UPPSALA
UNIVERSITET

PROJECT REPORT

Statistical and Visualization Methods on Evaluating Players in Football

Zonghao Lu, Jiayi Yang

Project in Computational Science: Report

January 2020

Institutionen för informationsteknologi

Abstract

Currently, the rating and ranking systems of football players are based on simple statistical data, such as number of goals or passes. This project in computational science offers a new solution to evaluating players' performance, by developing scientifically effective models, calculating player ratings against all the players playing the same position in different leagues, and visualizing players' performance as a readable scouting report. This project also takes into account how to improve the computing performance and evaluate the models.

Acknowledgements

The Project in Computational Science, *Statistical and Visualization Methods on Evaluating Players in Football*, is under the supervision of David Sumpter. Thanks to his ideas and stimulating support, we succeeded in developing the topic of the project, and in drawing conclusions of sufficient interest for the involved collaborators. *Hammarby IF Fotboll* provided the account to Wyscout database, the models developed by *Twelve* are the starting points of our project, experts from *B3 Consulting* showed us the future of football player scouting with advanced techniques. Here we would like to express our gratitude to them. Last, but not least, thanks to the course coordinator Maya Neytcheva, for her coordination over the course, and helpful advice and corrections on this report.

1 Background

Football (also known as soccer, particularly in North America) is widely acknowledged to be the world's most popular sport during the past decades [1]. It has a long history which can be dated to 600 AD and as a modern competitive sport, football never stops its evolution process. Recently, data analysis comes into stage and it becomes popular in many aspects. For example, data can quantify players' performance which helps the coach to make decisions. In addition, data can be an important reference when a transfer of a player takes place. Therefore, digging into the data and finding valuable information has a significant impact on modern football.

2 Introduction

In this project, we developed scientific and effective models on evaluating football players' performance based on different kinds of football data provided by Wyscout[1]. Different kinds of events have their own impacts in a football match. Further, we use a novel concept that called XGChain (expected goal chain) to describe all the events since football is consist not of single moves but of continuous passes. This concept is the main idea of our project. Having these models, player ratings against all the players playing the same position in different leagues can be calculated. Moreover, visualization is a significant part to make the computation results more readable and easy to understand. In the following parts, Section 3 is talking about how we pre-process our data and

[1] https://wyscout.com/

Section 4 focuses on the most important part modeling. Then, method and results are mentioned in Section 5 and Section 6 separately. The last two parts are conclusion and discussion.

3 Data Preparation

Opta and Wyscout are presently two of the biggest football data providers in the world. They have similar match data with only a few differences. In this project, we use data from Wyscout. The data can be divided into two parts, the player personal information and the match records. Each match is a JSON file containing all the events. Each event is described by different tags, such as team_id, time, event_type, start_position, and end_position. All these different events can be used to form the original status of the data.

In order to use the data more conveniently, a data pre-processing should be conducted before any other operations. Event chains are created by linking events by their start position and end position. If one event has the same start position as the end position of another event, it can be considered as a following event. The key is that we start a new chain if and only if one team has at least two consecutive events. After the pre-processing, every event should be included in chains. All personal data of the players is extracted in advance.

4 Modeling

Currently, the rating and ranking systems of football players are based on simple statistic, such as number of passes made, tackles or pass success rate. This kind of statistic is commonly used by both media and football clubs. It seems to be an intuitive and natural way to evaluate football players, but sometimes it is not as reasonable as expected. A player who tackles a lot might do so because he is badly positioned in the first place. Further, a player with a high pass success rate might always choose easier but ineffective pass option. Therefore, we need a more nuanced view of player performance. One solution, introduced by David Sumpter and *Twelve* [2], is to look at how the actions a player takes increase and decrease the chance of his team scoring, by using Markov Chains. In order to implement that, some statistical models are introduced.

4.1 Expected Goals (xG)

Expected Goals, also known as xG, is a predictive model used to assess every goal-scoring chance, usually expressed as a number between 0 and 1, on whether a given shot will result in a goal. Since there is a common rule on the football pitch, namely, *the more of the goal you can see when you shoot, the better your chance of scoring*, the angle can be seen as a starting point for expected goals models. Figure 1 shows shot success statistics in terms of the angle to the goal post, by using Opta's shot data over the last two seasons. The more red the colour, the higher the probability for a shot from this position during the past season. Shots from the bright red area have at least a 20% chance of going in. The circle marked by the colour change from yellow to orange is where the probability of scoring is around 5%. Further out the probability of scoring drops off dramatically.

