Statistics meets Sports - when figures are more than numbers

Christophe Ley

Ghent University

ASA 2019
September 27 2019

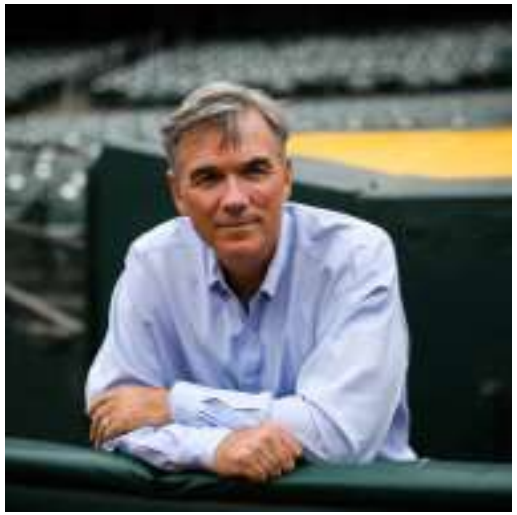
Outline

Plan

Father of sport analytics

In the 2001 after-season, Billy Beane is shocked by the bad performance of his Oakland Athletics baseball team.

Father of sport analytics



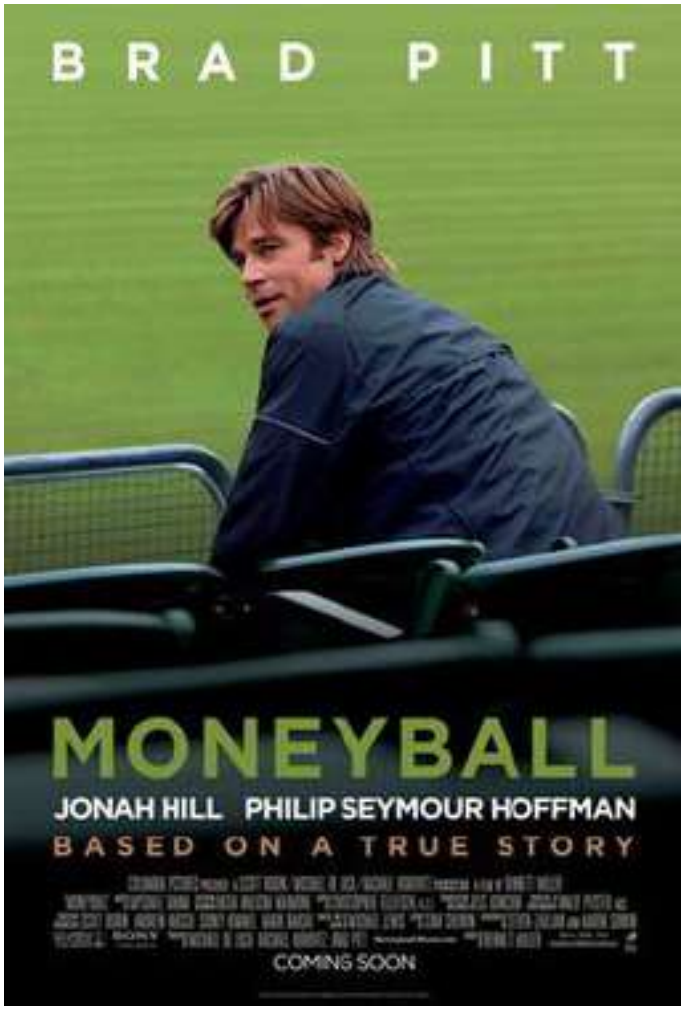
In the 2001 after-season, Billy Beane is shocked by the bad performance of his Oakland Athletics baseball team. He hires Yale economics graduate Peter Brand

Father of sport analytics

In the 2001 after-season, Billy Beane is shocked by the bad performance of his Oakland Athletics baseball team. He hires Yale economics graduate Peter Brand who uses sabermetrics to hire players.

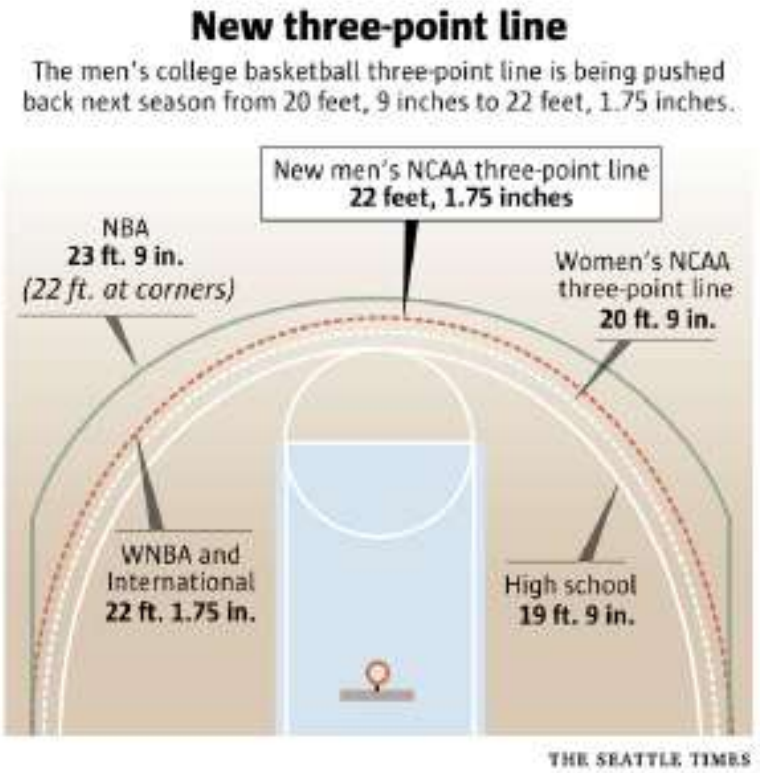
Big success $\Rightarrow$ revolutionized professional sport

Moneyball


BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
BASED ON A TRUE STORY
COMING SOON

