



Scientific Challenges and Problems in Football Analytics

Ioannis Ntzoufras

Department of Statistics

Athens University of Economics & Business



AUEB Sports Analytics Group

Founded on 2015



Research group in the Computational and Bayesian Statistics Lab of the Department of Statistics

Team Leaders: Dimitris Karlis and Ioannis Ntzoufras

Web page: <https://aueb-analytics.wixsite.com/sports>



Comp Bayes AUEB Lab

AUEB Sports Analytics Group

Events organized by the Group

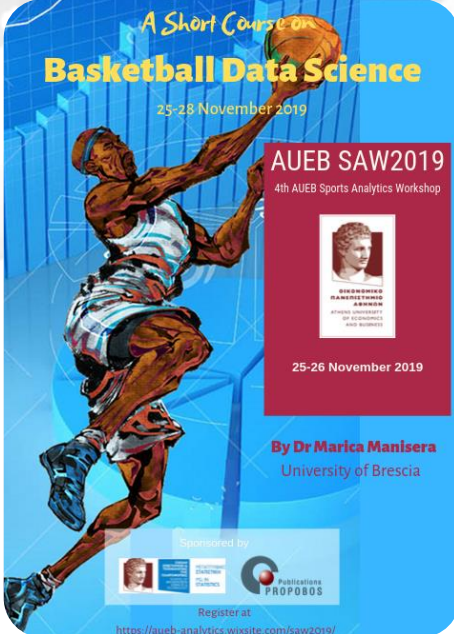


- Organized four [AUEB Sports Analytics Workshops](#) (2016-2019)
- Organized [MathSport 2019](#) International Conference on 1-3 July 2019 at Athens
- Short Courses on [Sport Economics](#) by Professor Stefan Kesenne
- Short Course on [Basketball Data Science](#) by Dr. Marica Manisera

AUEB Sports Analytics Group

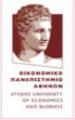
Sports Analytics Events

- [Upcoming SAW2021 – March/April 2021](#)
- [Past event: https://aueb-analytics.wixsite.com/saw2019](https://aueb-analytics.wixsite.com/saw2019)



A Short Course on
Basketball Data Science
25-28 November 2019


AUEB SAW2019
4th AUEB Sports Analytics Workshop



25-26 November 2019

By Dr Marica Manisera
University of Brescia

Sponsored by



Register at
<https://aueb-analytics.wixsite.com/saw2019/>



AUEB SAW2019
4th AUEB Sports Analytics Workshop



25-26 November 2019

Featuring
A Short course on
BASKETBALL DATA SCIENCE

Topics of the workshop

- Football Analytics and Machine Learning
- Basketball Analytics
- Basketball Performance Analysis
- Volleyball Prediction
- Sport Economics
- Competitive Balance in Football
- Football Scheduling

Register at
<https://aueb-analytics.wixsite.com/saw2019/>

Sponsored by



AUEB Sports Analytics Group

Research & Publications

- **European Training Network** on *Sports Analytics and Injury Prevention*
- Discussing with **ETN** “*Real-time Analytics for Internet of Sports*”
- **14 M.Sc. Theses** on Football Prediction, Player Evaluation, Competitive balance, Water Polo, Volleyball Modelling & Basketball Analytics (8 in Statistics, 3 in Business Analytics, 1 in Data Science, 1 in UCD & Stats, 1 in UCL)
- **2 Active Ph.D. Students** partially working on related topics
- **10 Publications** on *JRSSC, JRSSA, J. Management Mathematics, J Quantitative Analysis in Sports, Economics Bulletin, International Journal of Computer Science in Sport, Journal of Statistical Software*. (2 Submitted & 1 under submission)
- *We are considering the possibility of having a Sports Analytics M.Sc.*
- Our Publication on Football Modelling using the Bivariate Poisson Model is a **key publication** on the field.



Ioannis Ntzoufras

Professor in Statistics, Department of Statistics, [Athens University of Economics and Business](#)

Η διεύθυνση ηλεκτρονικού ταχυδρομείου έχει επαληθευτεί στον τομέα aueb.gr - [Αρχική σελίδα](#)
[Model and Variable Selection](#) [Bayesian computation](#) [Bayesian Modelling](#) [Sports modelling](#)

ΤΙΤΛΟΣ	ΠΑΡΑΤΙΘΕΤΑΙ ΑΠΟ	ΕΤΟΣ
Bayesian modeling using WinBUGS I Ntzoufras John Wiley & Sons	1519	2011
On Bayesian model and variable selection using MCMC P Dellaportas, JJ Forster, I Ntzoufras Statistics and Computing 12 (1), 27-36	504	2002
Analysis of sports data by using bivariate Poisson models D Karlis, I Ntzoufras Journal of the Royal Statistical Society: Series D (The Statistician) 52 (3 ...	394	2003
Factorial composition of self-rated schizotypal traits among young males undergoing military training NC Stefanis, N Smyrnis, D Avramopoulos, I Evdokimidis, I Ntzoufras, ... Schizophrenia Bulletin 30 (2), 335-350	180	2004
Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference D Karlis, I Ntzoufras IMA Journal of Management Mathematics 20 (2), 133-145	133	2009
Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R D Karlis, I Ntzoufras Journal of Statistical Software 14 (10), 1-36	5127	2005

Collaborations



- Sotiris Drikos, Past Manager of the Greek National Volleyball team
- Stefan Kesenne, Top Sports Economist, consultant of the Bosman case
- Yiannis Kosmidis, Warwick University UK
- Nial Friel, The Insight Centre for Data Analytics, University College Dublin
- Konstantinos Pelechrinis, University of Pittsburg
- Gianluca Biao, University College London
- Leonardo Egidi, University of Trieste
- Dr Christos Marmarinos (and Lazaros Papapdopoulos)



Introduction

Football/Soccer is the best sport for implementing Science/Statistics/Analytics

- Low number of events (so difficult to predict)
- High uncertainty (so difficult to predict)
- Very popular (because it is difficult to predict?)
- Very profitable (because it is difficult to predict?)
- High Financial Risk of investment (because passion becomes more important than numbers and science) – Professional Teams are usually acting as win-maximizers and not profit-maximizers



