MathSport International

Play-by-play data analysis for team managing in basketball

L. Grassetti, R. Bellio, G. Fonseca and P. Vidoni
Dept. of Economics and Statistics - University of Udine, Italy
E-mail: paolo.vidoni@uniud.it

Athens, July 3rd 2019

MathSport International 2019 Conference, Athens University of Economics and Business – Dept. of Statistics – Athens, July 1–3, 2019

TALK OUTLINE

BASKETBALL ANALYTICS

Two main purposes:

1. match outcome prediction (as in , KS06, M16, RP-C15, S14, , YL12 and , ZMS13)
2. ***analysis of performances:*** team, ***lineups*** and players

Performance analyses can be based on:

- **box score statistics** data
- or more **complex data collections**, such as the advanced play by play data used – in Deshpande and Jensen, 2016 – to study the individual player contribution to the match-winning probability of the team at different game moments (DJ16).
- ...

PLAYER PERFORMANCE LITERATURE ANALYSIS

The leading approach to player performance assessment is based on the so-called adjusted plus-minus (APM) method

- its basic formulation, corresponding to a linear model specification, was introduced in an influential contribution by Rosenbaum (2004 – R04) and recently re-discussed in Omidian (2011 – O11)

but

- the model specification entails **sparse design matrices** and **multicollinearity**.

For this reason

- the *Regularized* APM (**RAPM**) approach was formulated in Sill (2010 – S10)
- the RAPM method typically employs *ridge regression* for the estimation of player efficiency (as summarized in Englemann, 2017 – E17) but **other regularization methods** can also be adopted (Efron and Hastie, 2016 – EH16)

The method was adopted also for the analysis of the players of the *Major League Soccer* (KPM17) and the *National Hockey League* (see MacD11 and MacD12).

EMPIRICAL FRAMEWORK AND PRINCIPAL AIMS

The empirical analyses

- regard the **Italian Serie A basketball league**
- are based on **freely available data** from the first round of the 2018/2019 championship
- adopt a **model-based strategy** mimicking the APM and RAPM literature

The present work **aims** at

A1 focusing on the behaviour/effects of the **lineups**

A2 adopting an **alternative and more informative score computation**

A3 considering a more flexible **model-based approach to regularization**

...

To provide **some guidance** on whether **alternative match strategies** could have been adopted with **better performances**

MORE SPECIFICALLY

We propose to:

A1 change the point of view – from *players performance estimation* to **lineups specific effect**

From **215 players** to **1886 lineups** ⇒ more complex model specification

A2 use of a **performance index rating** (called score in the following)

Value	Relevant events
-1	missed free-throw, turnover or offensive foul
-0.5	missed shot (2 points or three points shots)
0.5	assist
1	steal, offensive or defensive rebound, block, scored free-throw or received foul
2	scored shot
3	scored three-pointer

A3 adopt a different method of estimation

Ridge regression and other *regularization methods* are typically used to face the overparameterization issues in APM estimation.

Our proposal is based on the **empirical Bayes** model-based approach.

ITALIAN BASKETBALL LEAGUE (SERIE A1) CHAMPIONSHIP 2018/2019

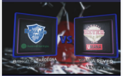
16/06/2019 - 20:45, Impianto: PalaSerradimigni

Banco di Sardegna Sassari 95 - 88 Umana Reyer Venezia

	T1	T2	T3	T4
Singoli q.	23-16	28-25	30-24	14-23
Prog.	23-16	51-41	81-65	95-88

Arbitri: Roberto BEGNIS - Michele ROSSI - Maurizio BIGGI

Highlights



Foto

Accedi al play by play della partita Intera Partita 1° quarto 2° quarto 3° quarto 4° quarto

