P(U|H)	P(U|A,R+)	P(U|H,R+)
2012	0.35	0.27	0.44	0.25	0.24
2013	0.36	0.25	0.47	0.16	0.13
2014	0.37	0.29	0.45	0.15	0.14
2015	0.37	0.30	0.44	0.22	0.20
2016	0.34	0.26	0.43	0.19	0.13
Average	0.36	0.27	0.45	0.19	0.17

5.

Conclusion

Conclusion

Goals

- Evaluate the luck and skill influence in sport leagues
- Estimate the teams’ skill

Contributions

- Φ coefficient
  - Which leagues and sports are more competitive
- Bayesian model
  - Which features have more influence on the skill

6.

Future Work

Trabalhos Futuros

Leagues

Adapt the φ coefficient to other kinds of leagues

Model

Add new features in the bayesian model

Sports

Expand the amount of sports studied

Hot Hand

Study the *Hot Hand* influence in the sport

Thanks!

Questions?

Link φ coefficient

Link NBA teams’ skill