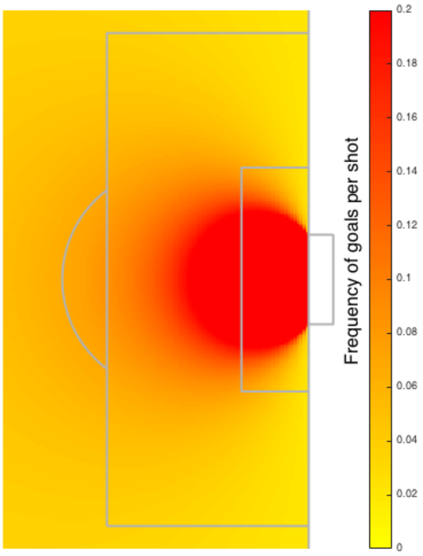


Figure 1: Frequency of goals per shot in terms of the angle

In addition to the angle, distance to goal and distance to centre of the pitch might be also included in our model, shown as Figure 2. This is a more complicated, and more accurate, picture of the shooting success. Instead of the circles we find when we look only at the angle to the posts, we get a sort of squashed out circle. In practice, this tells us that shooting from a bit wider out can still result in a goal.

From this point, a model can be fitted by using logistic regression as a traditional approach to predict probability of the goal. According to Sumpter[3], the best model for goal prediction is:

$$P\left(goal\right)= \frac{1}{1+e^{4.03-2.53\theta+0.12x+0.11x\theta-0.0069x^2}} ,$$

where θ is the post angle and x is the distance to the goal line.

The logistic regression model is clearly a proper starting point to determine the likelihood of a goal. Furthermore, based on the big data from the football pitch, much more variables could be taken into account, such as shooting part, passage of play, chance creation, or the feature of the shot.

In this project, a model developed by Twelve is evaluated, considering 30 parameters, including with the position of the shot, shooting part (head, left foot, right foot, or other body part), passage of play (regular play, fast break, set piece), how the chance has been created (Is that from a corner or a free kick? Is

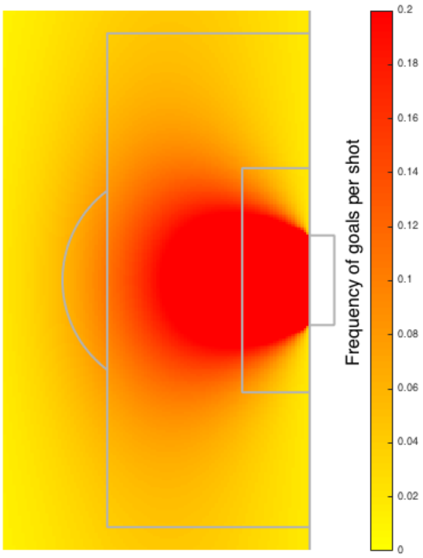


Figure 2: Frequency of goals per shot in terms of the angle and distance

a assist and a second assist involved?), the feature of the shot (strong or weak, swerve or not) and so on so forth.

Since the model from Twelve is used for analysing the data from Opta, on the other hand, for the data from Wyscout, it is needed to evaluate the model by adjusting the parameters to filter out the features that Wyscout doesn't record. As a result, shot position, shooting part, and chance creation (from corner or free kick) are considered in the fixed model. Also, direct shot from a free kick or penalty has its separate model.

4.2 Attack Impact

The attack impact model, also developed by Twelve, is a model to assess every attack event, i.e. pass, dribble, free kick, corner and throw-in, by a value between 0 and 1, to analyse and evaluate players' attack performance. It mainly base on the position of each attack action, some other variables may also be considered.

Among the different types of attack action, pass plays a much more important role, since it is the most common movement in the pitch[4] and the base for building an attack. Pass impact algorithm consists basically of two regressions that are fitted with the help of the historical data of several seasons, which includes tens of thousands of passes. All the matches used to train the model are broken down into sequences of possession, i.e., fragments of the game during which one of the teams holds the possession of the ball without losing it and without any stops in the play (due to fouls, throw-ins, offsides, etc.).