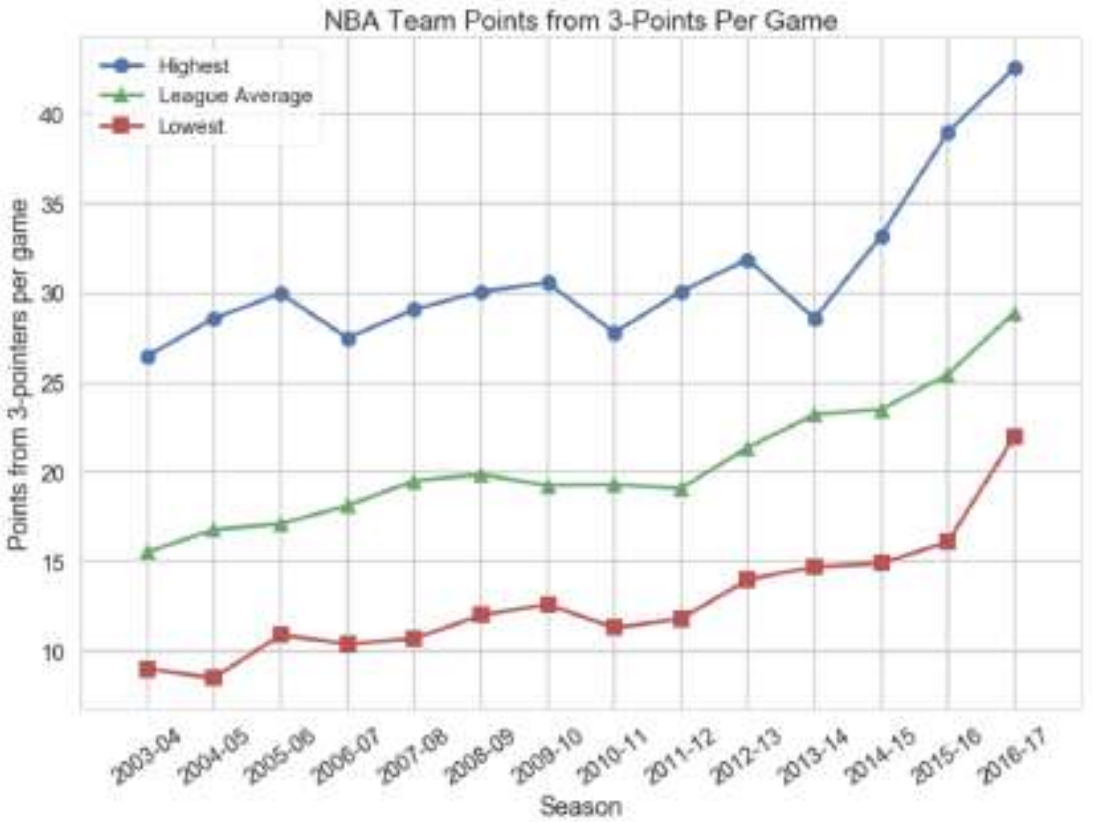
Basketball

The San Antonio Spurs head coach Gregg Popovich figured out that the 3-point line is closer in the corner.



Year	NBA Avg. Pct. of 3PA from Corners	Spurs Pct. of 3PA from Corners	Spurs Rank
2000-01	.239	.345	t
2001-02	.235	.364	t
2002-03	.257	.400	t
2003-04	.261	.363	2
2004-05	.272	.431	t
2005-06	.270	.408	t
2006-07	.271	.384	t
2007-08	.264	.454	t

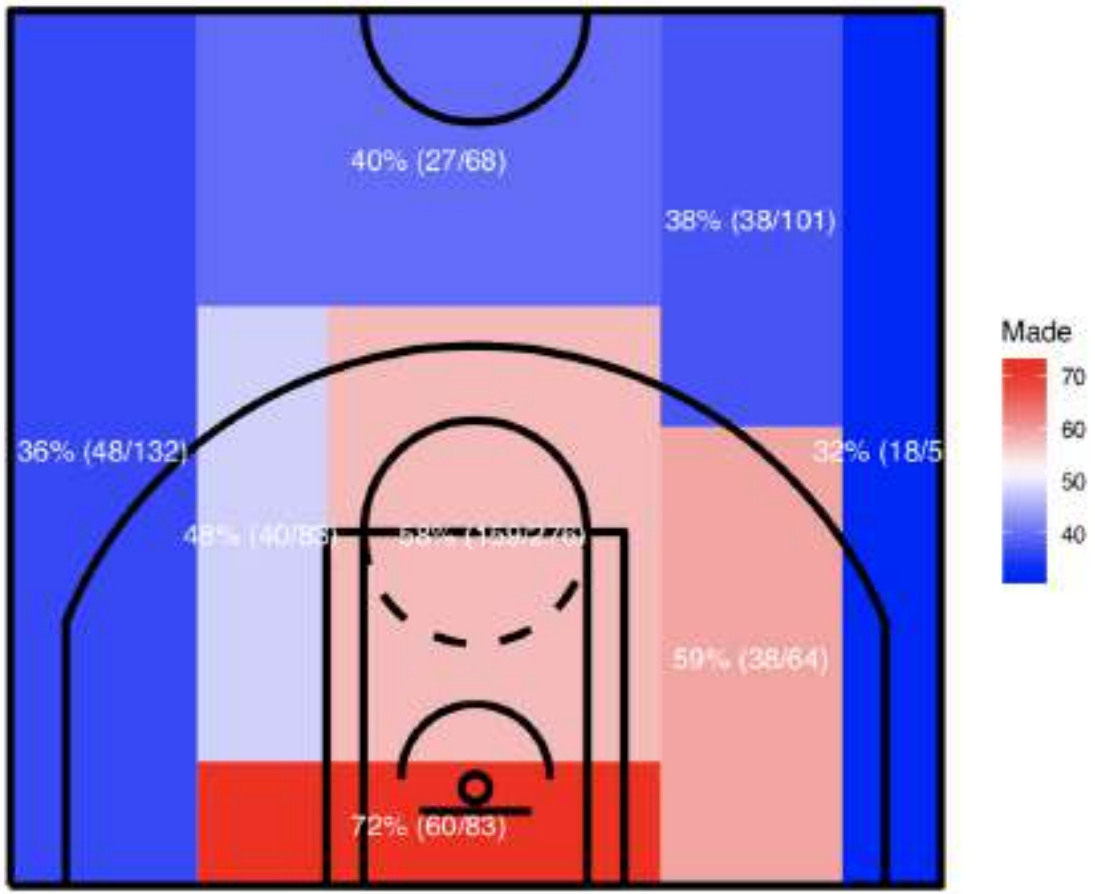
The 3-point revolution



SportVU

Innovative statistics based on speed, distance, player separation and ball possession.