Banco di Sardegna Sassari

Banco di Sardegna Sassari				Falli		Tiri da 2			Sc	Tiri da 3			Tiri Liberi			Rimbalzi			Stoppate		Palle		Ass	Valutaz.		+/-
	All.Pozzecco Gianmarco	Pt	Min	C	S	R	T	%		R	T	%	R	T	%	Off	Dif	Tot	Dat	Sub	Per	Rec		Lega	OER	
0	SPISSU Marco	9	17	3	0	3	4	75.0	0	1	1	100.0	0	0	0.0	1	2	3	0	0	3	0	3	8	1.13	9
• 2	SMITH Jaime	10	26	0	3	0	5	0.0	0	2	3	66.7	4	4	100.0	0	0	0	0	0	1	2	4	12	0.91	5
• 3	MCGEE Tyrus	11	22	3	4	0	1	0.0	0	1	2	50.0	8	8	100.0	0	4	4	0	0	2	0	1	13	1.22	1
6	CARTER Justin	12	15	2	3	3	4	75.0	1	0	1	0.0	6	6	100.0	2	1	3	0	0	3	4	0	15	1.09	4
8	DEVECCHI Giacomo	0	0	0	0	0	0	0.0	0	0	0	0.0	0	0	0.0	0	0	0	0	0	0	0	0	0	0.00	0
15	MAGRO Daniele	2	6	1	0	1	1	100.0	1	0	0	0.0	0	0	0.0	0	0	0	0	0	2	1	1	1	0.67	6
• 21	PIERRE Dyshawn	12	26	0	3	2	2	100.0	0	2	3	66.7	2	2	100.0	1	1	2	0	0	0	1	2	19	2.00	5
22	GENTILE Stefano	0	15	4	0	0	1	0.0	0	0	0	0.0	0	0	0.0	0	1	1	0	1	2	1	1	-5	0.00	-1
• 25	THOMAS Rashawn	19	31	2	8	5	7	71.4	0	1	4	25.0	6	9	66.7	1	7	8	0	0	3	0	1	23	1.03	7
33	POLONARA Achille	2	15	3	4	1	3	33.3	0	0	3	0.0	0	2	0.0	2	2	4	0	0	1	1	0	0	0.25	2
35	DIOP Ousmane	0	0	0	0	0	0	0.0	0	0	0	0.0	0	0	0.0	0	0	0	0	0	0	0	0	0	0.00	0
• 45	COOLEY Jack	18	27	3	9	6	9	66.7	1	0	0	0.0	6	8	75.0	4	5	9	0	0	2	0	3	29	1.20	-3
	Squadra	0	0	2	0	0	0	0.0	0	0	0	0.0	0	0	0.0	2	0	2	0	0	0	0	0	0	0.00	
	Totali	95	200	23	34	21	37	56.8	3	7	17	41.2	32	39	82.1	13	23	36	0	1	19	10	16	115	1.03	

Used software:

- R Statistics (`Rcore19`)
- `rvest` package (`W16`)
- `stringi` package (`G19`)

The **box score** info are used to

- check the results of the play-by-play data collection
- initialize the lineups construction (using the starting five info)

SERIE A1 – DATA COLLECTION

Banco di Sardegna Sassari	Min.	Umana Reyer Venezia
	1. min	
McGee Tyrus - Rimbalzo difensivo (1)		
		Bramos Michael - Tiro sbagliato da 3 punti
Cooley Jack - Tiro sbagliato da sotto		
		Watt Mitchell - Rimbalzo difensivo (1)
		Watt Mitchell - Tiro sbagliato da sotto
Cooley Jack - Rimbalzo difensivo (1)		
	2. min	
Pierre Dyshawn - Canestro da sotto	2-0	
Thomas Rashawn - Rimbalzo difensivo (1)		
		Haynes MarQuez - Tiro sbagliato da 3 punti
Cooley Jack - Fallo commesso (1)		
Cooley Jack - Palla persa (1)		
		Watt Mitchell - Fallo subito (1)
		Haynes MarQuez - Fallo subito (1)
McGee Tyrus - Fallo commesso (1)		
	2-1	**Haynes MarQuez - Tiro libero segnato**
	2-2	**Haynes MarQuez - Tiro libero segnato**
	3. min	
Thomas Rashawn - Fallo subito (1)		
		Mazzola Valerio - Fallo commesso (1)
McGee Tyrus - Fallo subito (1)		
		Mazzola Valerio - Fallo commesso (2)
Cooley Jack - Assist (1)		
McGee Tyrus - Tiro libero segnato	3-2	
McGee Tyrus - Tiro libero segnato	4-2	
	4-4	**Stone Julyan - Canestro da fuori**
Thomas Rashawn - Tiro sbagliato da 3 punti		
Cooley Jack - Rimbalzo offensivo (1)		
Cooley Jack - Fallo subito (1)		

