ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF STATISTICAL SCIENCES

"PAOLO FORTUNATI"

First Cycle Degree / Bachelor in Statistical Sciences

Curriculum Stats&Maths

FACTOR ANALYSIS FOR FOOTBALL PLAYERS' PERFORMANCES

(*Multivariate Analysis*)

Presented by:

Emanuele Tartaglione

0000883073

Supervisor:

Prof. Angela Montanari

SESSION I

ACADEMIC YEAR 20 / 21

To my grandfather, Benito, who gave me the passion for football.
To my parents, who gave me the passion for mathematics.
To my family, my friends and my girlfriend.

INDEX:

1. INTRODUCTION

In the last years, statistics, data analysis and data science have been spreading in the football world. The use of the statistics and data analysis in the sports, called "sport analytics" did not begin with football though. It started in the MLB (Major League Baseball), the whole world began to familiarize with the term "sport analytics" with the book: "Moneyball: The Art of Winning an Unfair Game", written by the journalist Michael Lewis, where he explains the Billy Bean's method. He was the General Manager (GM) of the Oakland Athletics from 1998 to 2016 [1]. In his first years as GM built a successful team on a minimal budget using statistics. Many statisticians refer to Billy Bean as the progenitor of the sport analytics.

Then it began to be used also in NBA (National Basketball Association), NHL (National Hockey League) and only afterwards in football.

There are two types of sport analytics, the on-field, and the off-field one. The former analyses the performance of the players on the pitch, in the football case: how many goal scoring opportunities where created, how many passes were performed, etc… The latter regards the business part, how to increase profitability, from the merchandise to the price of tickets.

The most used and, at the same time, the most important statistical measurement in football analytics is the xG (abbreviation that stands for "Expected Goals"). It is the most innovative metrics in the football performance analysis. Used to calculate the quality of score chances, it is the likelihood of a shot to be a goal. Through a complex dataset that contains all the goals of the past years in the principal leagues, the type of the shot is analysed. Several factors are considered: the type of assist, if the ball is kicked with the dominant foot or not (or the head), from which position the player hits the ball, etc… [2].

The use of the xG was a revolution in the football analysis. It was seen that, generally, the teams nowadays tend to shoot from closer position than the past. It is due to the introduction of this new metric, the "xG". The clubs noted that is it worth trying to get as close to the goal as possible before shooting. The xG is used to assess the players: a player that creates more chances has more opportunities to score a goal.

However, the clubs are not the only one to rely on data analysts. Recently, the football player K. De Bruyne decided to abandon his agent and rely on a team of data analysts to discuss his new contract with Manchester City [3]. The team showed his importance in the field and how he is crucial for his

team and even how he will be important in the future. He will now get paid £400.000 per week, an increase of £100.000 with respect to the previous salary.

This could be a revolution in the sports world, less power to the agents and more power to the statisticians! It may offer more job opportunities for analysts, moreover the data will talk for the players.

In the following paper I will analyse the performance of the midfielders of the "Serie A TIM", the major Italian Football League. The main idea is to apply factor analysis to establish which player performed better in the season 19/20, then compare the market values of the top players and see how they changed at the end of the season.

2. DATASET DESCRIPTION

The dataset contains information about the midfielders of the "Serie A TIM", the major Italian League. Data have been downloaded, in date 18/03/2021, from the website WyScout, a professional platform born in 2004 that contains statistics of players, matches and teams worldwide [4].

TABLE 1: Players' Variables

1. Player	12. xA	23. Red cards per 90
2. Team	13. Duels won, %	24. Head goals
3. Position	14. Foot	25. Shots
4. Age	15. Height	26. Shots on target, %
5. Market value	16. Weight	27. Goal conversion, %
6. Contract expires	17. Defensive duels won, %	28. Accurate crosses, %
7. Matches played	18. Areal duels won, %	29. Successful dribbles, %
8. Minutes played	19. PAdj Sliding tackles	30. Offensive duels won, %
9. Goals	20. PAdj Interceptions	31. Accurate passes, %
10. xG	21. Fouls per 90	
11. Assist	22. Yellow cards per 90	

The data variables are illustrated in Table 1. Some variables are easily understandable. Others need an explanation; we follow the definitions given by the WyScout Glossary [5]:

- Team: it is the current team of the players and not the one where he played the last season, thus in the results one or more teams not actually present in the Serie A TIM can appear.

- Age: current age of the players.
- xG: “Expected Goals”.
- xA: “Expected Assists”.
- PAdj Sliding tackles: PA stands for “Possession adjusted”. This is done because a player can contribute defensively only when his team is not in possession of the ball, but the other is. For example, if two players have done both 10 tackles in a match, the one whose team has got less possession of the ball will have a higher PAdj Sliding tackles than the other.
- PAdj Interceptions: same reasoning of the PAdj Sliding tackles.
- Goal conversion: percentage of shots resulted in a goal.

The list of the variables is not exhaustive, there are a lot of variables that may have had been considered, for example: accelerations, offsides, touches in the box, second assists, etc… I have chosen the most relevant ones, in my opinion.

Among all the players in the dataset, initially 249, only those that played at least 15 games and 1000 minutes in the season were selected to avoid a biased analysis. It is not worth to compare a player that played only a few games with one that played every match in the season. This led to a final dataset of 141 players…

The players chosen were only the players whose main role on the pitch is midfielder, this means that players, whose main position is not midfielder but can play as midfielder, were not chosen. For example, the website WyScout classifies as forwards players for instance P. Dybala or D. Kulusevski, who can play different positions (forward and offensive midfielder). As they play mainly as forwards, they were not chosen.

3. METHODOLOGY

My analysis was mainly performed with factor analysis (FA). My main aim is to see which were the best midfielders in the season 19/20. It is possible using an FA model to divide the variables in different factors that can be, for instance, offensive skills, defensive skills, etc… whereas the Thompson scores allows to see at the players’ ranking for each latent variable.