Main Topics Quantitative analysis of Football/Sports

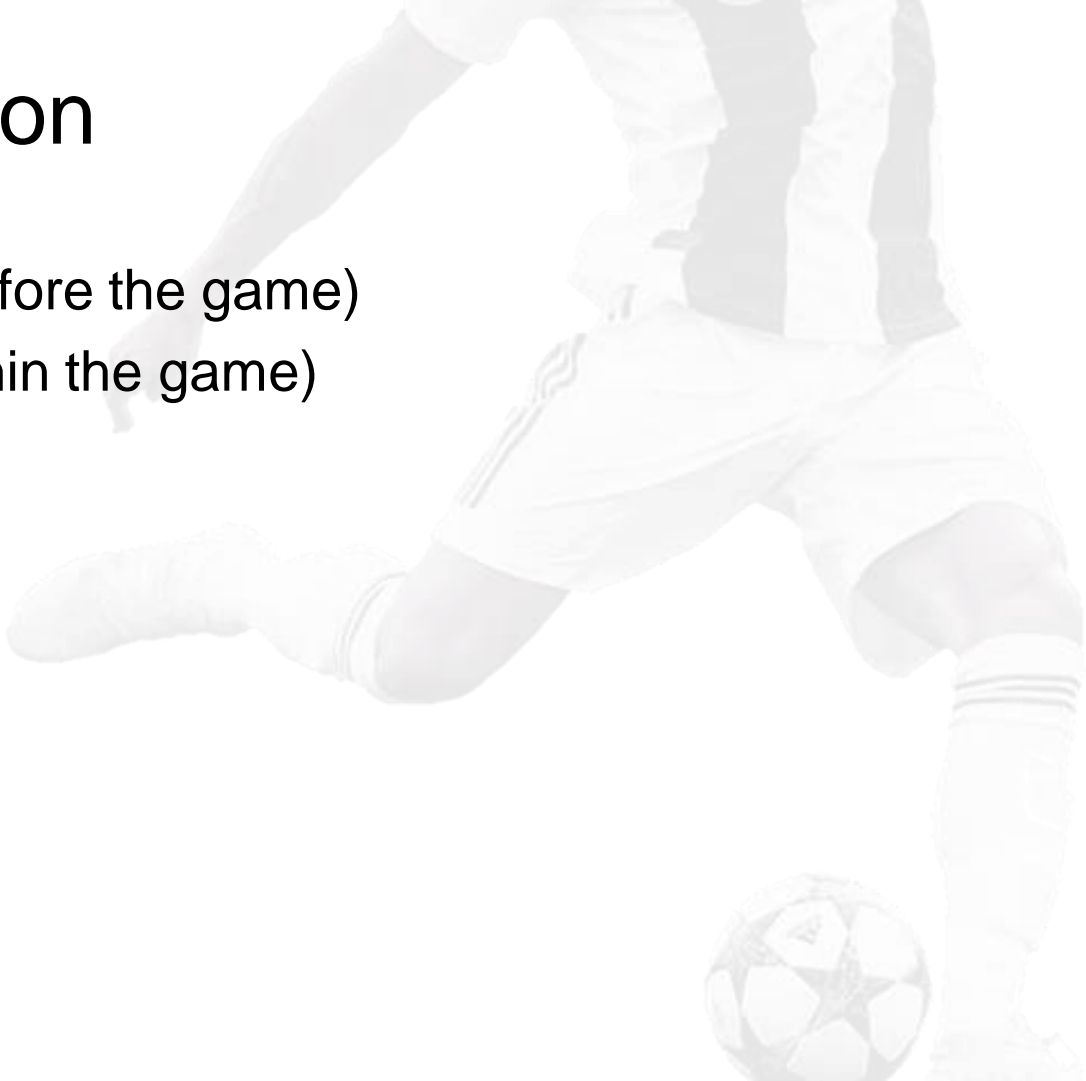
- Prediction
- Player Evaluation & Performance analytics
- Physical Metrics of Players in training
- Inline game metrics with wearables
- Scheduling
- Sports Economics & Competitive Balance
- Other (Passing Network Analytics, Referee effects, Red card effect, Home effect, Corruption Analytics, Analysis of substitution times)
- Extensive presentation in Bodosaki Lectures on Demand (BLOD):

<https://www.blod.gr/lectures/football-analytics-problimata-methodoi-kai-diaskedastiki-statistiki/>



Prediction

- Offline (before the game)
- Inline (within the game)



Offline Prediction



Modeling of

- Game Scores
 - Poisson based models and extensions
 - Modeling the difference using the Skellam model
- Final outcome of a game (Win/Draw/Loss)
 - Multinomial regression model
 - Bradley Terry Model

Models for Scores

Models for Counts (goals)

- Simple Poisson Model (Maher, 1982; Lee, 1992; Dixon & Coles, 1997, Karlis and Ntzoufras, 2000)
- Bivariate Poisson Model (Karlis & Ntzoufras, 2003)
- Negative Binomial Model (see e.g. Ntzoufras 2009)
- Skellam Model for the goal difference (Karlis & Ntzoufras, 2009)
- Poisson-log-normal random effects model (not the best for football counts; see e.g. Ntzoufras 2009)

Models for Scores



Such models allow us not only to predict a single football game but also (simulation-based results)

- Final League reproduction
- Estimate probabilities of winning a league, winning European tickets, or relegation.
- Estimate final rankings
- Estimate results under different scenarios/assumptions (by changing covariates i.e. conditions of the game)

Offline Prediction

Poisson Based models

- Vanilla model: home effect + teams attacking and defensive parameters
- Models with time evolved team parameters (time and form matters!)
- Additional covariates
 - Odds from betting teams (easily accessible – good covariates)
 - Team performance (ingame and before the game)
 - Information about events and formation (team strategy, formation, injuries etc.)
 - Economo-demographic variables (Stability, tradition, Budget, Player Value, Coach Value, Country of origin for European leagues)
 - Prior information (previous games between the teams)
 - Team form (e.g. performance in last 5 games)

Offline Prediction

The simple (vanilla) Poisson model

The model is expressed by

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\lambda_{ik}) && \text{for } j = 1, 2 \\ \log(\lambda_{i1}) &= \mu + \text{home} + a_{\text{HT}_i} + d_{\text{AT}_i} \\ \log(\lambda_{i2}) &= \mu && + a_{\text{AT}_i} + d_{\text{HT}_i} \quad \text{for } i = 1, 2, \dots, n, \end{aligned}$$

where n = number of games, μ = constant parameter; home = home effect; HT_i and AT_i = home and away teams in i game; a_k and d_k = attacking and defensive effects–abilities of k team for $k = 1, 2, \dots, K$; and K = number of teams in the data (here $K = 20$).

In full (balanced) round-robin leagues, the parameters can be easily calculated by considering averaged of scored/conceded goals for each team

Offline Prediction

Data for the simple (vanilla) model

- **Observations**
 - $2 \times$ Number of games (N)
 - Each game will occupy two lines/observations (one for home team and one for away team)
- **Response Variable:** Goals scored by each team in each game
- **Covariates**
 - **Home effect:** Binary for home and away teams (1 for home teams and zero otherwise)
 - **Scoring team:** Categorical factor for the team scoring the number of goals (the corresponding coefficient will estimate the attacking ability of each team)
 - **Team accepting goals:** Categorical factor for the team receiving the number of goals (the corresponding coefficient will estimate the defensive ability of each team).

Offline Prediction

Important Assumptions

- Dependence/Independence of Goals of a game
- Time dependent attacking and defending parameters
- What about draw inflation?
- What about Over-dispersion?
- Shall we focus on modeling scores or outcomes (win/draw/loss)?

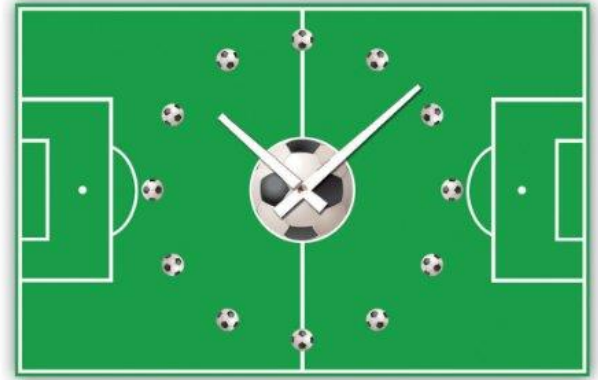
Checking the performance of the predictions

- Checking model fit and prediction using in-sample and out-of-sample measures

Prediction within the game

Modeling of

- Time to event (goal)
 - Survival analysis-based models
 - Dixon & Robinson (1998, *RSSD*)
 - Nevo and Ritov (2013, *JQAS*)
 - Boshnakov, Kharrat, McHale (2017, *Int. J. Forecasting*)
 - Work in progress by our team
- Model the probability of event for short intervals (every 1 or 5 minutes)
 - Using Binomial mixed models for repeated measures
 - Narayanan, S., Kosmidis, I., Dellaportas, P., 2020. **Bayesian modelling of flexible marked point processes with applications to event sequences from association football.** (*working paper*)



Player Evaluation



Aim

- Estimate the contribution of players in a team
- Rank, identify and reward best players
- Scouting – Early Identification of talents
- Estimate the future performance/value of a Player
- Help the manager to decide the best formation

Player Evaluation

Methods

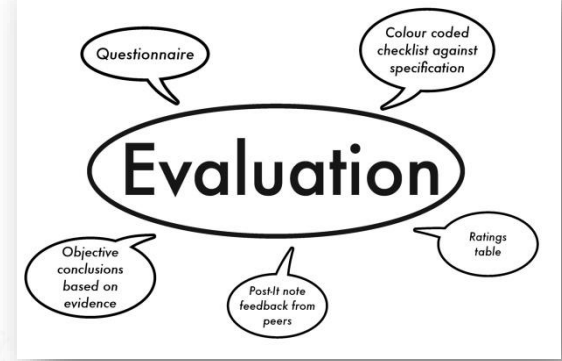
- Simple approach with binary indicators (plus-minus metric)
- Random effects
- Analysis based on Game Performance Indicators
- Expected Goals (xG) and Expected Assists (xA)
- Player Economic/Marketing Value and performance