The **play-by-play** info are used to identify

- **player and team** finalizing the play,
- **intermediate events** (substitutions, time-outs and so on),
- **outcome** of the play (points and scores),
- **quarter**,
- **minute** in the quarter,
- **home and the away teams**,
- identification of **plays** and **possessions**
- **match score and difference in points** (used to determine the "**status**" of the game)

THE DATASET STRUCTURE

play-by-play data

$\Downarrow$

ad hoc **data cleansing** precedure

$\Downarrow$

ready-to-use dataset

The single plays are finally **aggregated by shifts** and the obtained dataset presents the following characteristics:

- **3868 observations** (shifts) from **120 matches**
- the outcome variable is computed from the **home team point of view** and some contextual variables are collected.
- some shifts are excluded (points differential > 20 in the last 5 minutes of the match) $\Rightarrow$ Garbage Time

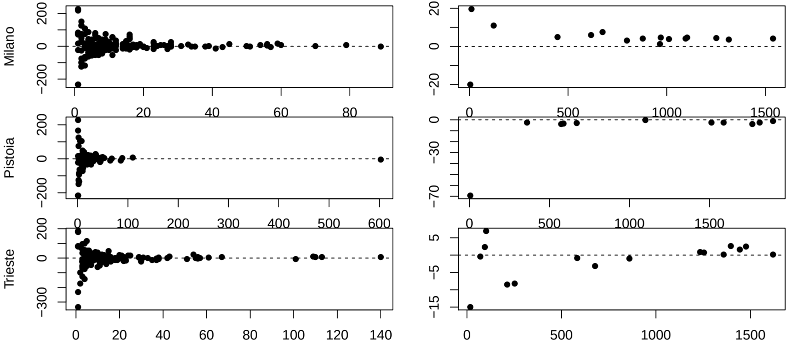
SOME SUMMARY STATISTICS – 1

Number of possessions by **lineups**

Team	No. of Lineups	Min.	1st quart.	Mean	Median	S.D.	3rd quart.	Max.
Avellino	91	1	4.00	25.10	12.00	50.65	23.50	374
Bologna	121	1	5.00	20.56	12.00	30.67	25.00	252
Brescia	115	1	5.00	21.61	11.00	32.02	25.50	238
Brindisi	73	1	6.00	37.58	14.00	69.73	39.00	431
Cantù	100	1	5.00	24.78	9.50	55.23	18.25	459
Cremona	80	1	7.00	33.67	13.00	54.73	37.00	369
Milano	172	1	4.00	14.90	9.50	15.82	18.00	89
Pesaro	60	1	5.75	40.63	10.50	116.91	33.00	888
Pistoia	99	1	6.50	24.00	11.00	62.07	21.00	603
R. Emilia	155	1	5.00	15.04	9.00	18.05	18.50	124
Sassari	180	1	3.00	14.17	7.00	27.47	15.00	224
Torino	137	1	6.00	19.19	11.00	26.03	21.00	214
Trentino	128	1	5.00	19.52	13.00	22.15	29.25	171
Trieste	139	1	5.00	18.22	10.00	23.59	20.50	140
Varese	65	1	5.00	39.35	16.00	118.83	42.00	957
Venezia	171	1	3.50	13.09	7.00	24.85	12.50	244

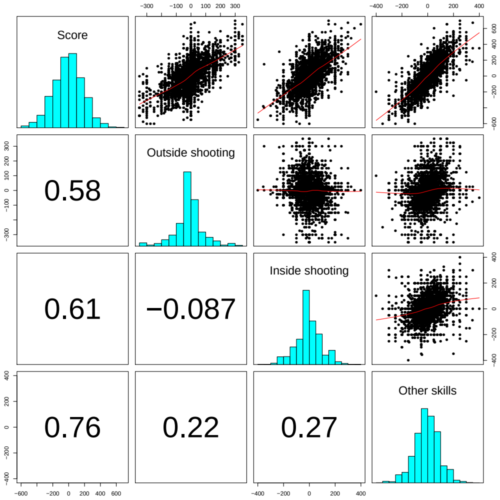
Some summary statistics – 2

Average score of each lineup (left panels) and each player (right panel) for three teams as a function of the number of possessions.