The first regression is obtained by assigning each chain a value between 0 (if the play ends without a shot) and 1 (if the sequence finishes with a goal). A value between 0 and 1 is also assigned if the play ends with a shot but without scoring, which is xG of the shot. Thus, a first logistic regression is used to obtain the probability of a certain pass (defined by its starting and ending position on the pitch) leading to a shot and a second regression to compute the probability of a shot leading to a goal, having a final probability of:

$$P_{pass}(goal) = P(goal \mid shot) \cdot P(shot) \; .$$

In this way we have a model that, given the starting and ending coordinates of a pass, gives us the probability of the team scoring a goal before the play which that pass belongs to ends. This probability can be used to assess how valuable the pass is. Figure 3 shows two examples of the pass impact algorithm applied to two different starting coordinates of the pass.[5]

Other models on dribble, free kick, corner and throw-in use similar algorithm to generate a value as an assessment on the importance of each movement in attack, by the probability of the team scoring a goal with that movement involved.

4.3 xGChain

With the xG model, each shots in a match could be valued accurate, and with attack impact model, we can get the expected contribution of each attack action involved in a goal. However, we aim at generating the actual contribution of the attack actions to the shots and goals, as another assessment of players' attack performance. Therefore, we introduce a new model, xGChain model.

4.3.1 The Possession Chain

To introduce xGChain model, we have to define the possession chain at first. As mentioned above, a possession chain is a sequence of consecutive on-the-ball events when the ball is under the effective control of a single team. A football game can then be seen as an (ordered) collection of sequences[6]. In our project, we define that a new possession chain starts by two consecutive touches of the same team, which means that just one touch does not break opponent's possession chain. Some samples of possession chains[7] are shown in Figure 4.

4.3.2 Contribution to xG

Now, if we found a possession chain ending with a shot, we can compute the xG of that shot, and calculate the attack impact (aI) of all the attack actions in that chain. Then, we calculate the weighted mean of the impact of each attack, as xGChain of that attack action, i.e., the contribution of that attack event to the xG of the shot,which is the end of the possession chain:

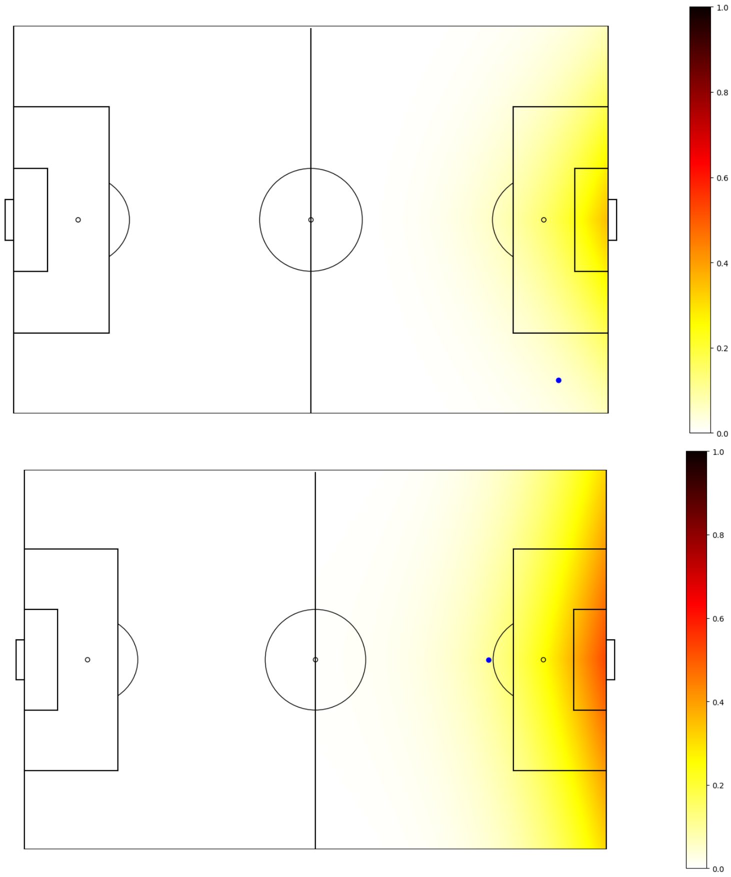


Figure 3: Pass impact for two different starting coordinates of a pass (blue dot). The figure is generated by applying the expected goals algorithm for the starting coordinates and a grid of points on the pitch as the ending ones to generate a heatmap.