NBA statistics made in Brescia



Formula 1

Monaco Grand Prix 2005 : Kimi Räikkönen won the race thanks to the real-time computations by their analysts (like how much fuel was there, how light was the car because of the reduced fuel, how much longer would the tires last, wind resistance, average lap times)

Sport Analytics - the success story

- every major professional sports team has sport analyticians
- "Analytics are the present and future of professional sports" says Leigh Stoneberg
- Fans are eager for sport-analytical content
- Websites dedicated to Sport analytics, see e.g. FiveThirtyEight.com started in March 2008 by Nate Silver
- New dedicated journal since 2015 : *Journal of Sport Analytics*

Plan

The remainder of this talk is based on the papers

Ley, C., Van de Wiele, T. and Van Eetvelde, H. (2019) Ranking soccer teams on basis of their current strength : a comparison of maximum likelihood approaches. *Statistical Modelling* **19**, 55–77

Groll, A., Ley, C., Schauberger, G. and Van Eetvelde, H. (2019) A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports, in press*

How to model a football game ?

Main idea :

- every team is given a strength parameter
- the score of a match is modelled via a bivariate Poisson distribution
- the Poisson parameters depend on the differences in strengths
- a "home effect" parameter is added if a team plays at home

There exist many distinct models, luckily the simplest turn out to be the best !

Independent Poisson model

$$\begin{aligned} Y_{ijm} &\sim Po(\lambda_{ijm}), \\ \log(\lambda_{ijm}) &= \beta_0 + (r_i - r_j) + h \cdot \mathbb{I}(\text{team } i \text{ plays at home}) \end{aligned}$$

n : number of teams

M : number of matches

Y_{ijm} : number of goals scored by team i against team j during match m

r_i, r_j : strength parameters of teams i and j (abilities)

h : home effect

β_0 : common intercept term

The basic idea goes back to Maher (1982).

This yields

$$\mathrm{P}(Y_{ijm} = x, Y_{jim} = y) = \frac{\lambda_{ijm}^{x}}{x!}\exp(-\lambda_{ijm}) \cdot \frac{\lambda_{jim}^{y}}{y!}\exp(-\lambda_{jim})$$

and consequently the likelihood can be written

$$L = \prod_{m=1}^{M}\left(\frac{\lambda_{ijm}^{y_{ijm}}}{y_{ijm}!}\exp(-\lambda_{ijm}) \cdot \frac{\lambda_{jim}^{y_{jim}}}{y_{jim}!}\exp(-\lambda_{jim})\right)^{w_{type,m}\cdot w_{time,m}},$$

with $w_{time,m}$ and $w_{type,m}$ two weight parameters.

The importance of a match is given by $w_{type,m}$ and equals 4 for World Cup matches, 3 for continental championship matches (e.g., EURO or African Cup of Nations), 2.5 for the qualifiers to the World Cup, and 1 for friendly games.

The time effect

$$w_{time,m}(t_m) = \left(\frac{1}{2}\right)^{\frac{t_m}{\text{Half period}}}$$

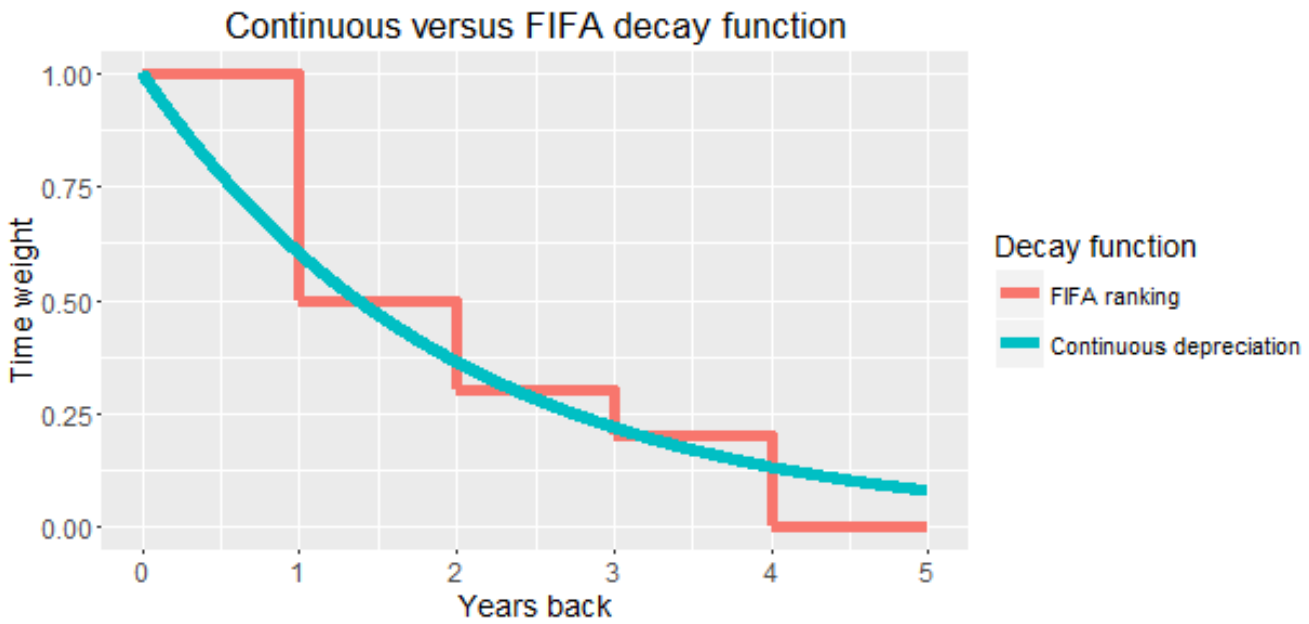


FIGURE – For a half period of 500 days

Crucial ! This weight is the main force of the new ranking.