Further details in the score analysis

- The score is a multidimensional measure of players/ lineups performance
- The components are mainly unrelated
- Separated analysis can help focusing on specific characteristics



MODEL SPECIFICATION: LINEUP EFFECTS

Following the classical RAPM literature, the model for the estimation of Lineup effects based on the score response variable is

$$y_t = \beta_0 + \mu_{h[t]} - \mu_{a[t]} + \eta_t \, , \tag{1}$$

where

- t identifies the shifts, $t = 1, \ldots, T$ ($T = 3868$)
- y_t is the **difference between the mean outcomes (scores or points) for the home team and for the away team** for each shift (the mean is over the number of possessions for each team).
- $h[t]$ and $a[t]$ identifies the **lineup for the home and away team** for shift t, respectively

- The model matrix corresponds to a **matrix of signed dummies**
- The model specification could include the effects of some additional **covariates** (using the model-based regularization procedure this is straightforward)
- The estimation is based on **weighted regression** (weights are the total number of possessions in the shift)

MODEL SPECIFICATION: PLAYER EFFECTS

The same kind of model specification can be adopted to estimate the *effects of the players*

- **directly** – changing the model matrix to account for the single players in the shift (each shift entails ten different players)

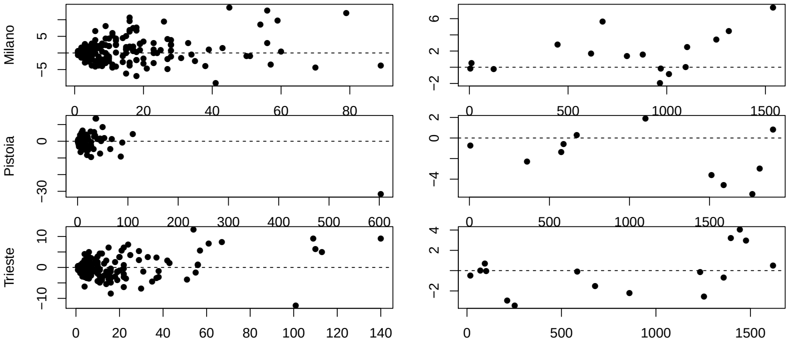
$$y_t = \beta_0 + \sum_{j=1}^{5} \gamma_{h_j[t]} - \sum_{j=1}^{5} \gamma_{a_j[t]} + \eta_t \,, \tag{2}$$

  - η_t denoting a normal error term
  - $\boldsymbol{\gamma}$ is the vector of player effects (with length $M = 212$)
  - $h_j[t]$ and $a_j[t]$ identify the j-th player involved in shift t, for home and away team respectively.

- or **indirectly** – basing on a dataset where y_t is substituted by the estimated lineup effects.

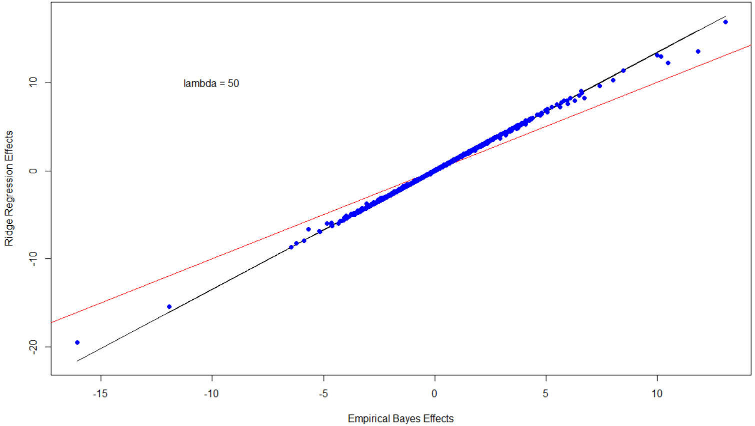
MODEL ESTIMATION RESULTS – 1

Estimated effects of each lineup (left panels) and each player (right panel) for three teams as a function of the number of possessions.