Player Evaluation

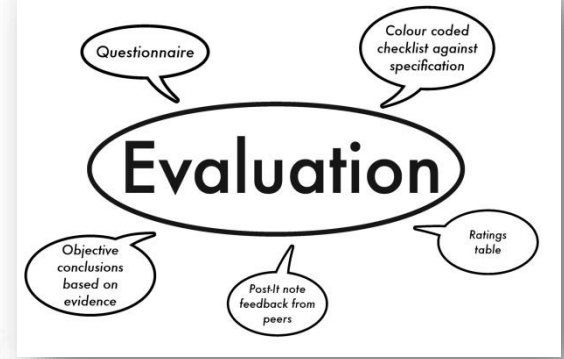
Methods (2)

- Simple approach with indicators
 - Build a model with indicators whether a player was in the field
 - Binary indicators for players
 - Difficult to build a dataset. Each game should be spitted in multiple lines according to substitution times
- Analysis based on Game Performance Indicators
 - Build a model to identify the importance of each event in the game (goals, shots, steals, passes, speed, stamina, area covered etc.)
 - Use model indicators to build an index of players
 - McHale, Scarf & Folker (2012, *Interfaces*) building different indexes based on different response measures



Player Evaluation

Methods (2)



Nice references on the plus-minus regression-based methods

- Kharrat, T., McHale, I. G., & Peña, J. L. (2020). [Plus–minus player ratings for soccer](#). *European Journal of Operational Research*, 283(2), 726–736
- Hvattum, L. M. (2019). [A comprehensive review of plus-minus ratings for evaluating individual players in team sports](#). *International Journal of Computer Science in Sport*, 18(1), 1–23.

LASSO Regression is one of the main tools due to the large number of features

Player Evaluation



Football Player Ratings
WITH
LARS MAGNUS HVATTUM

 **Football Player Ratings**
196 subscribers

SUBSCRIBED 

HOME VIDEOS PLAYLISTS CHANNELS DISCUSSION **ABOUT** 

Description

Welcome to Football Player Ratings!

On this channel you will find 1) videos describing and analyzing mathematical models for evaluating individual football players and 2) videos presenting and discussing ratings produced by our best player rating models.

If you are interested in the topic of how to evaluate football players, or just like to have some great arguments for why your favorite player is better than another player, consider subscribing for easy access to new videos.

I'll be happy to get any suggestions about specific players/teams/leagues/matches to analyze, so be sure to leave a comment to let me know what you want to see!

Stats

Joined 29 Oct 2017

7,427 views



<https://www.youtube.com/channel/UC64jAkIQX-hD3pSnnOmr2MA/>

Player Evaluation

Methods (3)

- Random effects
 - Use random effects to identify individual contribution
 - Goal Scoring: McHale & Szczepanski (2014, *JRSSA*)
 - Passing Skills: Szczepanski & McHale (2016, *JRSSA*)
- Player Economic/Marketing Value and performance
 - Saebo & Hvattum (2018, *Journal of Sports Analytics*): *Modelling the financial contribution of soccer players to their clubs*
 - Evaluating the efficiency of the association football transfer market using regression based player ratings (pre-print only)

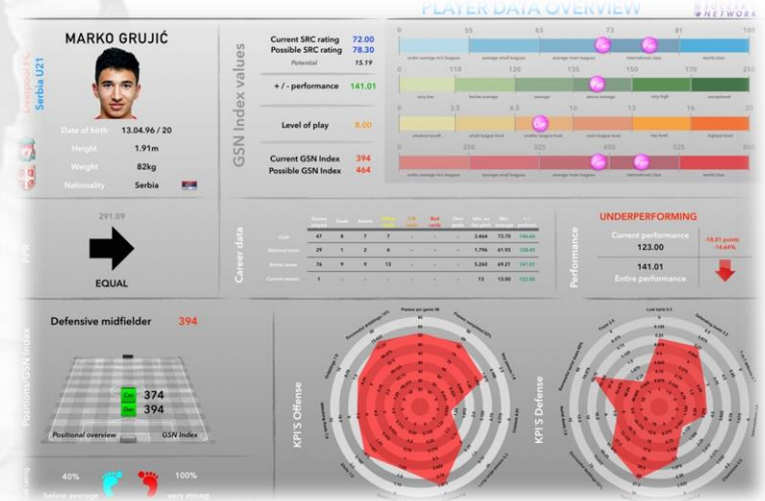


Player Evaluation

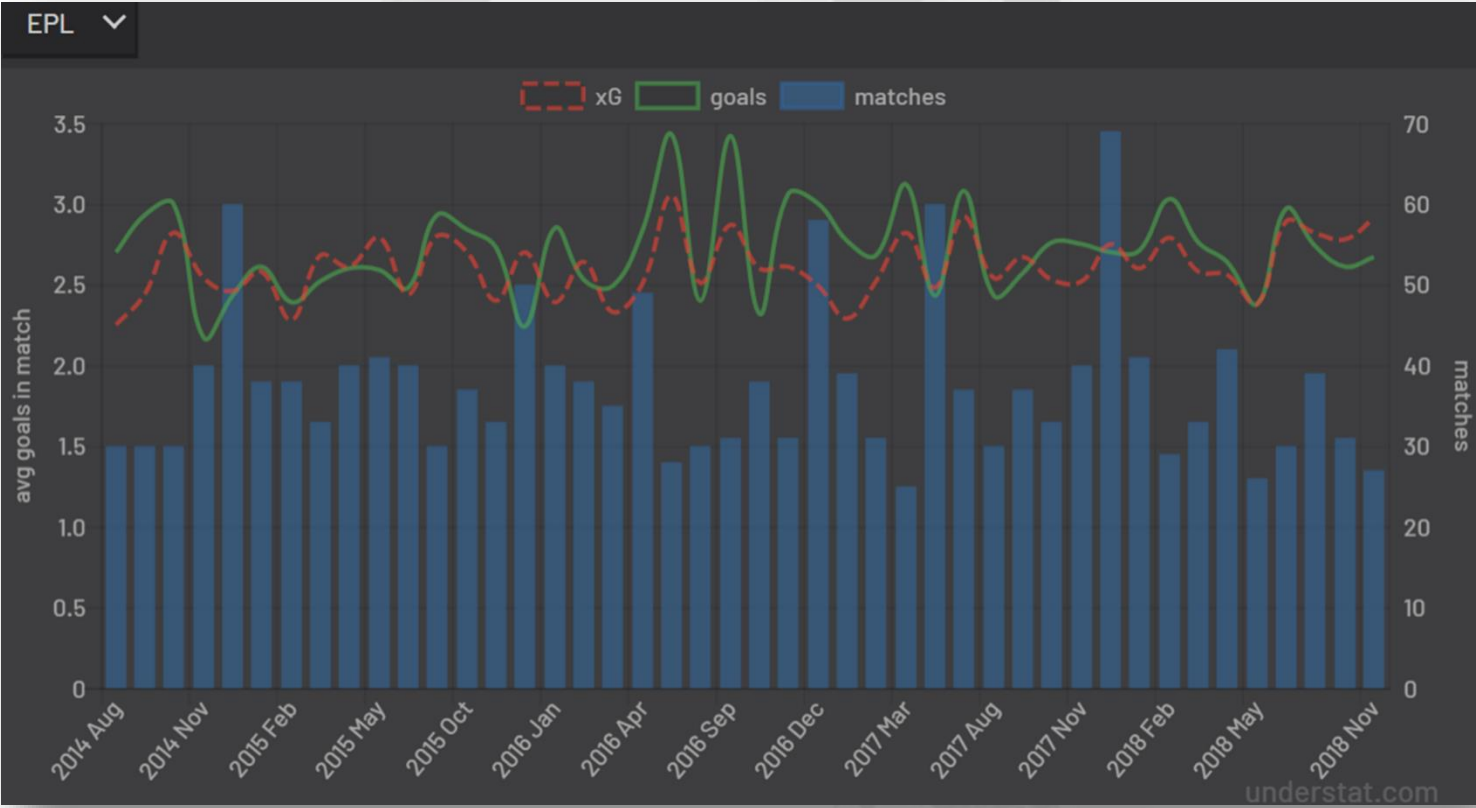
Methods (4)