Example

TABLE – Ranking of Premier League teams on 01.02.2018 for a Half Period of 200 days compared to the official ranking.

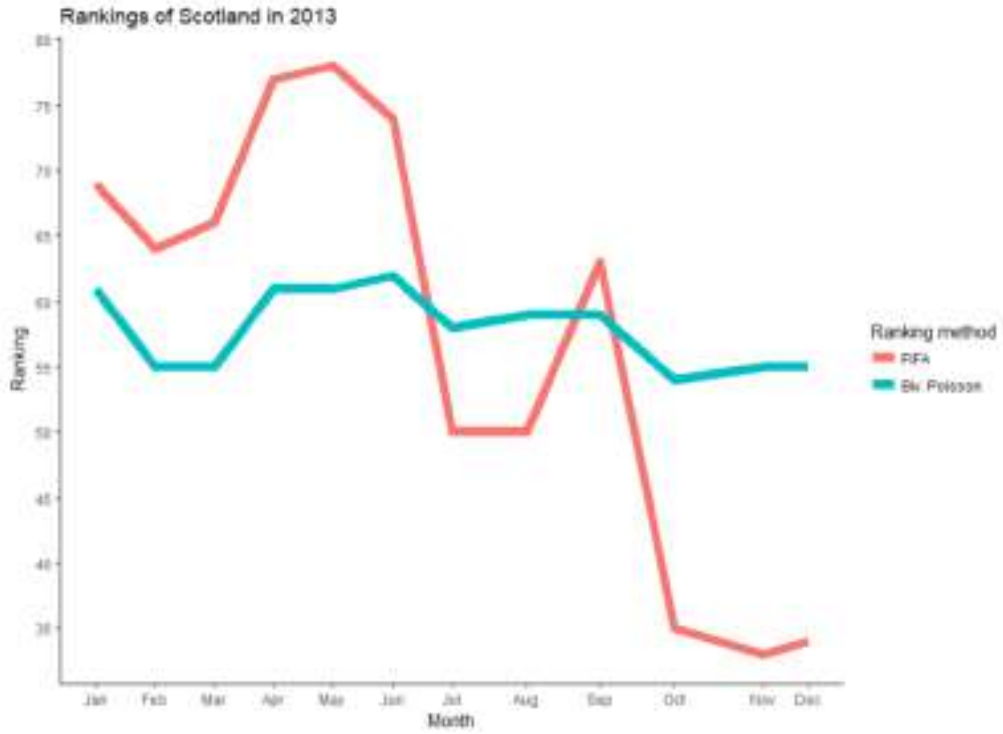
Position	Team	Strength	Team	Points
1	Manchester City	2.06	Manchester City	68
2	Liverpool	1.58	Manchester United	53
3	Manchester United	1.52	Chelsea	50
4	Tottenham Hotspur	1.49	Liverpool	50
5	Chelsea	1.42	Tottenham Hotspur	48
6	Arsenal	1.21	Arsenal	42
7	Leicester City	1.09	Burnley FC	35
8	Burnley FC	0.98	Leicester City	34
9	Bournemouth	0.95	Everton	31
10	Everton	0.88	Bournemouth	28
11	Crystal Palace	0.86	Watford	27
12	West Ham	0.86	West Ham	27
13	Southampton	0.84	Crystal Palace	26
14	Watford	0.82	Brighton and Hove Albion	24
15	Newcastle United	0.82	Huddersfield Town	24
16	West Bromwich Albion	0.80	Newcastle United	24
17	Swansea City	0.79	Stoke City	24
18	Brighton and Hove Albion	0.76	Southampton	23
19	Stoke City	0.67	Swansea City	23
20	Huddersfield Town	0.66	West Bromwich Albion	20

Important application : FIFA ranking of national teams

Notoriously problematic ranking !

Important application : FIFA ranking of national teams

Notoriously problematic ranking !
Example 1 : Ranking of Scotland in 2013



Half Period : 3 years

Example 2 : Ranking of Poland before the World Cup 2018

Position	Team	Strength	Team	Points
1	Brazil	1.76	Germany	1631(1631.05)
2	Spain	1.64	Brazil	1619(1618.63)
3	Argentina	1.64	Portugal	1446(1446.38)
4	Germany	1.61	Argentina	1445(1444.69)
5	Colombia	1.51	Belgium	1333(1332.55)
6	Belgium	1.50	Poland	1323(1322.83)
7	France	1.48	France	1226(1226.29)
8	Chile	1.46	Spain	1218(1217.94)
9	Portugal	1.43	Chile	1173(1173.14)
10	Netherlands	1.42	Peru	1160(1159.94)
11	Uruguay	1.37	Switzerland	1134(1134.5)
12	England	1.36	England	1116(1115.69)
13	Peru	1.33	Colombia	1095(1094.89)
14	Croatia	1.29	Wales	1072(1072.45)
15	Poland	1.28	Italy	1066(1065.65)

The bivariate Poisson model

Some people do not like the idea of *independence* between scores, and prefer using the bivariate Poisson distribution :

$$\mathrm{P}(Y_{ijm} = x, Y_{jim} = y) = \frac{\lambda_{ijm}^{x}\lambda_{jim}^{y}}{x!y!}\exp(-(\lambda_{ijm}+\lambda_{jim}+\lambda_{C_m}))\sum_{k=0}^{\min(x,y)}\binom{x}{k}\binom{y}{k}k!\left(\frac{\lambda_{C_i}}{\lambda_{ijm}\lambda_{jim}}\right)^{k}$$

where λ_{C_m} is a covariance parameter between the teams playing match m.