I ran a model including all the variables but the variables from 1 to 8 and from 14 to 16 (see Table 1) and I determined the number of factors required in order to explain the correlations in the observed data. Then, for each player, I estimated Thompson scores from the FA and used them to

create an overall ranking. Finally, I used the ranking to see which players performed better and compare their market values at the beginning and at the end of the season. To conclude my analysis, I calculated the sensitivity of some variables, seeing which influence they had in the original model. All the analysis were performed through the statistical software "R".

3.1. FACTOR ANALYSIS

It is a statistical technique aimed at explaining the correlation among many observed variables, in terms of a small numbers of unobserved, variables, called factors.
The origin of the FA is attributed to Spearman (1904), a psychometrician that was involved in studying the human intelligence. He observed the performance of his students and considering the correlation matrix between children's examination in Classics, French, and English, he noticed a very high correlation among the variables. He thought that it was due to a hidden, unobservable factor, that he called general ability or intelligence. The generalization of Spearman's model can be written in the following way:

$x_i = \Lambda_{i1} f_1 + \Lambda_{i2} f_2 + \cdots + \Lambda_{im} f_m + u_i + \mu_i$

Which, in matrix form becomes:

$$x = \Lambda f + u + \mu$$

where x is a p-dimensional random vector with expected value μ and covariance matrix $\sum$; Λ is the $p \times m$ factor loading matrix; f is the $mx1$ random vector of common factors and u is the $p \times 1$random vector of unique factors. Without loss of generality, assuming to deal with mean centred x variables, the μ can be omitted from the model: $x = \Lambda f + u$.

In order to reduce indeterminacy, some constraints may be applied:

- $E(f) = 0 \quad \& \quad E(u) = 0$: they mean that we work with mean centred data.
- $E(ff^T) = I$: the variances of the common factors are 1 and their covariances are 0.
- $E(uu^T) = \Psi$, where Ψ is a diagonal matrix. This mean that the unique factors are uncorrelated and may be heteroscedastic.
- $E(fu^T) = 0 \quad \& \quad E(uf^T) = 0$: the common factors are uncorrelated with the unique factors.

If these constraints hold, the covariance matrix $\sum$ can be decomposed as $\sum = E(xx^T) = \Lambda\Lambda^T + \Psi$.The diagonal elements can be also written as: $Var(x_i) = \sum_{ii} = \sum_{k=1}^{m} \lambda_{ik}^2 + \psi_{ii} = h_{ii}^2 + \psi_{ii}$

where the quantity $\sum_{k=1}^{m} \lambda_{ik}^2$ is the communality: the part of the variance of the observed variables that is explained by the common factors; while $\psi_{ii}$is the unique variance: the part explained by the unique factors.

When dealing with standardized variables the communalities are constrained to be between 0 and 1 as they explain the variance, and the uniquenesses are computed as $1 - h_{ii}^2$.

If the factor model holds, the factor lading matrix Λ also represents the covariance between the observed variables x and the common factors f . $Cov(x, f) = E(xf^T) = \Lambda$.

The "R" function used in this project is called "factanal". It works with standardized data, so it considers the correlation matrix and not the covariance matrix. This function performs maximum-likelihood factor analysis, meaning that the observed variables x are assumed to be distributed according to a multivariate normal distribution. I decide to add the "varimax rotation" to the model. The rotations improve the interpretability of the factors. This method changes the role of the variables in the model.

Once we run a model and we look at how many hidden factors are enough to explain the observe correlations, we could be interested in determining who performed best for each factor. This means that we can estimate a vector of factor scores f_j for each observation x_j.

For this analysis I used Thompson scores. Thompson's method defines the factor scores as a linear combination of the observed variables chosen to minimize the squared expected prediction error. Furthermore, it produces standardized, factor scores. I choose this method because it maximizes the "validity" of the estimates, even if the estimates are not unbiased. Once I estimate the factor scores, I create an overall ranking.

3.2. RANKING ESTIMATION

The main goal of this analysis is to compare the players' performances creating an overall ranking for each (j-th) player, including all the variables (Goals, xG, Assists, …). To do that, I calculated a weighted mean of Thompson scores, considering as weights the sum of squared loadings (SS loadings), that determines the value of each factor. Briefly, the formula is the following:

$$Ranking_j = \frac{\sum_i F_{ij} L_i}{\sum_i L_i} \qquad (1)$$

where F_{ij} is the Thompson score of the i-th factor referred to the j-th player and L_i is the value of the SS loadings of the i-th factor.

4. RESULTS

Before applying any kind of statistical technique, I assume that there are some correlations among the variables. The goals and the shots or the accuracy of the crosses and the assists can be correlated variables. I had a look at the correlation matrix.