Expected Goals (xG)

- We model every shot
- Response measure: is the probability of a shot resulting in a goal
- The sum of these probabilities will give the xG of a player and a team
- Similar for assists (xA)
- References:
 - <https://www.optasports.com/services/analytics/advanced-metrics/>
 - <https://understat.com/>



Player Evaluation



Expected Goals (xG): <https://understat.com/>

Player Evaluation

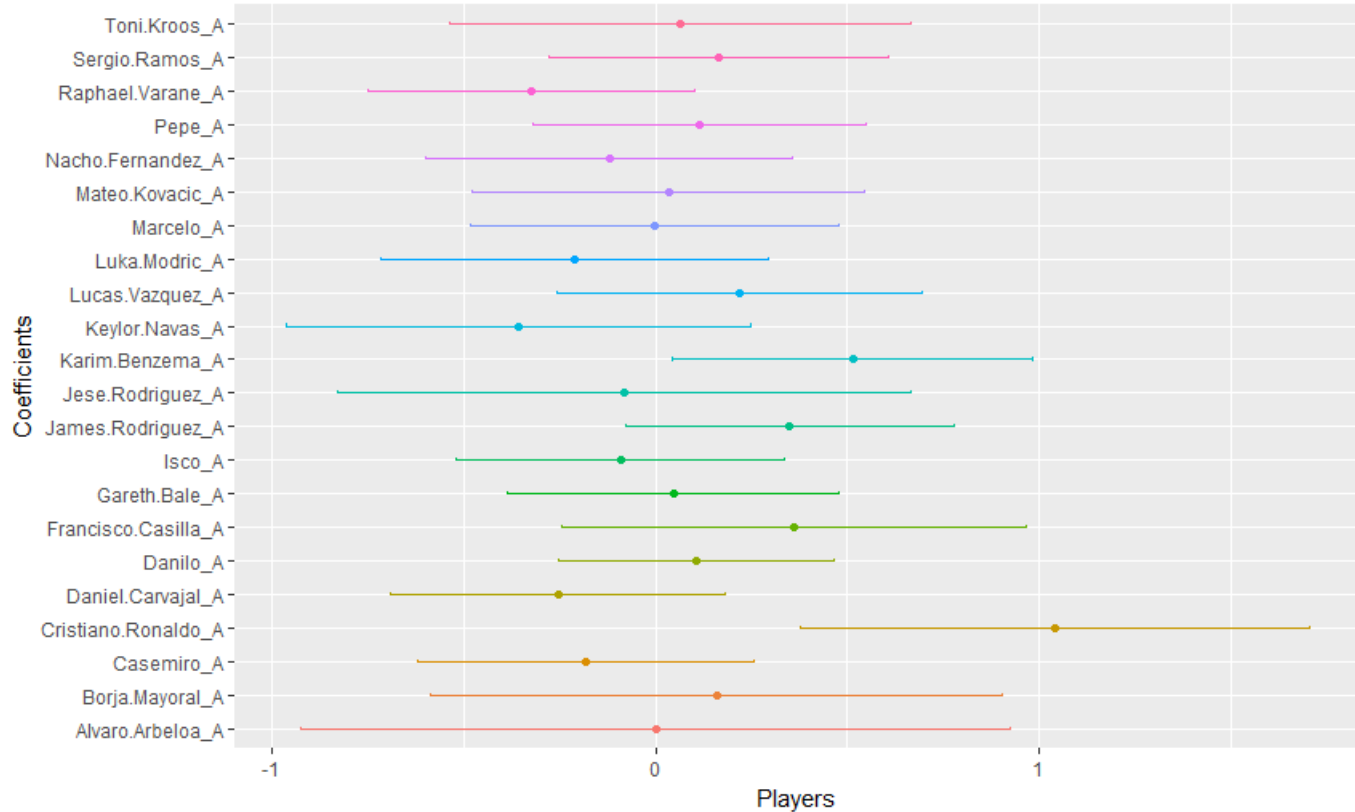
Example of the Simple approach with indicators

- 351 matches of the La Liga Season 2015/2016
- 954 goals (555 goals were scored by home teams, 399 conceded)
- 110 scored by Real Madrid, 34 conceded
- M.Sc. Thesis at AUEB by A. Mourtopallas