- Model proposed by Karlis and Ntzoufras (2003)
- They suggest various choices for the λ_{C_m}'s
- The best choice for this ranking is $\lambda_{C_m} = \lambda_b$ constant
- Comparison done in Ley et al. (2019) on the basis of prediction capacities

Other advantages of Poisson models

The two models are such that the goal difference $D_m = Y_{ijm} - Y_{jim}$ follows the Skellam law, and hence $\mathrm{P}(D_m > 0), \mathrm{P}(D_m = 0)$ and $\mathrm{P}(D_m < 0)$ are easy to compute.

Other advantages of Poisson models

The two models are such that the goal difference $D_m = Y_{ijm} - Y_{jim}$ follows the Skellam law, and hence $\mathrm{P}(D_m > 0), \mathrm{P}(D_m = 0)$ and $\mathrm{P}(D_m < 0)$ are easy to compute.

Besides the ranking, we can also make predictions with these models.

Plan

With the method termed "Ranking" we can already predict the World Cup. But would you be satisfied with this ?

With the method termed "Ranking" we can already predict the World Cup. But would you be satisfied with this ?

There are many more details to take into account !

With the method termed "Ranking" we can already predict the World Cup. But would you be satisfied with this ?

There are many more details to take into account !

Here are the covariates that we shall look at.

Covariates

- **Economic Factors :** GDP per capita, population

Covariates

- **Economic Factors :** GDP per capita, population
- **Sportive Factors :** bookmaker's odds (Oddset), FIFA rank, ELO rank, Ability parameters

Covariates

- **Economic Factors :** GDP per capita, population
- **Sportive Factors :** bookmaker's odds (Oddset), FIFA rank, ELO rank, Ability parameters
- **Home advantage :** host of the world cup, same continent as host, confederation

Covariates

- **Economic Factors :** GDP per capita, population
- **Sportive Factors :** bookmaker's odds (Oddset), FIFA rank, ELO rank, Ability parameters
- **Home advantage :** host of the world cup, same continent as host, confederation
- **Factors describing the team's structure :** (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad

Covariates

- **Economic Factors :** GDP per capita, population
- **Sportive Factors :** bookmaker's odds (Oddset), FIFA rank, ELO rank, Ability parameters
- **Home advantage :** host of the world cup, same continent as host, confederation
- **Factors describing the team's structure :** (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad
- **Factors describing the team's coach :** age, nationality, tenure

Covariates

- **Economic Factors :** GDP per capita, population
- **Sportive Factors :** bookmaker's odds (Oddset), FIFA rank, ELO rank, Ability parameters
- **Home advantage :** host of the world cup, same continent as host, confederation
- **Factors describing the team's structure :** (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad
- **Factors describing the team's coach :** age, nationality, tenure

All variables are incorporated as differences between the team whose goals are considered and its opponent !

Example

FRA 0 :0 URU

URU 1 :2 DEN

Team	Age	Rank	Oddset	...
France	28.3	1	0.149	...
Uruguay	25.3	24	0.009	...
Denmark	27.4	20	0.012	...
⋮	⋮	⋮	⋮	⋱

Example

FRA 0 :0 URU

URU 1 :2 DEN

Team	Age	Rank	Oddset	...
France	28.3	1	0.149	...
Uruguay	25.3	24	0.009	...
Denmark	27.4	20	0.012	...
⋮	⋮	⋮	⋮	⋱

Goals	Team	Opponent	Age	Rank	Oddset	...
0	France	Uruguay	3.00	-23	0.140	...
0	Uruguay	France	-3.00	23	-0.140	...
1	Uruguay	Denmark	-2.10	4	-0.003	...
2	Denmark	Uruguay	2.10	-4	0.003	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

The idea thus consists in using these factors as covariates in order to predict the number of goals scored at every match.

How can we best do this ? How to judge the importance of every covariate ?

The idea thus consists in using these factors as covariates in order to predict the number of goals scored at every match.

How can we best do this ? How to judge the importance of every covariate ?

We use a random forest trained on the World Cups 2002–2014.

Random forests

- **principle** : aggregation of a (large) number of **classification** / **regression trees**

Random forests

- **principle** : aggregation of a (large) number of **classification** / **regression trees**
- **final prediction** : the predictions of every tree are aggregated, either by majority vote (classification) or by averaging (regression)

Random forests

- **principle** : aggregation of a (large) number of **classification / regression trees**
- **final prediction** : the predictions of every tree are aggregated, either by majority vote (classification) or by averaging (regression)
- since we are interested in the number of goals scored by every team, we go for the regression version