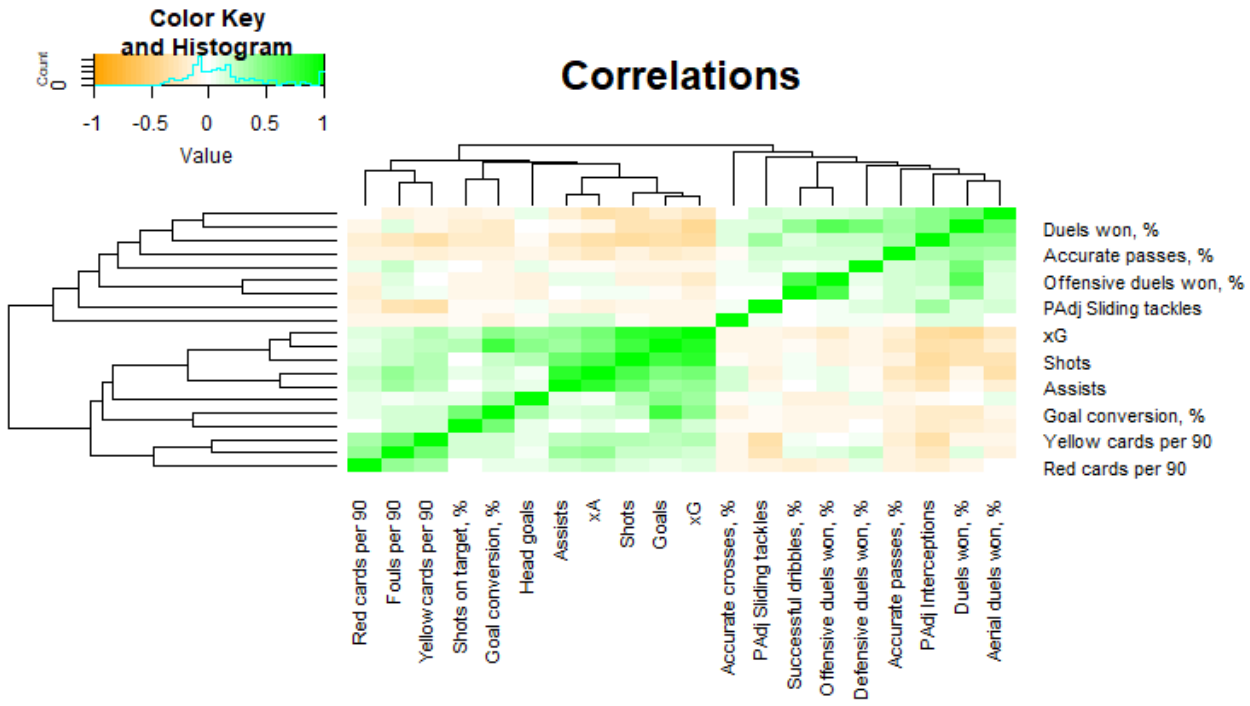

Fig.1: Correlations among all the variables

Figure 1 shows the correlation matrix, where it can be noticed that some variables are correlated, thus I decided to perform the FA: I ran a six factors model and the p-value, the probability associated too the test of the hypothesis that six factors are enough to explain the observed correlations, is 0.0748, therefore the hypothesis is not rejected. Moreover, the SS loadings are greater than 1, meaning that all the factors are significant. Theoretically, for a factor to have a positive reliability, the SS loadings must be greater than one (Kaiser's rule) [6].

```
                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
SS loadings       2.913   2.179   2.026   1.764   1.588   1.342
```

In the Fig. 2 it is shown the factor configuration after varimax rotation and the effect of the variables on the factors.

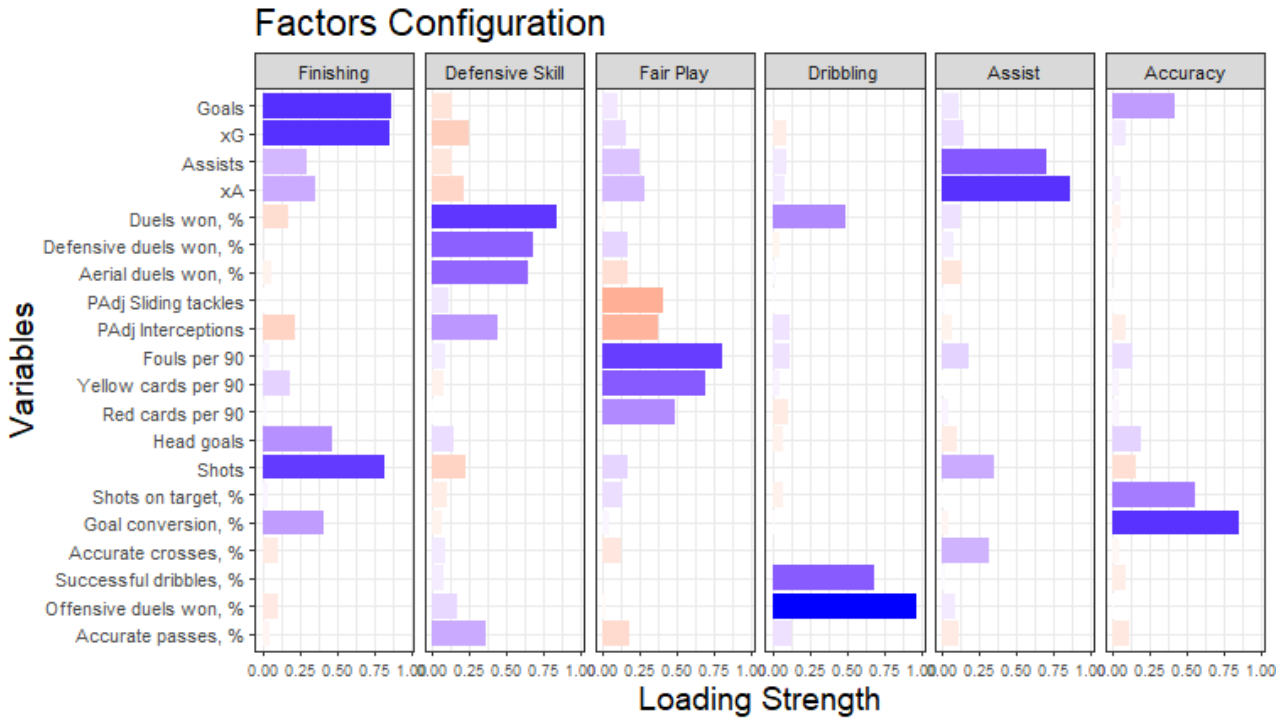

Fig. 2: Factors Configuration

An important thing to consider in the factor configuration is that the midfielders have different roles on the pitch, for sake of simplicity, they are divided in 4 parts: defensive midfielders, central midfielders, attacking midfielders and wingers. They have different qualities and tasks: defensive skills, offensive skills, a mix of the two, etc… depending on the position on the pitch. In this analysis I tried to compare all the midfielders regardless their position on the field.