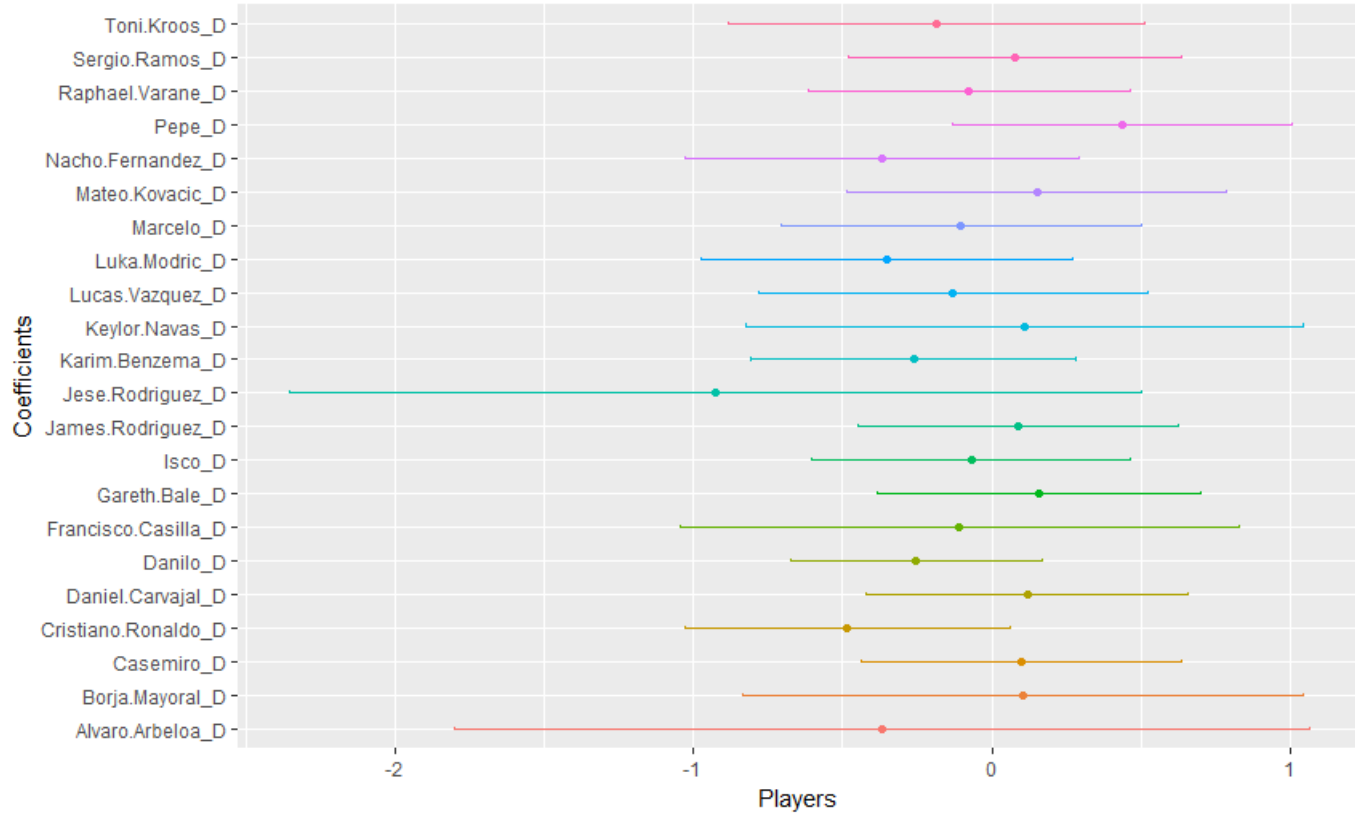
Player Evaluation

Players errorbars for the attacking ability

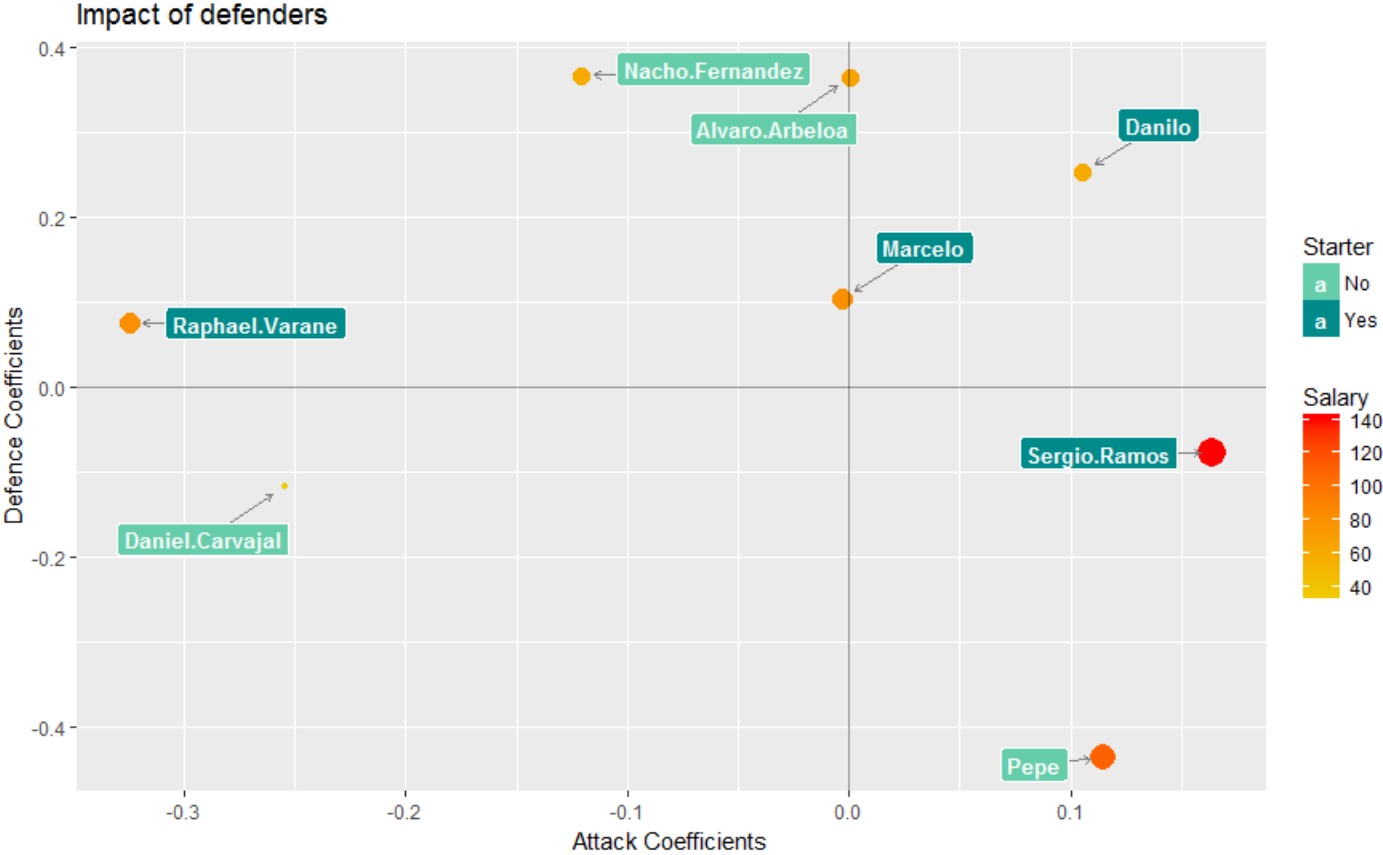


Player Evaluation

Players errorbars for the defensive ability

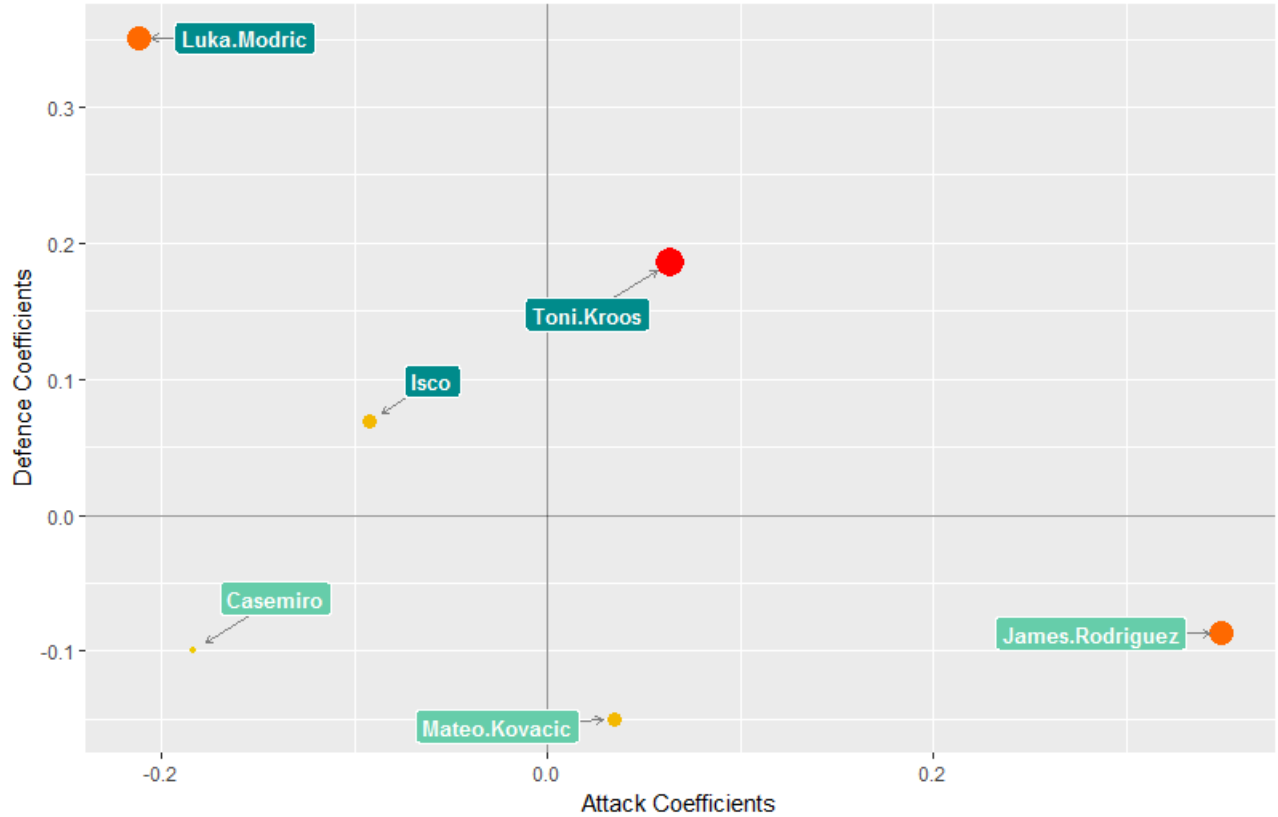


Player Evaluation



Player Evaluation

Impact of midfielders



Salary

150
125
100
75
50

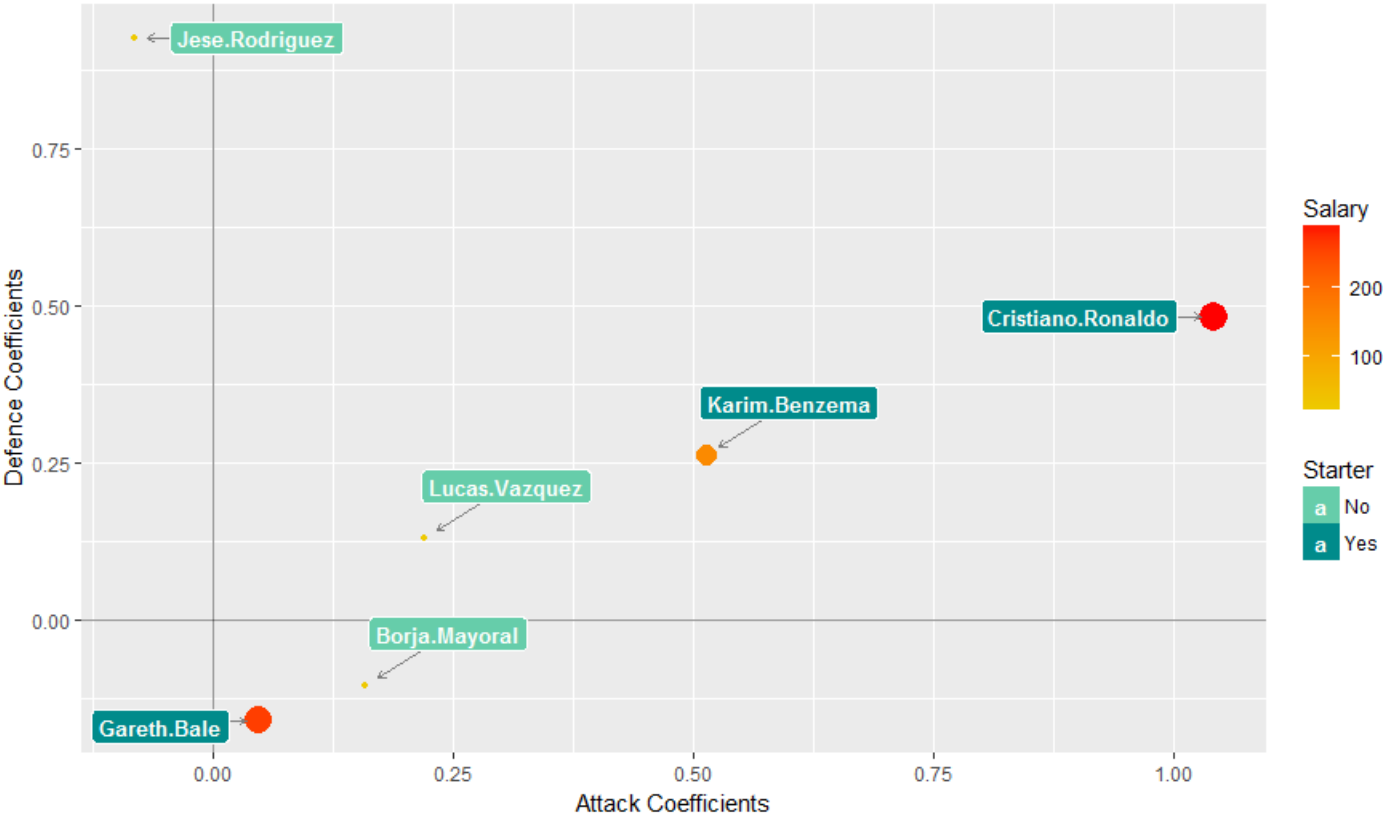
Starter

a No
a Yes



Player Evaluation

Impact of forwards



Real Madrid 2015/16 Player Evaluation



Conclusions



Cristiano Ronaldo was the key player of the team

Tony Kroos' impact was higher than we may have presumed



Nacho Fernandez improved since previous season (very high def contribution)

Lucas Vasquez was a very promising player (contributed positively in both attack and defensive dimensions with low salary)



Gareth Bale performed less than expected (overprized)

Pepe ⇒ low defensive contribution – high salary (overprized?)

Metrics for physical improvement and training

Aim

- Improve the physical condition of athletes
- Focus on specific skills and measure them
- Avoid injuries
- Improves the team by optimizing allocated training time and resources



Inline game metrics with wearables

The aim is to measure

- Movement of players in the game
- Speed and coverage
- Physical condition
- Physical and tactics performance

It helps

- Evaluate the performance of players and teams within a game
- The manager to decide formation and substitutions