$$xGC = \frac{aI}{\sum aI} \cdot P,$$
$$where \quad P = \begin{cases} 0 \quad if\ not\ shot \\ xG \quad if\ shot\ not\ scoring \\ 1 \quad if\ goal \end{cases} .$$

The xGChain model could be used as an additional assessment for evaluating players' attack performance, and give higher scores to the players who made more key passes.

4.4 Defensive Impact

In addition to attack movements, we also assess defensive actions. However, using the same idea from Markov Chains, to assign a value to each position on the pitch, corresponding to the probability that a team with possession of the ball at that point scores a goal, we can evaluate players' defensive actions based on how much they decrease the opponent's chance of scoring.

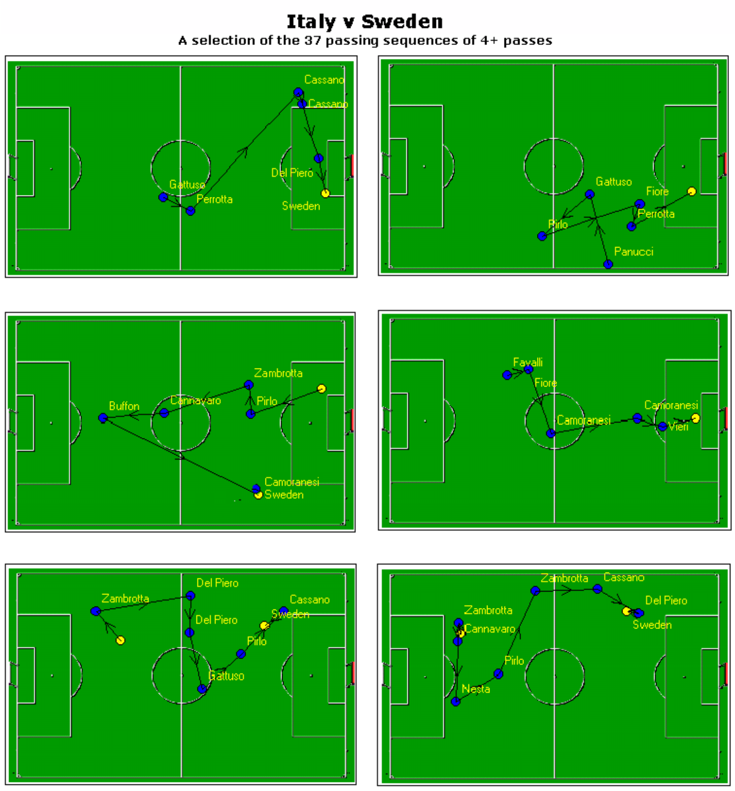


Figure 4: Samples of possession chains

In the defensive impact model developed by *Twelve*, a field impact has been calculated at first, which generates a value for each positions on the pitch. Then, for different types of defensive actions, e.g. tackle, clearance, interception and block, different weights are involve, by multiplied with the field impact of the position that the defensive action occurs. As a result, a value between 0 and 1 would be given to each defensive action as an assessment[2].

5 Method

To implement our computation models, Python is used for coding because it provides several useful and powerful visualization tools (libraries). Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hard-copy formats and interactive environments across platforms. A lot of different types of charts can be plotted by using this library with only a few lines of code which offers great convenience to us. Initially, event chains obtained from the pro-processing stage are read. Afterwards, all statistical computations are conducted on these chains in two for loops (one for all the matches and the other one for all the event chains) and the result is written to a csv file stored in the disk.

With the csv file recording all the information extracted from the event chains, many types of chart are tried, such as bar chart, pie chart, line chart and radar chart. Each type of them has different dimension, thus their expressive ability are also different. When making these figures, every attribute useful to help make a scouting decision is taken into consideration

Last but not least, we implement a parallel solution to accelerate the computation above. Because each json file is an independent match and has no relation with others which provides a perfect property to be processed in parallel. Python built-in library *Pool* is a user-friendly tool to help develop a parallel environment. Figure 5 illustrates the strong scalability and weak scalability of the parallel computation with Pool.