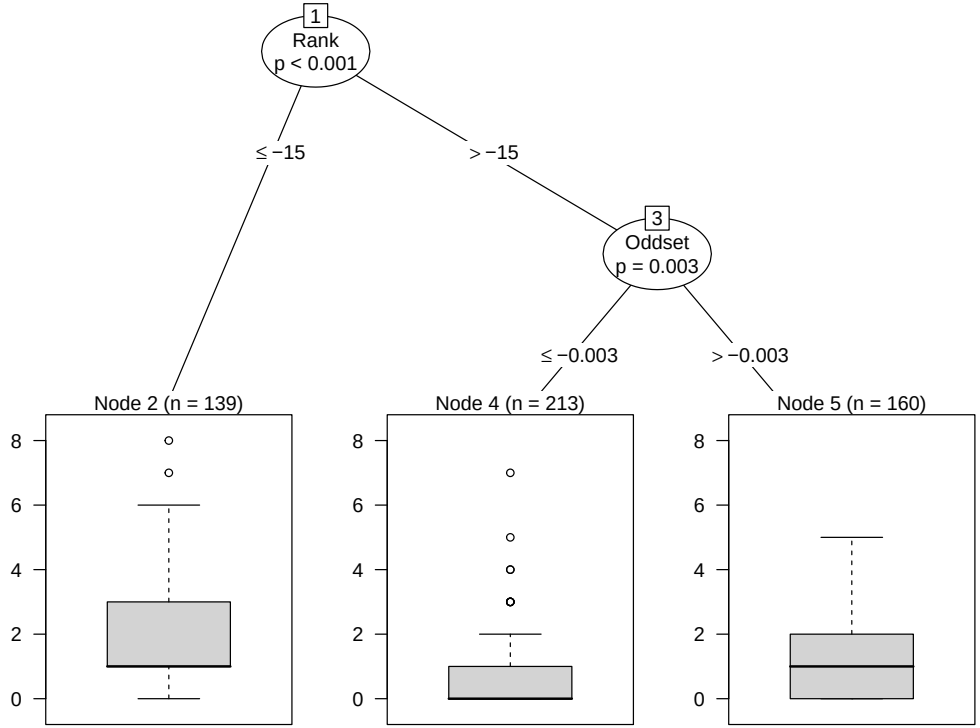
Random forests

- **principle** : aggregation of a (large) number of **classification** / **regression trees**
- **final prediction** : the predictions of every tree are aggregated, either by majority vote (classification) or by averaging (regression)
- since we are interested in the number of goals scored by every team, we go for the regression version
- for every tree : we search for the split leading to the largest difference (based on a certain criterion) between the new partitions

Random forests

- **principle** : aggregation of a (large) number of **classification / regression trees**
- **final prediction** : the predictions of every tree are aggregated, either by majority vote (classification) or by averaging (regression)
- since we are interested in the number of goals scored by every team, we go for the regression version
- for every tree : we search for the split leading to the largest difference (based on a certain criterion) between the new partitions
- observations within a same partition as similar as possible, observations in different partitions as different as possible

Random forests



Regression tree example for the data of the World Cups 2002 – 2014.

Random forests

- we build different regression trees
- two randomisation steps :

  1) the trees are based on **bootstrap samples**
  2) at every split we only consider a **(random) subset of the predictors** among which we look for the best split.

Random forests

- we build different regression trees
- two randomisation steps :

  1) the trees are based on **bootstrap samples**
  2) at every split we only consider a **(random) subset of the predictors** among which we look for the best split.

- in random forests the single trees are commonly not pruned.

Random forests

- we build different regression trees
- two randomisation steps :

  1) the trees are based on **bootstrap samples**
  2) at every split we only consider a **(random) subset of the predictors** among which we look for the best split.

- in random forests the single trees are commonly not pruned.

  $\Longrightarrow$ unpruned tree : nearly unbiased but high variance.

Random forests

- we build different regression trees
- two randomisation steps :

  1) the trees are based on **bootstrap samples**
  2) at every split we only consider a **(random) subset of the predictors** among which we look for the best split.

- in random forests the single trees are commonly not pruned.

  $\Longrightarrow$ unpruned tree : nearly unbiased but high variance.

- by combining many trees in this way, we reduce the correlation $\Longrightarrow$ predictions with **low bias and reduced variance**

Random Forests for Football

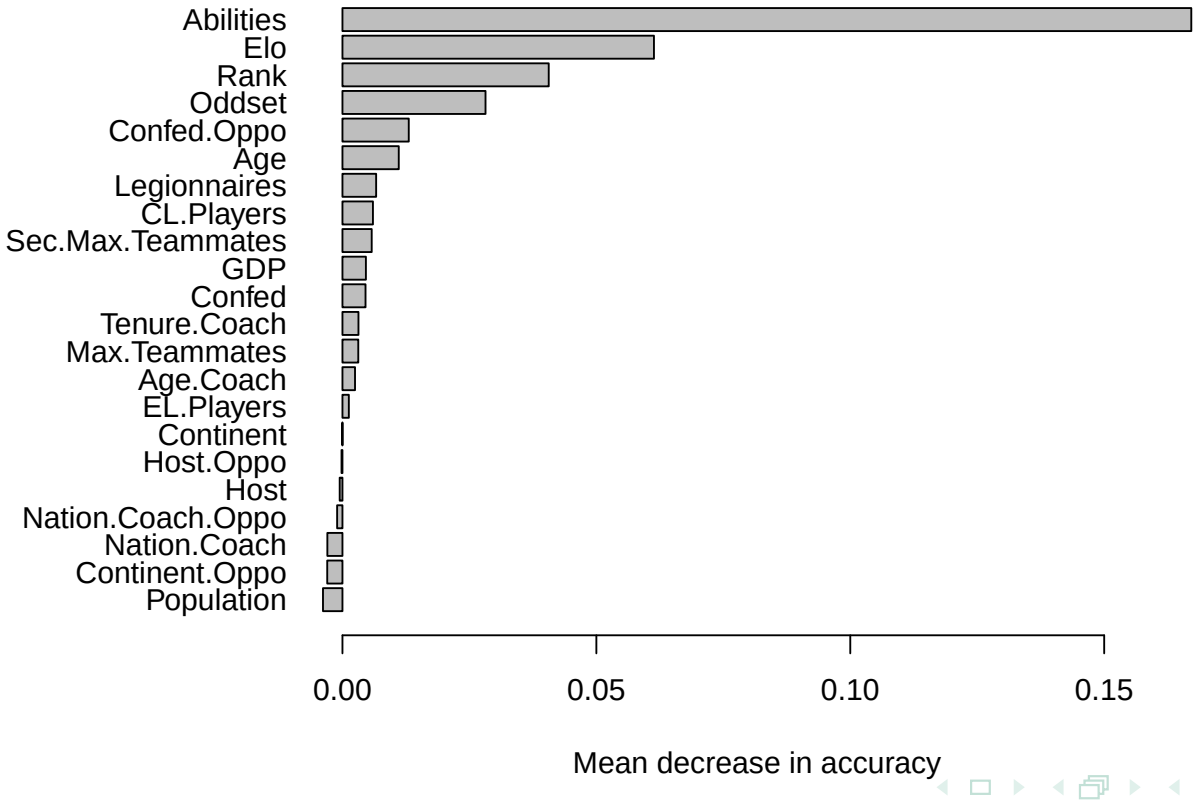
- response : metric variable *Number of Goals*
- prediction of new observation : covariate values are dropped down each of the regression trees, resulting in B predictions $\Longrightarrow$ average
- use predicted expected value as event rate $\hat{\lambda}$ of a Poisson distribution $Po(\lambda)$

Modus operandi of our World Cup prediction

- The abilities were estimated by the bivariate Poisson model with a half period of 3 years.
- All matches of the 228 national teams played since 2010-06-13 up to 2018-06-06 are used for the estimation $\Rightarrow$ more than 7000 matches.
- Predictors taken last week
- World Cup is simulated 100.000 times (special steps at knock-out stage)

Training/Learning Phase

Our random forest learns the importance of every covariate on the basis of the World Cups 2002, 2006, 2010 and 2014.