I try to give a meaning to the factors (Table 2): the first factor seems to represent the attacking skills, of course they are the most important because a player is good if he scores goals and does assists, even if he is very bad at defending. The variables common to the second factor are the defensive skills and the passing accuracy, the opposite of the attacking skills. The other factors express, in order: aggressivity, offensive qualities, capability to be an assist man and shooting skills. I named the factors as: finishing, defensive skill, fair play, dribbling, assist and accuracy.

TABLE 2: Variables that most influence each factor

Factor number	Common variables
1.	Goals, xG, Assists, xA, Shots, Goal conversion, Head goals
2.	Duels won, Defensive duels won, Aerial duels won, Interceptions, Accurate passes
3.	Yellow cards, Red cards, Fouls
4.	Successful dribbles, Offensive duels won, Duels won
5.	Accurate crosses, Assists, xA
6.	Goals, Shots on target, Head goals, Shots on target, Goal conversion

Figure 3 shows how each variable influences the factors. We can see that the most powerful variables are Goals, xG, Assists, xA, Duels won and Shots. There are other variables that have a low effect, such as Sliding tackles, that mostly covers only the Fair Play factor, or Accurate crosses, covering only the Assist factor. Later in the analysis (in 4.2.) I will show the sensitivity of the fouls, the yellow cards, and the red cards: the variables influencing most the Fair Play factor.

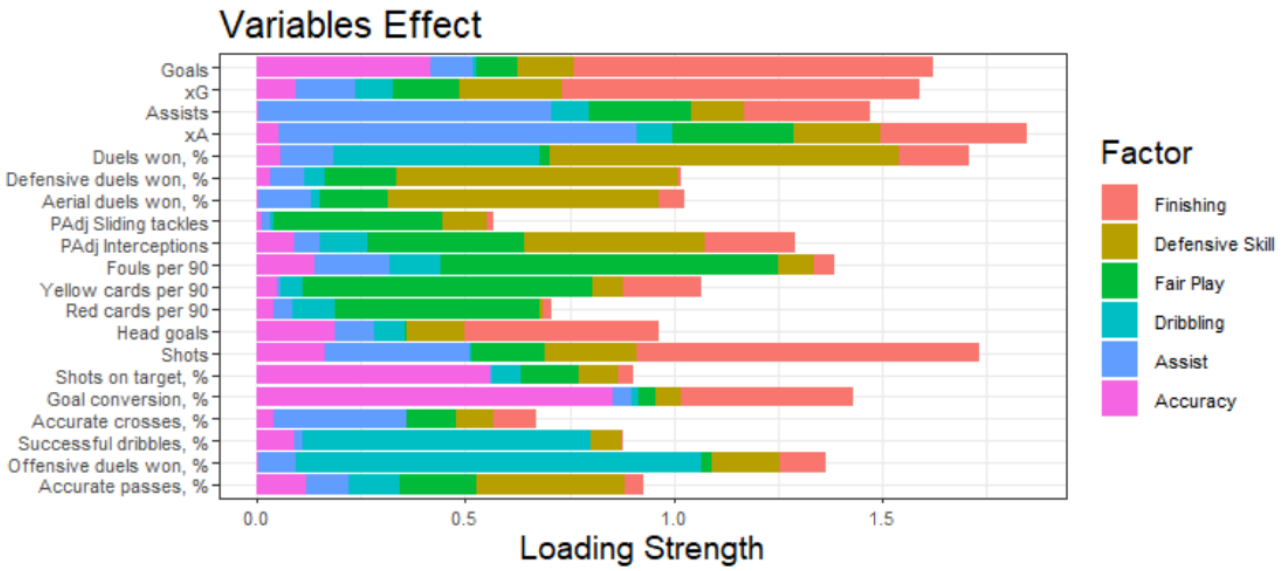

Fig. 3: Factor loadings

4.1. PLAYERS' PERFORMANCES AND MARKET VALUES

My main goal is to create a ranking of the players: firstly, I look at the factors ranking using the Thomson estimators as mentioned in the paragraph 3.2. Table 3 shows the final ranking.

TABLE 3: The Best 6 players in the Serie A TIM 19/20

	Player	Team	Age	Ranking
1.	A. Gómez	Sevilla	33	1.276
2.	R. Gosens	Atalanta	26	1.100
3.	Luis Alberto	Lazio	28	0.921
4.	E. Pulgar	Fiorentina	27	0.851
5.	M. Mancosu	Lecce	32	0.818
6.	J. Kurtic	Parma	32	0.805

Table 3 shows the top 6 best performing players in the 19/20 season, according to the ranking that I created. A. Gómez is the best player according to my ranking and he also won the Serie A Awards as best midfielder that year [7]. R. Gosens, Luis Alberto and A. Gómez were also in the Serie A Team of the Year that season [8].

The last season A. Gómez played for Atalanta and J. Kurtic played half season for SPAL and the other half for Parma, but here it is shown the name of the current team.

Table 4 displays the statistics of these players (for example considering Goals, xG, Assists and xA, basically the characteristics that most influenced the first factor):

TABLE 4: Goals and Assists of the 6 best players of the Serie A TIM 19/20

	Player	Goals	xG	Assists	xA	Ranking
1.	A. Gómez	7	8.10	13	12.92	1.276
2.	R. Gosens	9	5.74	6	3.53	1.100
3.	Luis Alberto	6	8.03	12	10.27	0.921
4.	E. Pulgar	7	7.30	6	7.80	0.851
5.	M. Mancosu	14	15.62	2	2.13	0.818
6.	J. Kurtic	4	3.16	1	4.47	0.805

We can see that these players' statistics are very good. It means that the computation of the ranking worked well. The only player with the statistics, listed in Table 4, under the average is J. Kurtic. It would be good to explore the other variables of this player.

Duels won, % 55.5	Defensive duels, % 58.0	Aerial duels won, % 54.6
Successful dribbles, % 63.0	Offensive duels won, % 56.1	Accurate passes, % 82.9

Looking through the variables, I chose to report here only the better ones. He has very good statistics regarding the percentage of duels won and pretty good passing accuracy. This explains why he is in a high position (6^{th}) in the final ranking even though he did not score a lot of goals and he gave only one assist to his teammates.