League and Contest Scheduling



AIM

- Fair scheduling
- Eliminate bias due to the sequence of games
- Strengthen competitiveness (related with next slides)
- Incorporate constraints (incl. other sports, safety issues, other events, tv requirements etc.)

HOW?

- Using Operational Research and optimization methods
- Hybrid search methods
- Validate using simulation methods from Statistical models

Sports Economics & Competitive balance

Competitive Balance

- A balanced league increases the interest of the fans and improves the athletic product (Uncertainty of Outcome Hypothesis)
- The notion of a balanced league is not yet very well defined
 - Equal Strength between all teams? or
 - Equal Strength between best teams (or the teams with the highest number of fans?)
- A lot of work on the topic by AUEB Sports Analytics Group
 - Dr. V. Manasis PhD and work on the topic (joint work)
 - Seminar by Prof. Karlis at Roses of U.Reading <https://youtu.be/HZfGTPYeSnQ>

Sports Economics & Competitive balance

What league do we want to see?

- All fans like the fact that a weaker team occasionally wins a game or a league

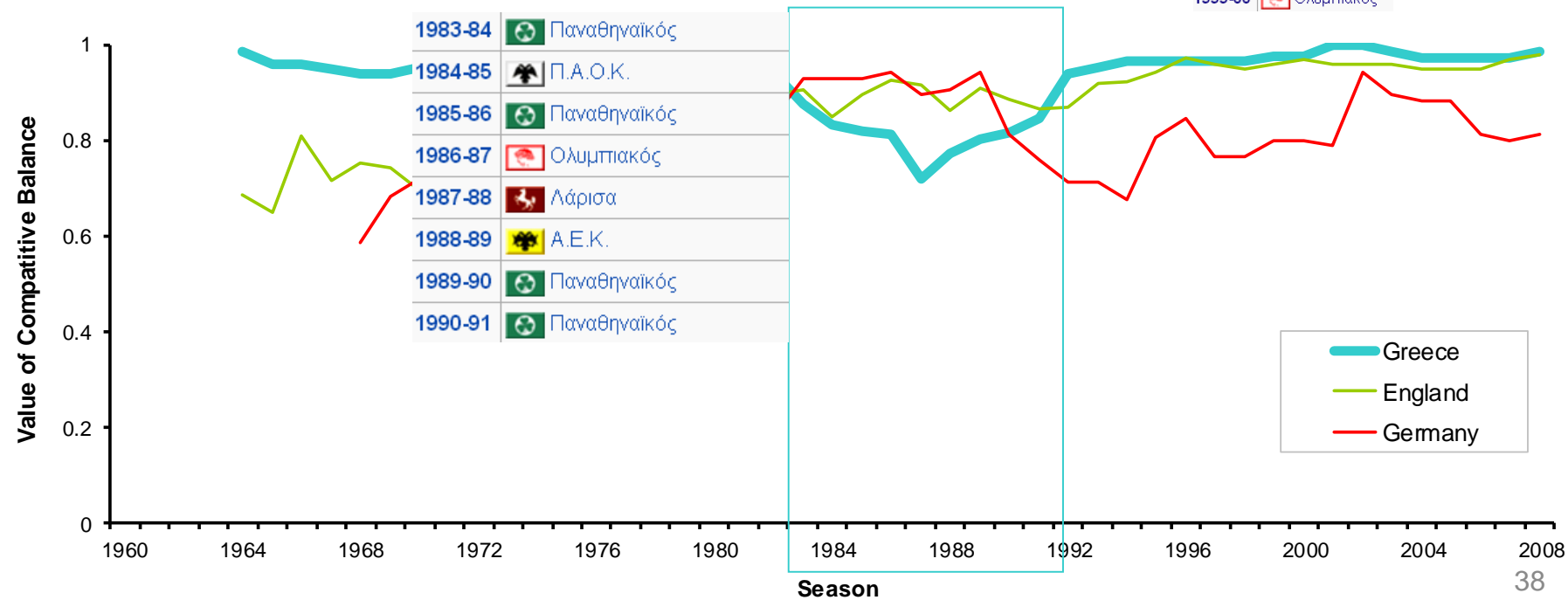
- May neutral fans follow the weakest team
e.g. Greece in Euro 2004

But

- They do not like their team to loose
- They like or they are willing to pay an expensive ticket to see a final with high ranked and expensive teams
e.g. Bayern-Barcelona