Winning probabilities

		Round of 16	Quarter finals	Semi finals	Final	World Champion	Oddset
1.	ESP	80.5	61.2	38.0	22.7	13.7	11.8
2.	GER	78.0	49.0	30.4	19.1	11.5	15.0
3.	FRA	77.8	49.9	32.1	18.5	10.8	11.8
4.	BRA	75.0	44.1	28.0	17.6	10.3	15.0
5.	BEL	75.9	52.5	30.1	17.7	9.9	8.3
6.	ENG	73.1	49.8	26.6	14.7	7.5	4.6
7.	ARG	71.8	39.9	22.3	11.1	5.4	8.3
8.	CRO	66.4	33.5	18.3	8.5	3.8	3.0
9.	POR	61.1	39.8	18.7	8.0	3.2	3.8
10.	COL	71.4	32.2	15.5	7.4	3.2	1.8
11.	SUI	55.4	29.5	14.1	6.6	2.9	1.0
12.	URU	82.7	38.5	16.9	7.1	2.8	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Most probable group stage

Group A 25.0%	Group B 27.7%	Group C 23.8%	Group D 22.8%
1. URU	1. ESP	1. FRA	1. ARG
2. RUS	2. POR	2. DEN	2. CRO
KSA	MOR	AUS	ICE
EGY	IRN	PER	NGA

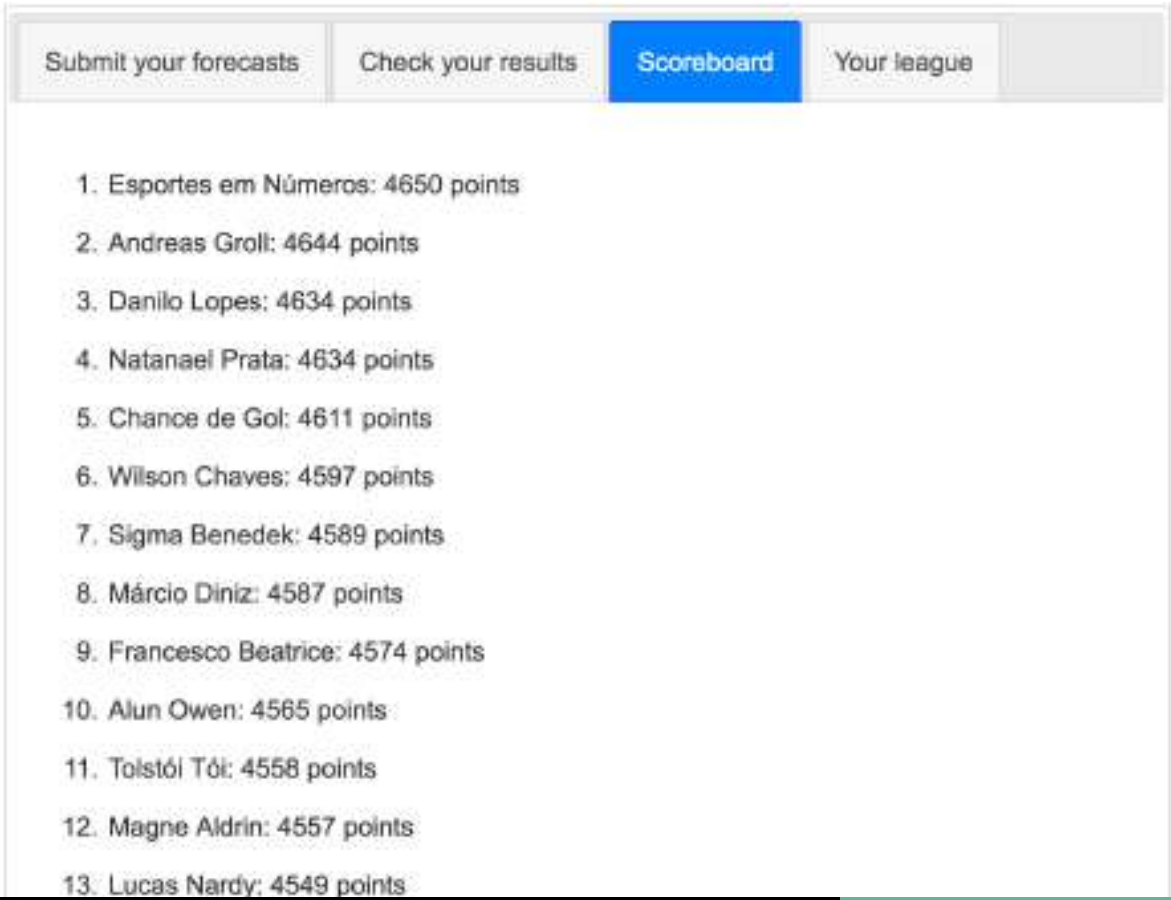
Group E 21.5%	Group F 23.3%	Group G 26.7%	Group H 19.7%
1. BRA	1. GER	1. BEL	1. COL
2. SUI	2. SWE	2. ENG	2. POL
CRC	MEX	PAN	SEN
SRB	KOR	TUN	JPN

Performance I

	Likelihood	Class. Rate	RPS
Hybrid Random Forest	0.442	0.609	0.190
Bookmakers	0.438	0.562	0.194

