Similarity of football players using passing sequences

Alberto Barbosa[1], Pedro Ribeiro[1], and Inês Dutra[2]

[1] INESC-TEC & DCC/FCUP, University of Porto, Portugal
[2] CINTESIS & DCC/FCUP, University of Porto, Portugal

Abstract. Association football has been the subject of many research studies. In this work we present a study on player similarity using passing sequences extracted from games from the top-5 European football leagues during the 2017/2018 season. We present two different approaches: first, we only count the motifs a player is involved in; then we also take into consideration the specific position a player occupies in each motif. We also present a new way to objectively judge the quality of the generated models in football analytics. Our results show that the study of passing sequences can be used to study player similarity with relative success.

1 Introduction

Association football is one of the most popular team sports in the world. Traditionally, studying many aspects of the game has been relying on the empiric experience of coaches and scouts. However, some aspects of the game have been the subject of research in many different fields of science, given the growing availability of data related to football matches. Player injury forecasting [14], team behaviour visualisation [9,18], talent discovery [5,13,15] and transfer market analysis [3,7,8] are some of the features of the football industry that have been continuously receiving attention from different fields of computer science.

This work has two major contributions: the study of passing networks using network motifs and the study of player similarity, using different ways to measure the quality of the results and models obtained.

2 Related Work

Research on association football has been growing in popularity and many aspects of the game have been receiving recent attention, like visualising and analysing team formations and their dynamics [17] or predicting match results [1]. Given the subject this paper, we will mainly focus on research that delved into studying motif based patterns in passing networks.

A passing network can be seen as a graph where the nodes represent the players and there is an edge going from player A to player B if player A successfully passed the ball to player B. Milo et al. defined network motifs as "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomised networks" [10]. Later, Gyarmati et al [4] defined flow motifs. Considering a passing sequence, a flow motif is a subsequence of the

passes where labels represent distinct players without identity. In the context of this paper, all motifs are flow motifs. We will next make a short description previous research on this topic.

Bekkers and Dabadghao [2] applied network motif methodology to football passing networks in data comprising 4 seasons of 6 major football leagues. They were able to identify unique play styles for both teams and players.

Håland et al. [6] studied the Norwegian elite league of football, concluding that passing can be modelled using networks and sequences of passes can be mapped into flow motifs. Their most relevant finding was that although more compact motif types (with fewer different players involved) had a lower likelihood of leading to shots, no connection between the ranking of a team and their distribution of flow motifs was clear.

Peña et al. [12] also applied flow motifs to football passing networks and clustered players according to their participation in different flow motifs, trying to identify unique play styles. The outlier in their analysis was the former FC Barcelona midfielder Xavi Hernández, which formed a singleton cluster due to his unique passing style and its influence in Barcelona's play style.

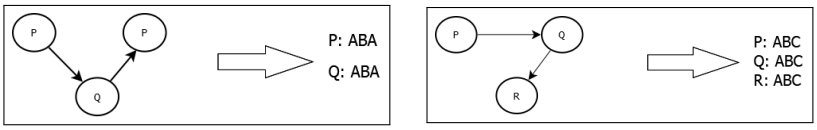
Gyarmati et al. [4] propose a quantitative method to evaluate the styles of football teams through their passing structures, through the study of network motifs, concluding that FC Barcelona's tiki-taka does not consist of uncountable random passes but has a finely constructed structure instead.

Wiig et al. [16] identify some players as key passers and/or passing recipients using network analysis on passing networks of teams in Norway. Some interesting conclusions of their work consist on showing that offensive players tend to have a high closeness centrality measurements and high PageRank values for pass recipients, whereas defenders tend to have high PageRank values for passers.

3 Data Description

The data used in this work was retrieved from a public data set containing spatio-temporal events in association football [11]. The data set contains events from the 2017/2018 season of the top tier leagues in Spain, England, Italy, Germany and France. In addition to that, data of the World Cup 2018 and of the European Cup 2016 is also provided. However, our analysis embraces the club competitions only, due to higher number of matches and events they provide, having the potential to yield more trustful and data reliable conclusions than small competitions with less data to extract knowledge from.

We pre-processed the raw data events to transform them into sequences of passes between the players involved and to cut the players that did not participate in, at least, 80% of the games that season. The later decision was made because, since we only have one season of match and event information, we preferred to work with players that had more consistent data regarding their passing and play styles and also it was important to our evaluation model to work with players that had played throughout the season consistently.



(a) ABA motif (b) ABC motif

Fig. 1: Processing passing sequences into motifs. In (a), we have a sequence of passes from player P to player Q and then to P again. This is processed as both players participating in an ABA motif. In (b) we have a different sequence of passes involving three different players: P passes to Q which then passes to R, all of them participating in an ABC motif. When we have bigger sequences of passes, these are divided into motifs of size k and processed in a similar way.