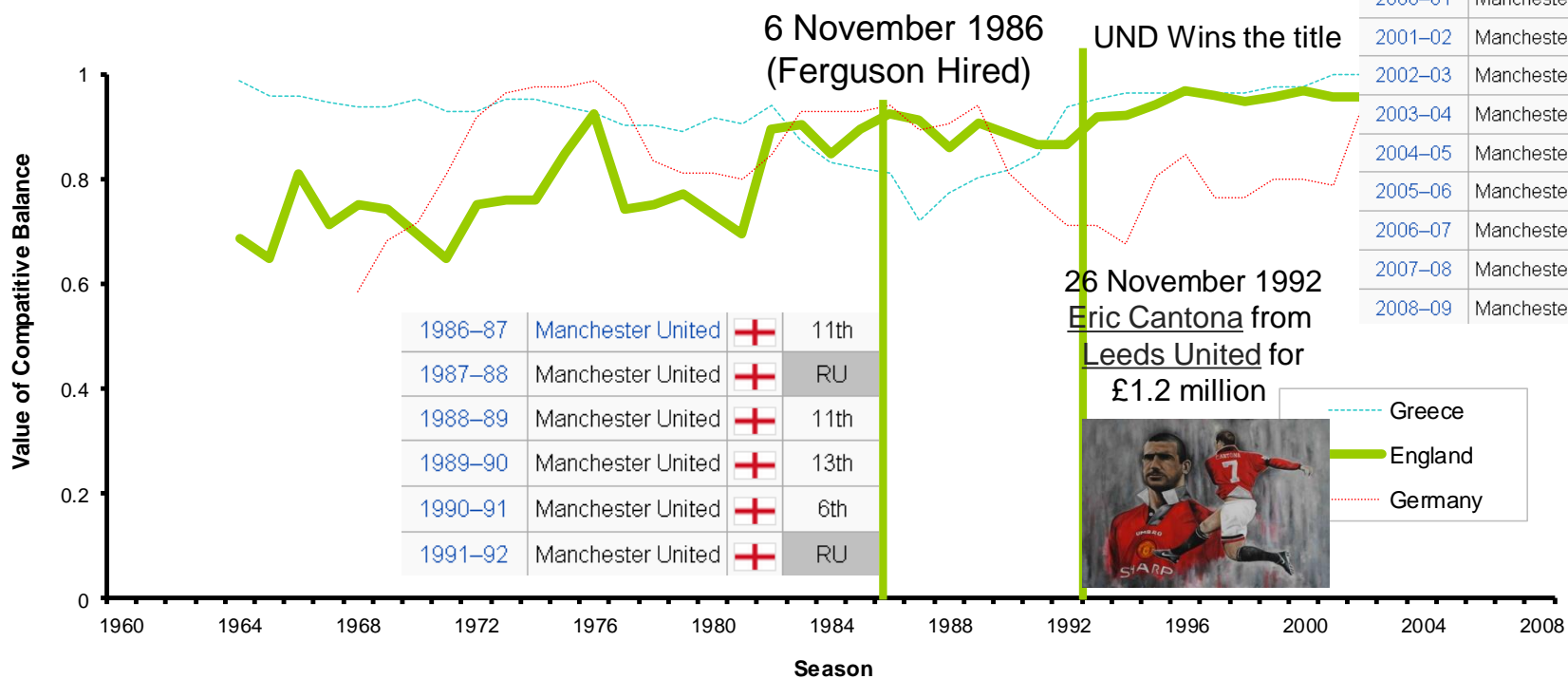
Moving Averages of lag five for DN_1 (Champion) from 1959-20



ManU won 13 out of 17 leagues for the period 1992-2009 and it was not ranked in lower position than 3rd.

3 cases in England ⇒ promoted team ⇒ won the championship:
 Ipswich (1961) & Nottingham (1997) & Leicester (2015-16 – not in the Figure)

Moving Averages of lag five for DN_1 (Champion) from 1959-2008



1992-93	Manchester United	+	W
1993-94	Manchester United	+	W
1994-95	Manchester United	+	RU
1995-96	Manchester United	+	W
1996-97	Manchester United	+	W
1997-98	Manchester United	+	RU
1998-99	Manchester United	+	W
1999-2000	Manchester United	+	W
2000-01	Manchester United	+	W
2001-02	Manchester United	+	3rd
2002-03	Manchester United	+	W
2003-04	Manchester United	+	3rd
2004-05	Manchester United	+	3rd
2005-06	Manchester United	+	RU
2006-07	Manchester United	+	W
2007-08	Manchester United	+	W
2008-09	Manchester United	+	W



— Greece
 — England
 — Germany

Premier League after 13 games of the 2015/16 season (when Leicester won)



Premier League after 13 games
of the 2015/16 season
(when Leicester won)



1968/1969		FC Bayern München
1967/1968		1. FC Nürnberg
1966/1967		Eintracht Braunschweig
1965/1966		TSV 1860 München
1964/1965		SV Werder Bremen
1963/1964		1. FC Köln

1984/1985		FC Bayern München
1983/1984		VfB Stuttgart
1982/1983		Hamburger SV
1981/1982		Hamburger SV
1980/1981		FC Bayern München
1979/1980		FC Bayern München
1978/1979		Hamburger SV

1996/1997		FC Bayern München
1995/1996		Borussia Dortmund
1994/1995		Borussia Dortmund
1993/1994		FC Bayern München
1992/1993		SV Werder Bremen
1991/1992		VfB Stuttgart
1990/1991		1. FC Kaiserslautern
1989/1990		FC Bayern München

2008/2009		VfL Wolfsburg
2007/2008		FC Bayern München
2006/2007		VfB Stuttgart
2005/2006		FC Bayern München
2004/2005		FC Bayern München
2003/2004		SV Werder Bremen

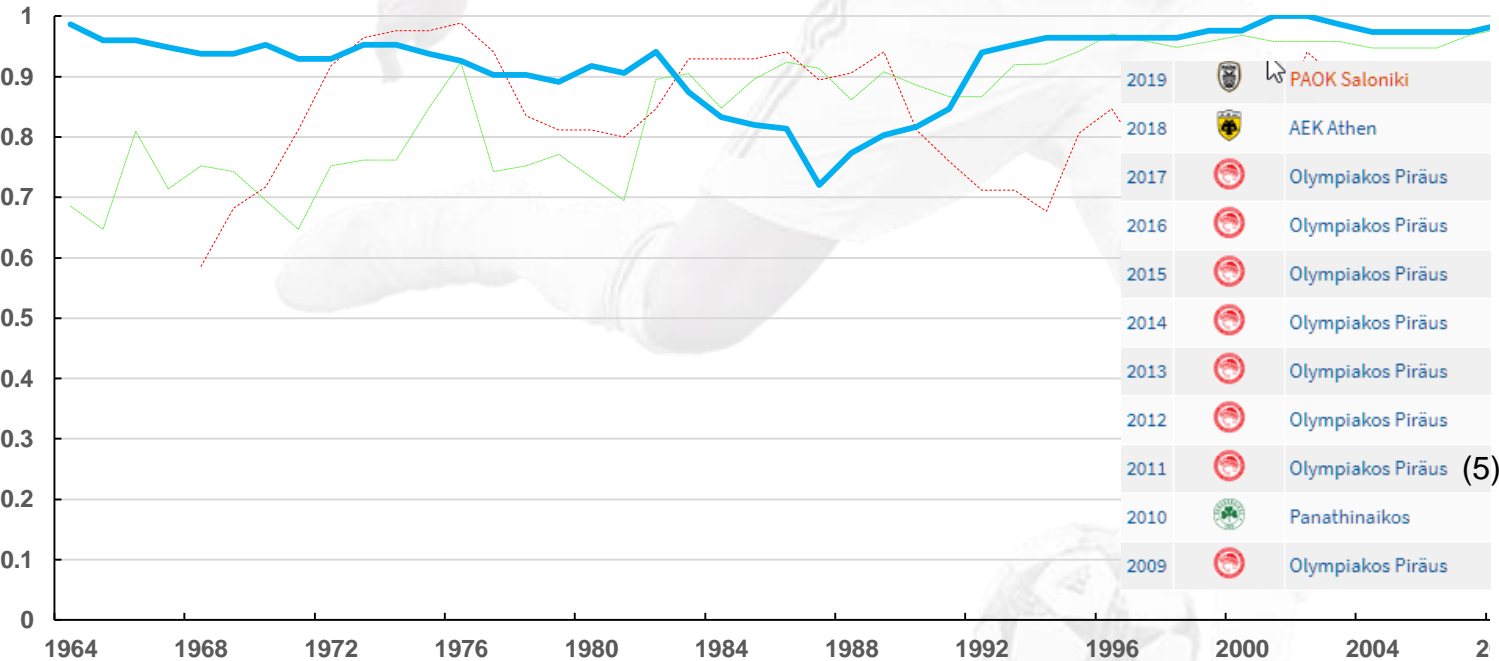
Moving Averages of lag five for DN_1 (Champion)