Performance II

Final standing in forecast competition fifaexperts.com (> 500 participants) :



Performance III

Final standing in forecast competition Kicktipp (with colleagues) :

Gesamtübersicht

Spieltagspunkte ▼

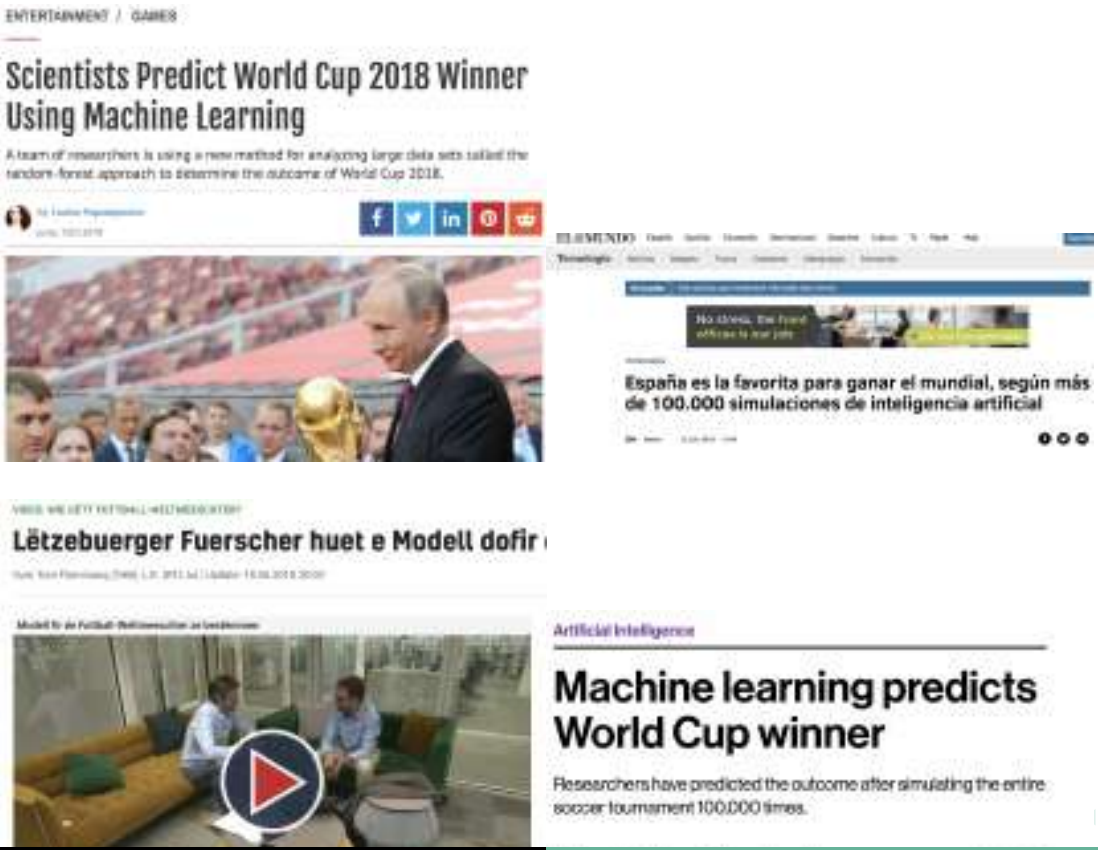
		Spieltage													
Pos	Name	1	2	3	4	5	6	7	Ac	Vi	Ha	Fi	B	S	G
1	stats_model	14	13	14	9	12	10	19	13	7	4	4	28	2,50	147
2	Hendrik	20	14	9	9	11	5	8	12	9	4	0	28	1,83	129
3	Katharina	12	11	9	10	15	10	11	16	7	3	2	20	1,50	126
4	Katrin	12	14	8	6	12	4	15	18	7	4	2	24	0,83	126
5	Lukas	10	12	9	6	9	6	4	15	7	3	6	32	1,00	119
6	Jona	10	9	6	10	9	6	11	12	8	6	7	24	1,00	118
7	Hilsi	16	8	7	7	10	2	6	14	9	7	2	24	1,50	112
8	Borussenengel	13	10	10	11	14	2	5	14	5	4	2	16	1,00	106
9	Christina	8	14	7	4	12	4	8	18	4	0	4	12	0,83	95

Figures became definitely more than numbers...

Before the start of the World Cup 2018, we put our research paper about the prediction of the WC 2018 online. Business as usual.

Figures became definitely more than numbers...

Before the start of the World Cup 2018, we put our research paper about the prediction of the WC 2018 online. Business as usual. But then it went viral...



However...

Journalists, or more generally non-statisticians, have trouble interpreting such tournament predictions !

Spain 13.7%, Germany 11.5%, France 10.8%...

Models for match-by-match analysis exist already, cf. fifaexperts.com.

However...

Journalists, or more generally non-statisticians, have trouble interpreting such tournament predictions !

Spain 13.7%, Germany 11.5%, France 10.8%...

Models for match-by-match analysis exist already, cf. fifaexperts.com.

We proposed a general method to evaluate any type of tournament predictions in :
Ekstrøm, C.T., Ley, C., Van Eetvelde, H. and Brefeld, U. (2019) Evaluating one-shot tournament predictions. *Submitted*

Plan

What should/could be done ?

- Make sense out of the shear amount of available data !
- Filter out useful information.
- How should data be collected by sports clubs ? Which data can help them ?
- How to take into account psychological aspects ?
- How to prevent injuries that can cost fortunes (and health) ?
- Make a stronger link between sports world and Data scientists/statisticians.
- Inter-disciplinarity

An attempt to address challenges

We are currently setting up a European network called S-TRAINING :

Sports – **T**raining and **R**esearch in D**A**ta Sc**I**ence Methods for A**N**alytics and **IN**jury Prevention **G**roup

Research Institutes Involved

Non-academic partners, sport clubs and journal (e.g., Journal of Sport Analytics) are part of the network.

If you are interested : christophe.ley@ugent.be.

Thank you for your attention !