4 Methodology and Results

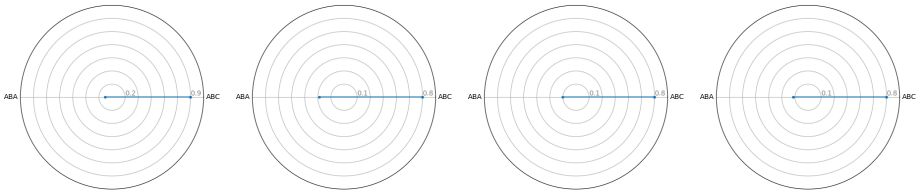
Our methodology and results will be described along each other as we discuss each of the three different approaches we followed in solving the problem of assessing player similarity in association football.

Our initial approach was to count the frequency of different size 3 and size 4 network motifs for each of the players involved. Such motifs were extracted from the passing sequences as in Figure 1. These motif sizes offer a good initial compromise between having enough data to analyse (larger passing sequences are less frequent) and having a relatively small number of different motifs that will constitute the features that we incorporate in our proposed distance metric, depicted in Equation 1. As said, we compute the distance between players A and B by considering M, the set of all motifs of size 3 and 4. P_m represents the normalised number (between 0 and 1) of motifs that player P was involved in. When normalising P_m we divide the number of times a player was involved in a motif of type m by the total number of times a player has been involved in all motifs in M.

$$D(A,B) = \sqrt{\sum_{m \in M} (A_m - B_m)^2} \tag{1}$$

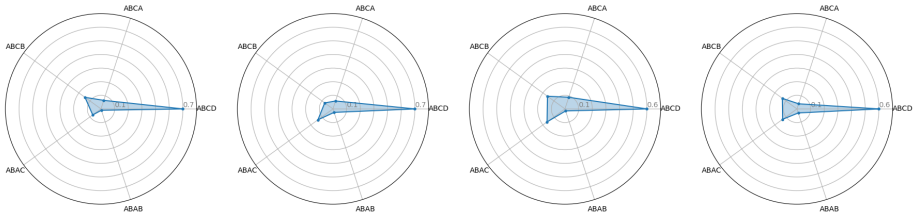
After counting the number of times a player was involved in a specific passing pattern, we computed the distance from him to every other player, as in Equation 1, to every other in order to see which players were the most similar to each other.

We evaluate our results in two different ways: visually and objectively. For a visual evaluation, we draw radar plots that mirror the participation of the player in all different motifs considered and visually compare their similarity. We also check the most similar players and use our domain knowledge to interpret the results. Since we analysed over 560 players, it would be impossible to present the radar plots of all the players and we choose a sample of 4 players to showcase the empirical analysis we performed, namely the goalkeeper Ederson (Manch-



(a) Ederson (b) Cédric Soares (c) Jorginho (d) Messi

Fig. 2: Radar plots of the four example players considering only size 3 motifs.



(a) Ederson (b) Cédric Soares (c) Jorginho (d) Messi

Fig. 3: Radar plots of the four example players considering only size 4 motifs

ester City), the defender Cédric Soares (Southampton), the midfielder Jorginho (Napoli) and the forward Lionel Messi (Barcelona).

We felt that a visual comparison of radar plots was not enough since we wanted a more objective way to score and compare different approaches. Our evaluation model was achieved by dividing our data set in two halves (first and second half of the season). With those two new data sets, we performed the same task of counting the motifs each player was involved in and we calculated the distance between all players. The evaluation score is given by the number of players that have themselves in the top-10 of similarity in different halves. This model is based on the intuition that the same player in the same season should have a similar behaviour when it comes to passing and play styles.

For each experimental setting, we performed two different tests: first we tested our approach with data from each league individually and then with data from the top-5 leagues in Europe simultaneously. This was done to test the ability of our model to deal with more data and noise, since the task of identifying similar players gets harder as we increase the amount of players to compare.

In Figures 2 and 3 a visual analysis of the radar plots of some players showed us that it was very hard to clearly obtain a fingerprint of the play style of the player merely through it. With size 3 motifs, we can barely notice any difference between different players, and the main observation that we can extract is that the fraction of ABC patterns seem to be higher in every player, which seems reasonable (it is much more likely that 2 consecutive passes go through 3 different players that two players passing between each other). Also, a difference in the fraction of ABA and ABC passing patterns can be noticed in some players (for

	1st	2nd	3rd	4th	5th
Ederson	Douglas Costa	J. Alonso	K. Koulibaly	Campaña	I. Diop
Cédric Soares	Y. Sabaly	D. Suárez	K. Naughton	M. Olsson	Cédric Soares
Jorginho	E. Zukanovic	E. Pulgar	S. Missiroli	I. Bebou	Jorginho
Messi	M. Kruse	D. Yedlin	F. Guilbert	P. Faragò	C. Traoré

Table 1: Top-5 similar players to our 4 example players according to their participation in size 3 passing sequences. We compared the passing patterns of each player in the 1st half of the season to the patterns of the entire set of players (of all leagues) in the 2nd half of the season.