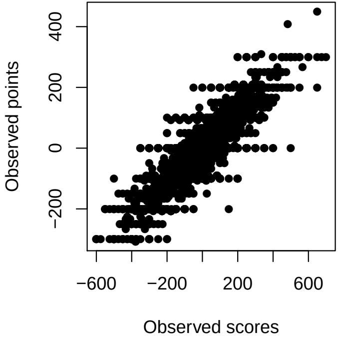
The estimation has been carried out by means of the `hglm` R package `R10`.

THE RELATIONSHIP BETWEEN EMPIRICAL BAYES AND RIDGE REGRESSION APPROACHES

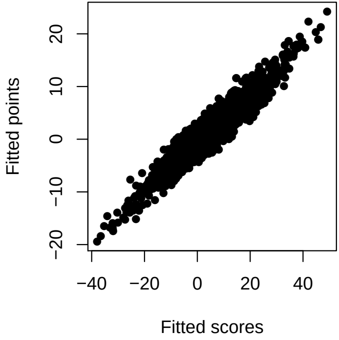


THE RELATIONSHIP BETWEEN POINTS AND SCORES
ESTIMATION RESULTS

Observed **points** vs **scores** for the shift data

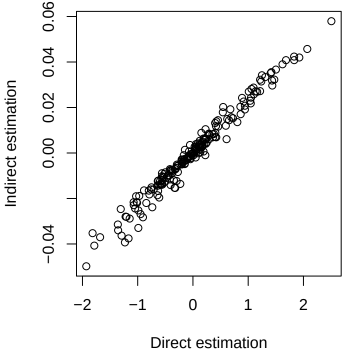


Fitted **points** vs **scores** based on the estimated model for lineup effects

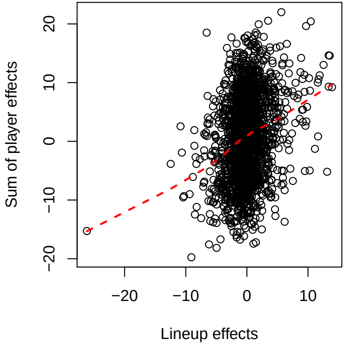


Players or Lineups?

Indirect (two-step) vs **direct estimation** of player effects

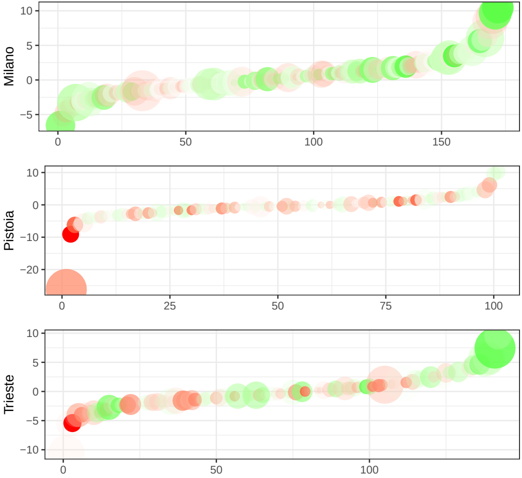


Estimated lineup effects vs **sum of the player effects of each lineup**



USING THE ESTIMATED RESULTS – 1

- Bubble plots for the **sorted estimated lineup effects**
- the **color scaling** denoting the **sum of estimated player effects** (green for higher values, red for lower ones)
- The **bubble size is proportional to the number of possessions** played by lineups.