An interesting feature to observe is that the players' ranking is distributed according to a normal with mean=0 and variance=0.172 (Fig. 4). It is almost obvious that it is a normal, because in the computation of the scores using the factor analysis, the 'factanal' algorithm supposes that the factors are multivariate normally distributed.

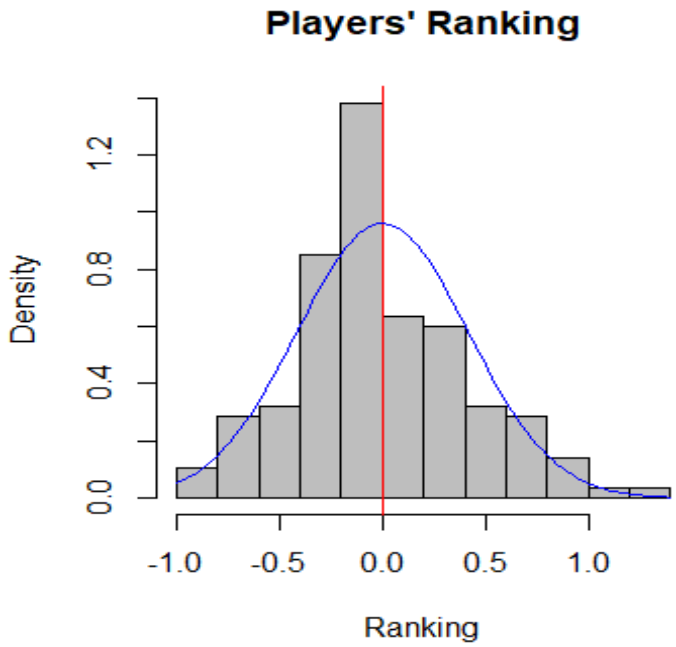

Fig. 4: Distribution of the Ranking

The mean=0 allows to divide the players: good and bad performers.

Among all the players, I want then to have a look at the ranking of the young players, being currently less than 24 years old (Table 5). I added two values: MV2019 and MV2020, the market values respectively on June 6th, 2019, and in August 25th, 2020. These two data were drawn from the website Transfermarkt.com [9].

TABLE 5: The best 6 young players of the Serie A TIM 19/20

	Player	Team	Age	Ranking	MV2019 (in mln €)	MV2020 (in mln €)
1.	J. Boga	Sassuolo	24	0.596	6.0	25.0
2.	R. Orsolini	Bologna	24	0.501	15.0	22.0
3.	M. Pessina	Atalanta	23	0.321	1.5	12.0
4.	Fábian Ruiz	Napoli	24	0.266	50.0	50.0
5.	M. Svanberg	Bologna	22	0.180	4.0	5.0
6.	A. Diawara	Roma	23	0.137	15.0	20.0

The best three players: J. Boga, R. Orsolini and M. Pessina did an impressive season, and as expected, their market value increased significantly (Table 5). Pessina was playing for Hellas Verona on a loan, he showed off and Atalanta called him back from the loan. J. Boga and R. Orsolini remained respectively at Sassuolo and Bologna, but it would have been a good choice to sell them if an offer greater than their market value arrived.

Fabián Ruiz did a better season than the mean of the midfielders, but its market value did not increase. There exist other factors that influence the players' value, one of them is the blazon of the team. In the last season Napoli classified 7th and thus did not reach a placement to be part of the Champions League (only the first four take part in the Champions League. This had a great impact on the blazon of the team and consequently on the players' team value.

The football market is a worldwide business, the mechanism behind it is complex, but I decided to examine two transfers, the first one of Suso, from Milan to Sevilla, and the latter of Allan, from Napoli to Everton.

Jan 29, 2020, Suso decided to Join Sevilla. The formula was a loan of 1 million, with a redemption of 21 million circa.

Player	Team	Age	Ranking
Suso	Sevilla	27	0.425

Suso's MV in that period was around 30 million [9], so it should have been a good investment considering the positive ranking and the age at that time (26).

Another thing to consider is that at that time, Milan was having a negative budgetary, and it may have been a deal that has benefited both clubs.

I do not want to enter in merit about market decisions of clubs because there are a lot of reasons to sell or buy a player for a market value that could be less or higher than the true one. Maybe a club is in financial trouble, or a player has issues with the teammates or with the coach. The salary of a player also plays an important role in the decision of buying or selling him.

From what the journalists report, in January, Allan had a discussion with Napoli’s president, and he was sold to Everton in the summer.

Player	Team	Age	Ranking
Allan	Everton	30	-0.011

Purely from Allan's performances, it would have been good to buy Allan for a value close to his true value. The transfermarkt value estimate was around 28 million and Everton bought him for 25 million circa. I would say that it was a good deal for Everton.

4.2. SENSITIVITY OF THE VARIABLES

In the model I ran, there are some variables that may be not useful to classify the skill of the players: fouls per 90, yellow card per 90 and red card per 90. These three variables are “behavioural variables”, and not always have a negative influence, sometimes it may be useful to do a foul or to take a yellow card for a good reason. I tried to run a model without these variables, calculate again the ranking according to the new model using the Eq. (1) and then look at the Ranking Difference (new ranking – old ranking).

I ran a model with 5 factors because according to the factor configuration of the first model, I removed the variables that most influenced the Fair Play factor. The new model presented a p-value of 0.00165, less than the acceptance rate of 0.05.

Although the p-value is low, this model follows the Kaiser’s rule [6], thus it is reliable, and I decided to use it to make comparisons with the 6-factors model

	Factor1	Factor2	Factor3	Factor4	Factor5
SS loadings	2.855	2.320	1.718	1.708	1.349

Fig. 5 shows the configuration of the second model.