As we can observe from these two figures, neither of them performs well. It can be assumed that when reading data from the disk, frequent IO operations have a high overhead which limits the speedup performs well.

6 Results

The scouting system we developed is expected to be used by the football industry practitioners, so that, the spreadsheet (the csv file) with values of the performance of the players per match needs further analysing, and to be visualized in a more intuitive and readable way.

Therefore, we tried different eays to visualize the results, stored in a spreadsheet. The attributes included in the spreadsheet could be divided into two groups: information and impact. For information, it contains match information (Team, Opponent, Match Time, Home/Away) and player information (Name, Age, Minutes Played, Position), by which the players in same age or same position could be compared easily. For impact, there are attack impact (xGChain, Assists, Total Pass, Defensive/Middle/Final Third Pass, Throw-in, Corner, Free Kick, Dribble Won, Dribble Lost), shot impact (xG, Goals), and defensive impact (Clearance, Block, Interception, Tackle Won, Tackle Lost, Total Heading, Defensive/Middle/Final Third Heading, Yellow Cards, Fouls, Goalkeeper). All the attributes, mentioned above, can be visualized in the same way as the following samples.

In Figure 6 we compare Vladimir Rodic, Muamer Tankovic and Imad Khalili among all the forwards in Allsvenskan. The figure shows the cumulative ranking points of shot perfermance. The y-axis refers to cumulative percentile ranking, in which a player on 100 is the best, while 0 is the worst. The x-axis shows the amount of minutes played. Overall, such a figure shows the level of a player in the league, and compares it within a group of players that readers may be interested.

Figure 7 shows players' pass contribution in different areas on the pitch, by which the readers can learn players' pass performance in high pertinence. For example, although J. Andersen has more total pass impact, M. Tankovic has much more pass contribution in the final third, which is more important for a