Using the estimated results – 2

Team rankings based on the estimated lineup effects

Teams	Score-based lineup effect	Rank	Outside shooting	Rank	Inside shooting	Rank	Other skills	Rank
Avellino	0.232	7	0.648	2	-0.290	13	0.029	7
Bologna	-0.217	11	0.204	7	-0.020	8	-0.154	15
Brescia	-0.115	10	-0.582	13	0.136	6	0.053	6
Brindisi	0.336	5	0.415	5	-0.449	16	0.227	1
Cantù	-0.592	13	-0.985	16	0.218	3	-0.075	11
Cremona	0.782	1	0.540	4	-0.058	10	0.226	2
Milano	0.646	2	0.564	3	0.163	4	0.086	5
Pesaro	-1.199	16	-0.600	14	-0.335	15	-0.279	16
Pistoia	-0.655	14	-0.381	12	-0.239	12	-0.128	14
R. Emilia	-0.107	9	-0.106	8	-0.016	7	-0.012	8
Sassari	0.477	3	-0.303	10	0.432	1	0.157	3
Torino	-0.527	12	-0.704	15	0.146	5	-0.101	12
Trento	-0.674	15	-0.259	9	-0.330	14	-0.125	13
Trieste	0.030	8	0.346	6	-0.213	11	-0.018	9
Varese	0.302	6	-0.314	11	0.410	2	0.105	4
Venezia	0.477	4	0.924	1	-0.032	9	-0.022	10

The same kind of analysis can be conducted considering the estimated player effects.

SUMMING UP

The proposed approach

- uses **only freely available data**
- generalises the existing works:
  - using a specific efficiency measure (score vs points)
  - estimating the more informative lineup effects (which also include the player effects)
  - adopting an alternative model estimation strategy (adopting hierarchical generalized linear model specification – Empirical Bayes estimator for the random effects)

CONCLUSIONS

Using the **estimated effects** one can

- determine the net efficiency of the lineups
- splitting the effect into three different aspects of the play
- evaluating also the players net efficiency

These pieces of **information** can be used to

- guide the choice of the lineups that can best face the opposing ones
- determine the estimated team rankings
- compare the different players (considering a net measure of their efficiency)
- predict the outcome of an hypothetical shift during a future game

ONGOING RESEARCH – EUROLEAGUE ANALYSIS

- The website has a **Dynamic Structure**
- More sophisticated methods for data scraping are needed
- **RSelenium** is the way (`H19`)
- Or a *Java script* in **Selenium**

2018-19 | Playoffs | Game 3

PANATHINAIKOS OPAP ATHENS 82

REAL MADRID 89

APRIL 23, 2019 CET: 20:00
LOCAL TIME: 21:00 OLYMPIC SPORTS CENTER ATHENS

Report Video Boxscore Play by play Graphic stats Shooting chart Quotes Photo Gallery

Referees: LAMONICA, LUIGI; BOLTAUZER, MATEJ; MOGULKOC, EMIN
Attendance: 18182

By Quarter	1	2	3	4
Panathinaikos OPAP Athens	18	23	17	24
Real Madrid	23	13	22	31

End of Quarter	1	2	3	4
Panathinaikos OPAP Athens	18	41	58	82
Real Madrid	23	36	58	89

PANATHINAIKOS OPAP ATHENS

							Rebounds						Blocks		Fouls		
#	Player	Min	Pts	2FG	3FG	FT	O	D	T	As	St	To	Fv	Ag	Cm	Rv	PIR
0	THOMAS, DESHAUN	30:46	11	4/5	0/2	3/3	1	3	4			2			3	2	9
5	LANGFORD, KEITH	15:47	10	3/5	1/2	1/2		1	1	2		1			5	3	6
6	**PAPAGIANNIS, GEORGIOS**	14:10	4	2/6			1	2	3			1			1	1	2
10	**PAPAPETROU, IOANNIS**	31:15	9	2/3	1/2	2/4	1	4	5		2		1		2	3	14
14	**GIST, JAMES**	7:28	4	2/3				1	1					1	3		
15	VOUGIOUKAS, IAN	24:52	8	3/7		2/2	2	3	5	2	1		2	1	2	6	17
19	LEKAVICIUS, LUKAS	9:17	3	0/2	1/2									1	3	1	-3
23	KILPATRICK, SEAN	3:12			0/2		1		1								-1
24	**LOJESKI, MATT**	27:40	16	2/3	3/4	3/3		1	1	3					3	2	17
33	**CALATHES, NICK**	31:23	17	3/9	3/7	2/2		2	2	7	3	2		1	2	3	17
43	ANTETOKOUNMPO, THANASIS	2:06										1					-1
44	MITOGLOU, KONSTANTINOS	2:04													1		-1
	Team							1	1								1
	Totals	200:00	82	21/43	9/21	13/16	6	18	24	14	6	7	3	4	25	21	77
				48.8%	42.9%	81.3%											