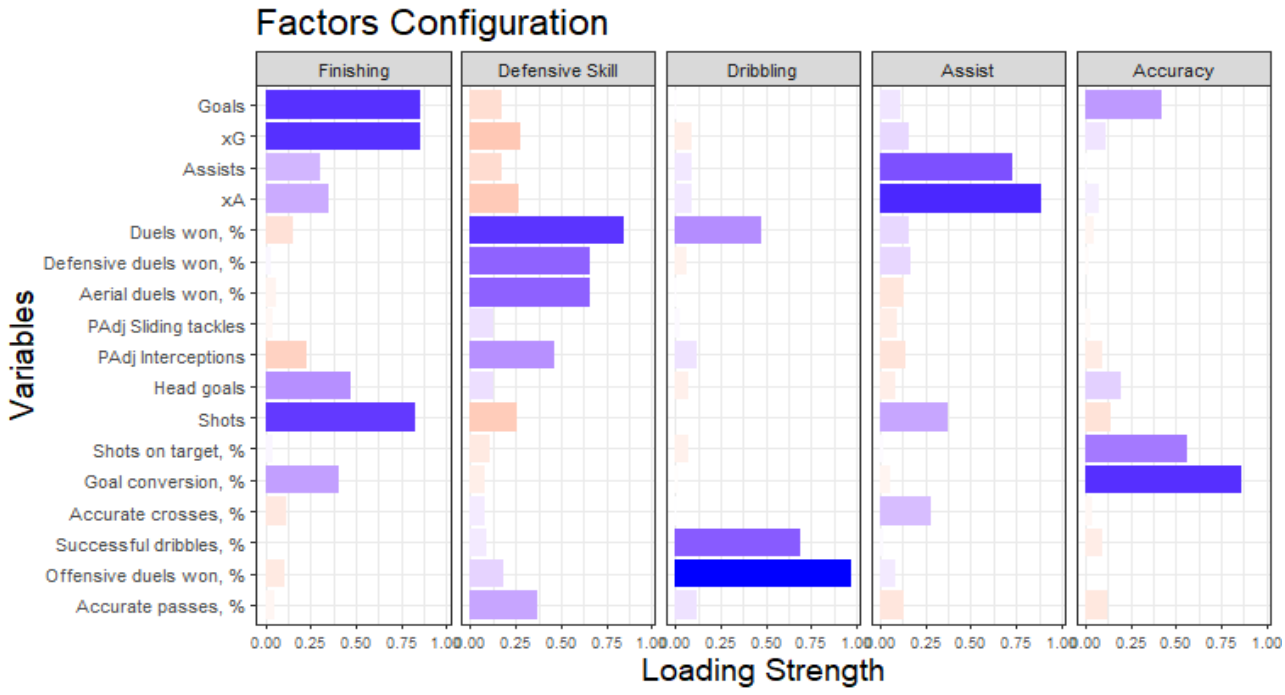

Fig. 5: Factors Configuration of the second model

We can see (Fig. 5) that is very similar to the first factor configuration, without the Fair Play factor, so I decide to accept it as the model to calculate again the ranking.

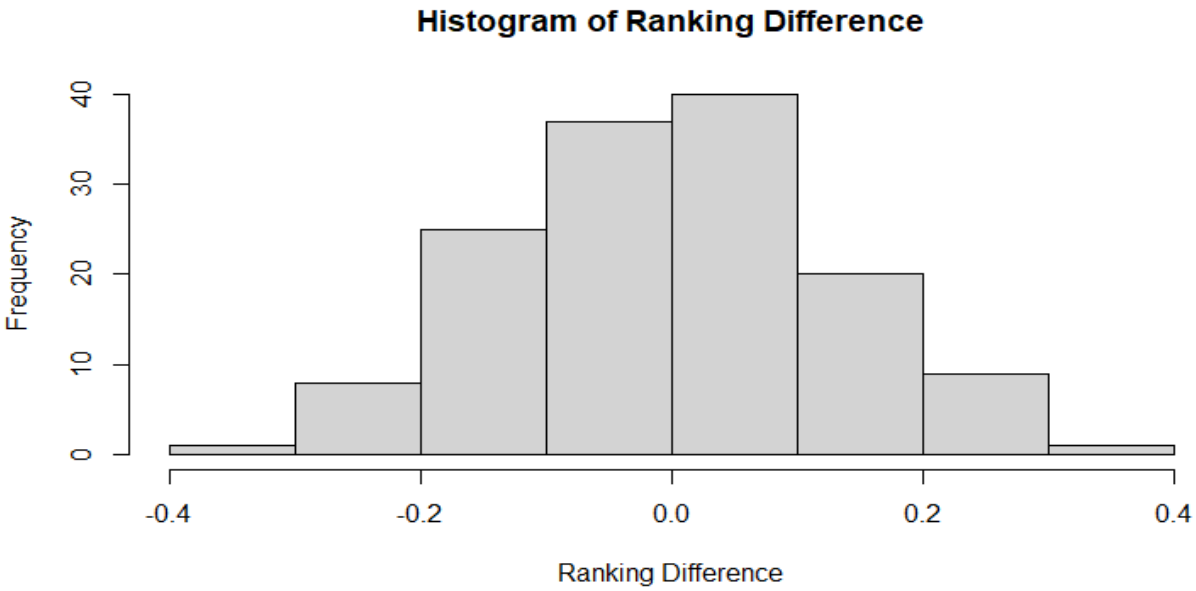

Fig. 6: Distribution of the Ranking Difference

From the Fig. 6, we see that generally it does not change a lot the ranking of the player but for some players it does. The overall sensitivity (OS) is 0.13， calculated as:

$$OS = \sqrt{mean((Ranking\ Difference)^2)}$$

Table 7 shows the players who performed better considering the model without the three omitted variables:

TABLE 7: The best 6 players of the Serie A TIM 19/20 according to the new ranking

	Player	Team	Age	Ranking
1.	R. Gosens	Atalanta	26	1.285
2.	A. Gómez	Sevilla	33	1.248
3.	E. Pulgar	Fiorentina	27	1.040
4.	D. Berardi	Sassuolo	26	0.999
5.	J. Kurtic	Parma	32	0.902
6.	Luis Alberto	Lazio	28	0.872

R. Gosens jumped in the first position according to the new ranking (Table 7), meaning that he benefited a bit from the variables' omission considering that its past ranking (Table 3) value was 1.100. We find in the fourth D. Berardi, meaning that he did very good if we do not consider the fouls and the cards.