One case => promoted team => won the championship: Kaiserslautern in 1998

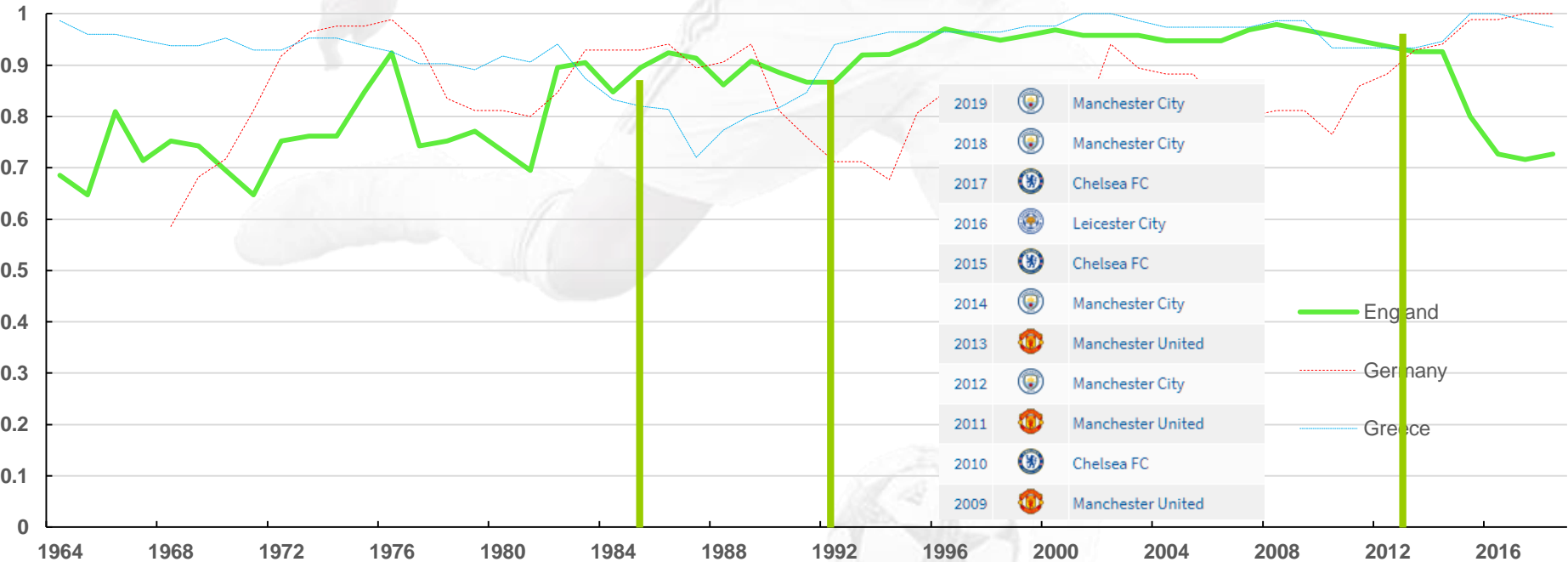
What happened next? Greece: 2009-2019

Moving Averages of lag five for DN_1 (Champion) from 1959/60 - 2018/19



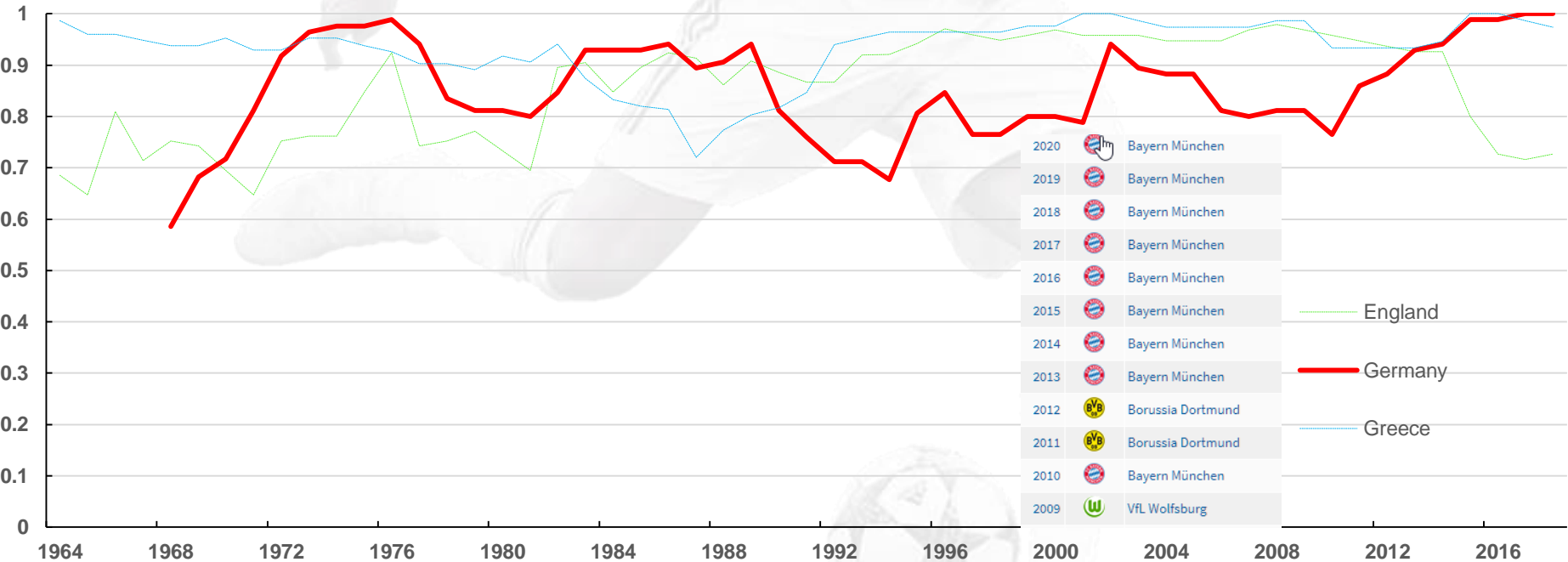
What happened next? England: 2009-2019

Moving Averages of lag five for DN_1 (Champion) from 1959/60 - 2018/19



What happened next? Germany: 2009-2019

Moving Averages of lag five for DN_1 (Champion) from 1959/60 - 2018/19



Concluding remarks

To conclude with

- **Prediction** is important for fans (in terms of betting) \Rightarrow increases profits of bet companies and interest for the sport product (in macro perspective)
- **Inline prediction** is important for fans (in terms of betting) \Rightarrow increases profits of bet companies and interest for the sport product (Media – TV, Radio, Internet).

Concluding remarks

- **Player performance and evaluation** ⇒ Of main interest for: the fans (Player Ranking), Teams (Scouting, Future Performance and Value), Companies (Sponsoring), Players (A lot of money from all previous), Coaches/Managers (Selection of better players)
- **Physical Measurements** (Training and Games): It is related with player evaluation. Main value to help managers/coaches to improve their teams. In macro perspective also the teams financial position is also improving.
- **Scheduling and Competitive Balance**: More Fair and Balanced contests lead to better overall product and more profit to all teams.

**THANK
YOU**



**NO Matter
How Many**

games ~~**Goals**~~

You

predict ~~**Save**~~

People Always

Remember

The One You

Miss.

**WEB SPORTS
ANALYTICS
GROUP**



**INNOVATIVE RESEARCH
& ANALYTICS**