Head coach: PITINO, RICK

ESSENTIAL BIBLIOGRAPHY – PLUS-MINUS LITERATURE

DJ16 Deshpande, S.K. and Jensen, S.T.: Estimating an NBA player's impact on his team's chances of winning. J. Quant. Anal. Sports, **12**, 51–72 (2016)

E17 Engelmann, J.: Possession-based player performance analysis in basketball (adjusted +/- and related concepts). In: Handbook of Statistical Methods and Analyses in Sports, pp. 231-244. Chapman and Hall/CRC (2017)

KPM17 Kharrat, T., Pena, J.L. and McHale, I.: Plus-minus player ratings for soccer. arXiv preprint arXiv:1706.04943, (2017)

MacD11 Macdonald, B.: A regression-based adjusted plus-minus statistic for NHL players. Journal of Quantitative Analysis in Sports **7.3**, (2011)

MacD12 Macdonald, B.: Adjusted plus-minus for NHL players using ridge regression with goals, shots, fenwick, and corsi. Journal of Quantitative Analysis in Sports **8.3** (2012)

M16 Manner, H.: Modeling and forecasting the outcomes of NBA basketball games. Journal of Quantitative Analysis in Sports, **12**, 31–41 (2016)

O11 Omidiran, D.: A new look at adjusted plus/minus for basketball analysis. MIT Sloan Sports Analytics Conference [online], (2011)

S10 Sill, J.: Improved NBA adjusted +/- using regularization and out-of-sample testing. In: Proceedings of the 2010 MIT Sloan Sports Analytics Conference (2010)

ESSENTIAL BIBLIOGRAPHY – BASKETBALL ANALYTICS

EH16 Efron, B. and Hastie, T.: Computer Age Statistical Inference. Cambridge University Press, Cambridge (2016)

KS06 Kvam, P. and Sokol, J.S.: A logistic regression/Markov chain model for NCAA basketball. Naval Research Logistics (2006)

R04 Rosenbaum, D.: Measuring how NBA players help their teams win. Retrieved from `http://www.82games.com/comm30.htm` (2004)

RP-C15 Ruiz, F.J.R. and Perez-Cruz, F.: A generative model for predicting outcomes in college basketball. J. Quant. Anal. Sports, **11**, 39–52 (2015)

S14 Shen, K.: Data analysis of basketball game performance based on bivariate poisson regression model. Computer Modelling & New Technologies, **18**, 474–479 (2014)

YL12 Yang, J.B. and Lu, C.-H.: Predicting NBA Championship by learning from history data. Proceedings of Artificial Intelligence and Machine Learning for Engineering Design (2012)

ZMS13 Zimmermann, A., Moorthy, S. and Shi, Z.: Predicting NCAAB match outcomes using ML techniques - Some results and lessons learned. Proceedings ECML 2013, 69–78 (2013)

Essential bibliography – R-libraries

G19 Gagolewski M. *et al*: R package stringi: character string processing facilities. `http://www.gagolewski.com/software/stringi/` (2019)

Rcore19 R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

RSA10 Ronnegard, L., Shen, X. and Alam, M.:hglm: A package for fitting hierarchical generalized linear models. The R Journal **2**: 20–28 (2010)

W16 Wickham H.: rvest: easily harvest (scrape) web pages. R package version 0.3.2. `https://CRAN.R-project.org/package=rvest` (2016)

H19 Harrison J.: RSelenium: R Bindings for 'Selenium WebDriver'. R package version 1.7.5. `https://CRAN.R-project.org/package=RSelenium` (2019)