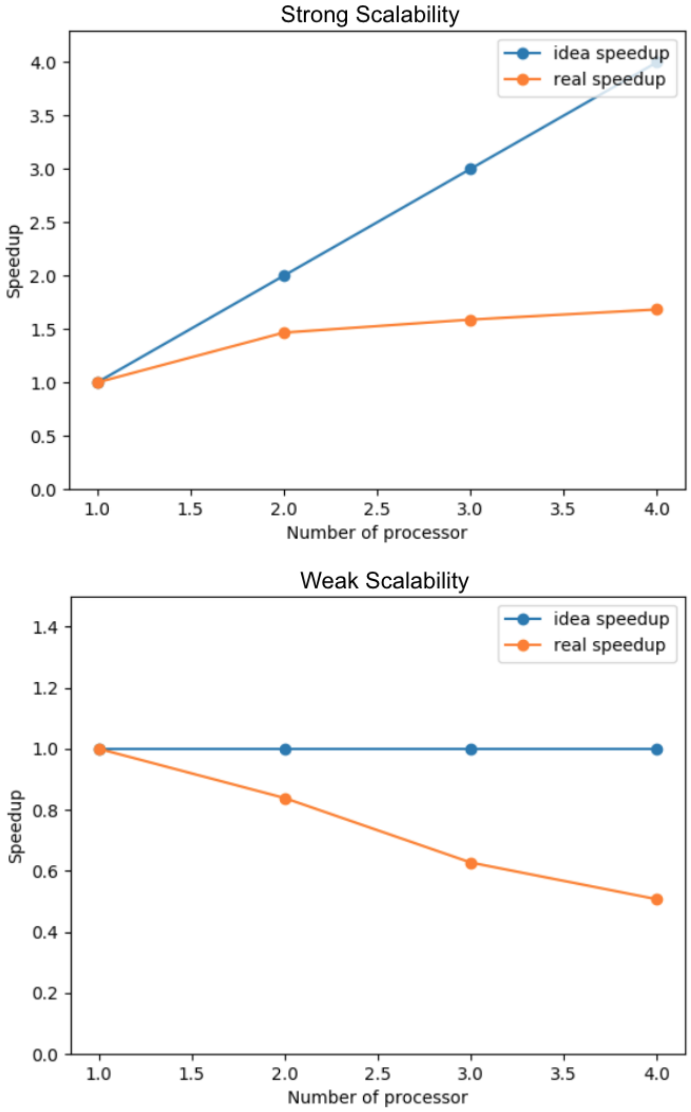


Figure 5: Main name

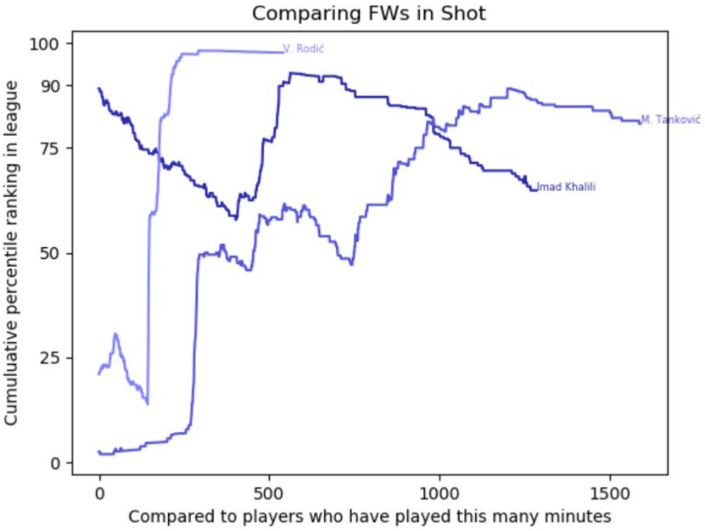


Figure 6: Sample of the line diagram

forward.

The radar chart (Figure 8) is a kind of all-in-one solution to evaluate players. It clearly shows all the attributes that the readers concern in one or two figures. In this example, the performance of Muamer Tankovic is shown in two radar charts, one in attack, another in defensive. The values in radar charts are represented in cumulative percentile ranking, 1 means the best and 0 the worst. Because of their intuitive and information-rich contents, radar charts are well received by scouts and sports directors. Some radar charts plotted from our implementation have been used in real-world player transfer process.

7 Discussion

7.1 Limitations

The key of the scouting system designed in this project, are the models. Although it gives insightful results, it still has potential to be improved. For example, the defensive impact model evaluates players' defensive performance assessing the defensive actions. However, for defenders, especially central backs, the positions they occupy when the opponent possesses the ball are important, which can efficiently decrease opponent's chance of scoring, and certainly should be assessed as the players' defensive contribution. But, the information of the players without ball are not included in either Opta or Wyscout data. Therefore, players' defensive performance analysed in this project is only one aspect of players' defensive ability.

Another limitation, or challenge, in such scouting assessments is how to evaluate the models themselves. Evaluating players is a subjective work, and we are