Player	Fouls per 90	Yellow cards per 90	Red cards per 90
D. Berardi	1.89	0.2	0.03

As we can see D. Berardi has an average of two fouls per match and a yellow card every five match.

In the following sets of graphs (Fig. 7 and Fig. 8) I will show the players more affected from the fair play variables. Figure 7 represents the players who performed better in the second model.

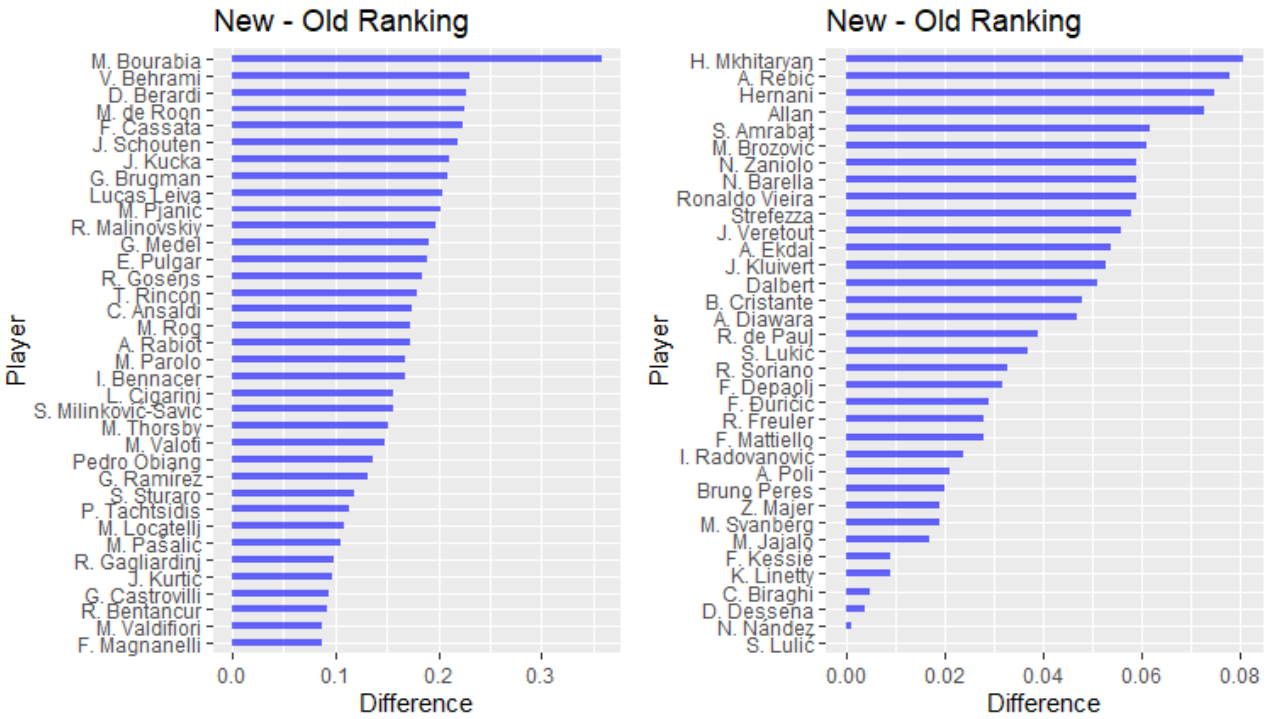

Fig. 7: Ranking Difference of each player (part 1)

The players M. Bourabia, V. Behrami, D. Berardi would have seen their ranking value increased by more than 0.2 if the fouls and the card were omitted in the initial model. For some players, the most aggressive ones, the sportsmanship is an issue. They could be top class players if only they learnt how to behave in some situations. Sometimes, they get yellow or even red cards just for protesting referee's decision. For a further analysis, it would be interesting to consider in a model the cards taken for this reason (Fig. 7).

Figure 8 represents the players who performed better in the first model (or the one who, in the second model, lost ranking value).