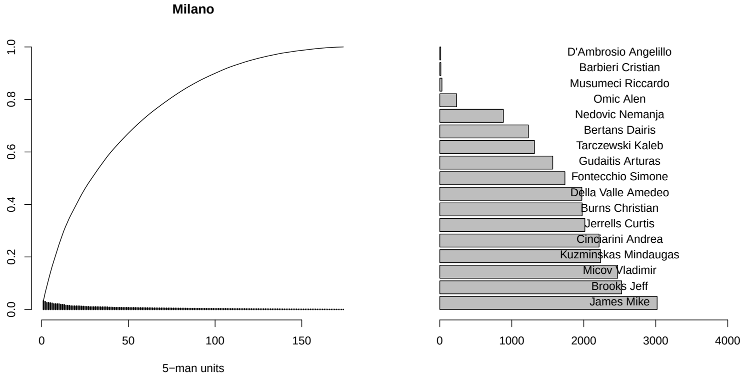
Thank you for your attention

SOME SUMMARY STATISTICS - NUMBER OF POSSESSIONS BY **players**

Team	No. of Players	Min.	Mean	Median	S.D.	Max.
Avellino	14	3	815.71	657.50	666.79	1927
Bologna	13	21	956.92	880.00	608.99	1761
Brescia	13	2	955.77	1001.00	505.51	1730
Brindisi	12	3	1142.92	1128.50	805.88	2223
Cantù	12	117	1032.50	879.00	791.59	2010
Cremona	12	5	1122.50	1307.00	742.83	2055
Milano	16	5	800.62	922.50	461.15	1539
Pesaro	10	98	1219.00	1383.00	845.07	2188
Pistoia	11	6	1080.00	1100.00	668.12	1897
R. Emilia	18	1	647.50	652.50	430.69	1276
Sassari	13	105	981.15	1122.00	606.12	1830
Torino	14	104	938.93	883.50	498.43	1911
Trentino	12	135	1041.25	1219.00	505.45	1648
Trieste	16	18	791.25	769.00	602.47	1620
Varese	12	1	1065.83	821.50	810.66	2224
Venezia	14	17	799.29	745.00	501.91	1482

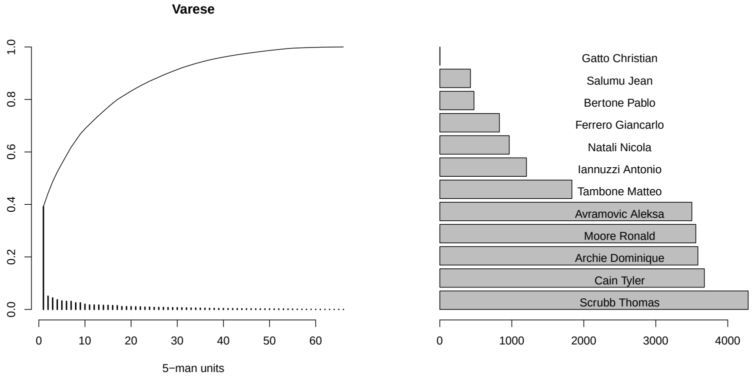
THE DISTRIBUTION OF THE NUMBER OF PLAYS – 1

Distribution of the number of plays for Milano team.



THE DISTRIBUTION OF THE NUMBER OF PLAYS – 2

Distribution of the number of plays for Varese team.



USING THE ESTIMATED RESULTS – 3

Team rankings based on the estimated player effects

Teams	Score-based player effect	Rank	Outside shooting	Rank	Inside shooting	Rank	Other skills	Rank
Avellino	0.300	7	0.569	4	-0.149	11	0.056	7
Bologna	-0.378	11	0.459	5	-0.042	7	-0.572	15
Brescia	-0.237	10	-0.763	14	0.129	6	0.192	6
Brindisi	0.382	6	0.351	6	-0.337	14	0.498	3
Cantù	-1.099	13	-1.316	16	0.264	4	-0.234	11
Cremona	1.241	4	0.644	3	-0.065	10	0.618	2
Milano	1.747	1	0.854	2	0.269	3	0.467	4
Pesaro	-1.803	16	-0.575	11	-0.332	13	-0.688	16
Pistoia	-1.694	15	-0.728	13	-0.423	16	-0.530	13
R. Emilia	-0.137	8	-0.000	8	-0.051	8	-0.024	8
Sassari	1.476	2	-0.687	12	0.695	1	0.895	1
Torino	-1.078	12	-1.086	15	0.176	5	-0.344	12
Trento	-1.468	14	-0.291	10	-0.412	15	-0.562	14
Trieste	-0.170	9	0.305	7	-0.205	12	-0.127	10
Varese	0.436	5	-0.262	9	0.398	2	0.233	5
Venezia	1.408	3	1.845	1	-0.063	9	-0.114	9