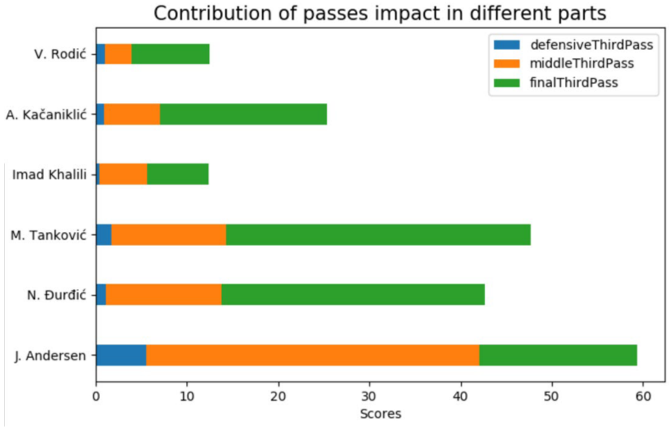


Figure 7: Sample of the bar chart

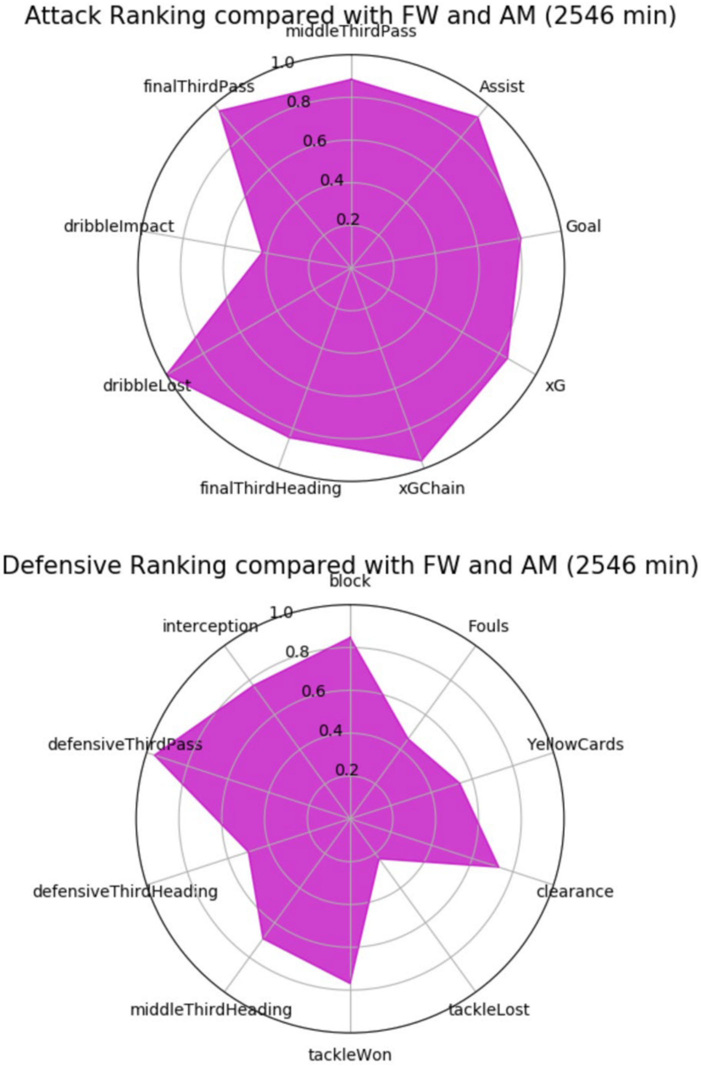
trying to simulate it in a scientific way, but obviously, there is not a standard answer to the result. During our project, we referred to experts and some player scoring and ranking issued by media, to verify the efficiency of our system, but this will still be a problem in future work.

7.2 Future Work

Recently, position information of all the player on the pitch during the whole match recently start to be collected by new technology. This kind of tracking data, containing opponents' behaviour, could be used in modelling, not only for the defensive impact model as mentioned above, but also for the xG model, to assess the likelihood of goals more accurate. The improvement of modelling may lead scientific player scouting to a giant leap.

8 Conclusion

During this project, several scientific and effective models on evaluating football players' performance have been developed. These models are used to calculate player ratings against all the players playing the same position in different leagues. However, it is not successful to conduct a parallel implementation in this situation. Then the results of players' performance could be visualized in the form of readable figures, to be included in scouting reports.



Attack Ranking compared with FW and AM (2546 min)

Defensive Ranking compared with FW and AM (2546 min)

Figure 8: Sample of the radar chart

References

[1] R. Giulianotti, “Football,” *The Wiley-Blackwell Encyclopedia of Globalization*, 2012.

[2] D. Sumpter, “Using markov chains to evaluate football player’s contributions,” https://medium.com/@Soccermatics/using-markov-chains-to-evaluate-football-players-contributions-57a107cc09e6, accessed January, 2020.

[3] ——, “The geometry of shooting,” https://medium.com/@Soccermatics/the-geometry-of-shooting-ae7a67fdf760, accessed January, 2020.

[4] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, “Actions speak louder than goals: Valuing player actions in soccer,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, ser. KDD â19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1851–1861. [Online]. Available: https://doi.org/10.1145/3292500.3330758

[5] F. J. Peralta Alguacil, “Modelling the collective movement of football players,” p. 57, 2019.

[6] M. Kwiatkowski, “Towards a new kind of analytics,” https://statsbomb.com/2016/08/towards-a-new-kind-of-analytics/, accessed January, 2020.

[7] “Possession chains and passing sequences,” https://soccerlogic.com/2016/08/09/possession-chains-and-passing-sequences/, accessed January, 2020.