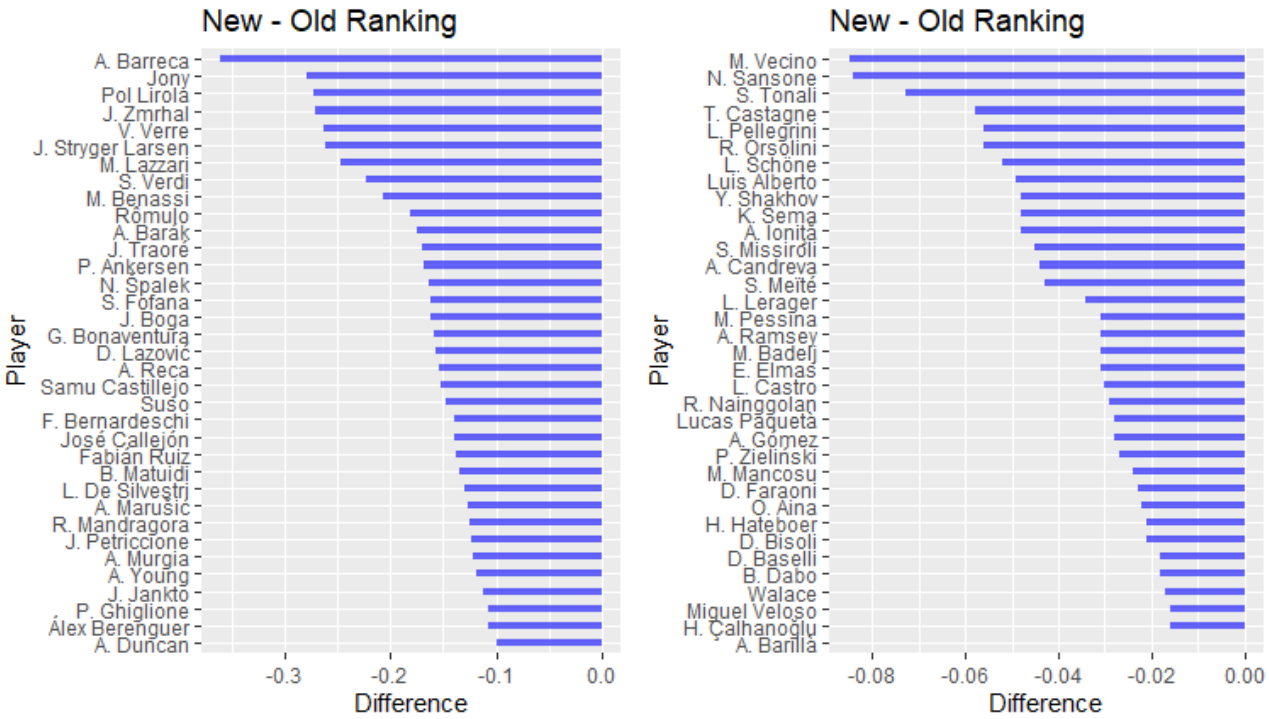

Fig. 8: Ranking Difference of each player (part 2)

Players like A. Barreca, Jony and Pol Lirola were instead influenced by the omission of the variables in a negative sense. They did not do many fouls and took few cards (Fig. 8).

5. DISCUSSION AND CONCLUSIONS

5.1. DISCUSSION

A more complex analysis would require more time. There are a lot of variables that were not considered.

- Set pieces: free kick, corner kick, penalty and throw in.
- Ball possession and pass accuracy: how many touches a player does in the game and the different type of pass accuracy. The general pass accuracy was considered, but one could have considered the long, medium, or short pass accuracy, the second assists a pass into the penalty area, etc…
- Fair Play variables: protest fouls (as mentioned before), penalty fouls, fouls suffered, simulation fouls.

- Counterattack and offensive variables: acceleration, clearances, progressive run, touch in the box.

With the inclusion of these variables, the efficiency of the model can be increased. From this analysis is possible to divide the players for positions and furthermore to consider in the model the defenders and the forwards, calculating a ranking for each position on the pitch, an example of how the players should be divided can be: central defenders, defensive midfielders, offensive midfielders, defensive wings, offensive wings, and forwards.

5.2. CONCLUSION

In this paper I analysed the midfielders' performances of the "Serie A TIM" 19/20. Firstly, I choose a dataset with the most characterising variables for a midfielder. Then I chose a method, the factor analysis, to perform a statistical analysis on my data.

I expected that some players would have been at the top of the ranking, as A. Gómez, R. Gosens and Luis Alberto, who played a fantastic season. However, it was a quite impressive discovery the 6th placement of J. Kurtic, but looking at his statistics carefully I saw that he had a very good percentage in duels won. The best younger midfielder was J. Boga, who also quadrupled his market value.

Consequently, I commented the transfers of Suso and Allan, saying that for those prices and considering only the last season, they were good market decisions by the teams who bought them.

Afterwards, I wanted to see what would have changed if the fair play factor were not included in the model, computing a new ranking for all the players. In this new ranking D. Berardi reached the 4th position, meaning that he did a lot of fouls and took some yellow and red cards. My analysis allows us to see whether market choices of a club are mainly based on statistical indicators or not. In the cases analysed previously the clubs did not look at the performances of the singles players but to other factors as market value, contract expiry or bad relationship with the coach, as in A. Gómez case, not mentioned before, but sold to Sevilla during the 20/21 season because of a row with Atalanta's coach Gian Piero Gasperini, as rumors said [10].

The market value of a player depends on many variables: performances, age, current team, contract length. However, market decisions are not made only looking at the statistics of a player. Sometimes, players can be a good deal for a team, but the same player can be a bad one for another. It might depend on the current team line-up, for instance, if a team has already two good

goalkeepers it does not make any sense to pay a lot for a third one. The deal can also depend on the economic situation of a club, on the legal restrictions of the league, on the player's salary or his will to play for a team or for another.

6. MODEL AND DATA

The R code and the dataset are available at:
https://github.com/EmanueleTartaglione/Bachelor-Thesis

7. REFERENCES

[1] https://www.wikipedia.org/

[2] https://www.goal.com/en/news/what-is-xg-football-how-statistic-calculated/h42z0iiv8mdg1ub10iisg1dju

[3] https://www.si.com/soccer/manchestercity/news/report-identified-man-city-star-as-significantly-underpaid-data-analytics-crucial-during-negotiations-over-new-deal#:~:text=Apr%2011%2C%202021-,Manchester%20City%20midfield%20star%20Kevin%20De%20Bruyne%20used%20data%20analytics,term%20future%20to%20the%20Etihad.

[4] https://wyscout.com/

[5] https://dataglossary.wyscout.com/

[6] Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141–151.

[7] "Gli MVP della stagione 2019/2020" (PDF) (in Italian). Lega Serie A. 4 August 2020. Retrieved 4 August 2020.

[8] "Gran Galà del Calcio: The top XI". Football Italia. 19 March 2021. Retrieved 19 March 2021.

[9] https://www.transfermarkt.com/

[10] https://www.theguardian.com/football/2021/jan/26/sevilla-sign-alejandro-papu-gomez-from-atalanta-in-cut-price-deal

https://rpubs.com/danmirman/plotting_factor_analysis

https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1226&context=pare

Anderson, T.W. (2003). An Introduction to Multivariate Statistical Analysis. Third Edition, Wiley.