	1st	2nd	3rd	4th	5th
Ederson	Allan	Hradecky	B. Dibassy	C.Schindler	Maxi Gómez
Cédric Soares	I. Gueye	K. Naughton	G. Bonaventura	S. Ascacíbar	Á. Correa
Jorginho	Zielinski	J. Lascelles	M. Veljkovic	M. Politano	K. de Bruyne
Messi	E. Pulgar	L. Suárez	S. Papastathopoulos	B. Oczipka	Malcom

Table 2: Top-5 similar players to our 4 example players according to their participation in size 4 passing sequences. We compared the passing patterns of each player in the 1st half of the season to the patterns of the entire set of players (of all leagues) in the 2nd half of the season.

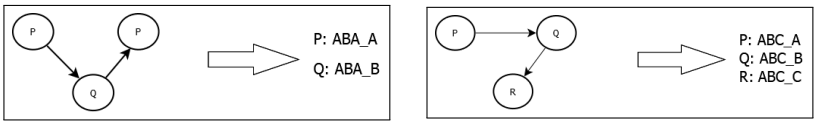
	England	France	Germany	Italy	Spain	All
Size 3	14,61%	17,20%	10.60%	13.20%	21,02%	2.86%
Size 4	36.59%	28.74%	26.38	28.20%	28.50%	8.57%

Table 3: Accuracy values obtained when analysing the flow motifs of sizes 3 and 4 without considering the specific position of each player in each motif.

example between the ABA frequency in Ederson radar plot against the ABA frequency in Cédric Soares).

The size 4 motif plots give us more visual information regarding the play style of each player, with richer and more varied topological patterns. For instance, we can notice that defensive players like Ederson and Cedric Soares tend to have a higher percentage of participation in ABCD plays than more offensive players like Jorginho and Messi. This can be due to fact that more dynamic and offensive players tend to participate in more types of plays and tend to appear in more than one passing role in small sequences of passes. Even so, we can hardly notice really significant differences between different players that play very different roles in a game of football.

To complement the visual analysis of the results, we studied a sample of the 5 most similar players reported by the algorithm according to their participation in size 3 and size 4 motifs, as presented in Table 1 and Table 2, respectively. The majority of players that are reported as being the most similar to the players in our study sample have really different play styles (for example, in Table 1 Douglas Costa (a winger) is reported to be the most similar player to Ederson



(a) ABA motifs (b) ABC motifs

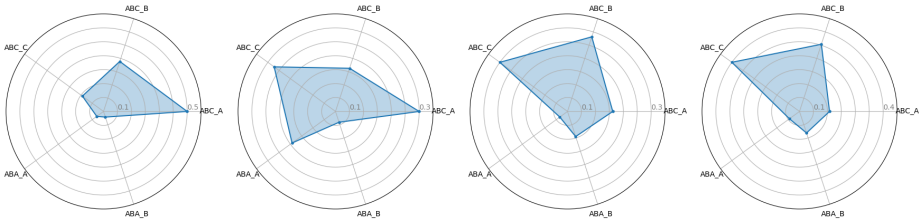
Fig. 4: Processing passing sequences into motifs considering the specific position each player represents in the motif. In (a), we have a sequence of passes from player P to player Q and then to player P again. This is processed as player P participating in an ABA_A motif (position A in ABA motif) and as player Q participating in an ABA_B motif (position A in ABA motif). In (b) we have a different sequence of passes between three players and the corresponding position motifs. As before, bigger passing sequences are divided into motifs of size k.

(a goalkeeper)). Even though, there seems to exist an improvement when we take into consideration size 4 motifs instead of only size 3 motifs, as in Table 2. The fact that Hradecky, a goalkeeper, is reported as one of the most similar players to Ederson, also a goalkeeper, and the fact that Messi and L. Suárez, both forwards in FC Barcelona, are also being reported as similar seems to indicate some sort of improvement when using size 4 motifs instead of size 3 motifs, as the radar plots showed.

Our objective evaluation model also mirrors the intuition that the empirical analysis based on the radar plotting of the results. The average accuracy of this initial approach when running only with a single league was roughly $15,32\%$ for size 3 motifs and of $28,68\%$ for size 4 motifs and the accuracy obtained when considering all top-5 European Leagues simultaneously was of 2.86% for size 3 motifs and $8,57\%$ for size 4 motifs, as shown in detail in Table 3. These results show that our initial approach, as the visual analysis seemed to show, had serious problems in capturing a player's unique play style.

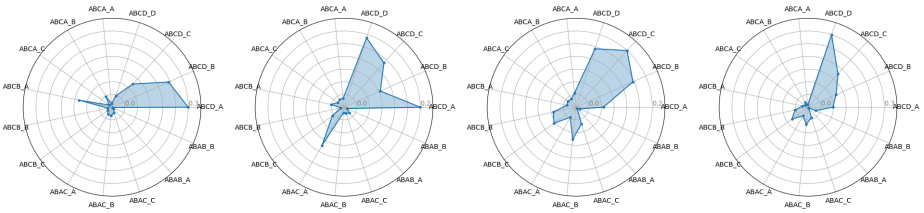
Given the poor practical results in our initial approach, we decided to enhance it by looking not only to the motifs each player is involved in, but also which specific position in the motif they occupy, or the orbit of each player, as shown in Figure 4. This would provide more variability and more features to help separate players from each other but would also help in identifying, for example, if a player is more of a play starter or a play finisher or if a player tends to be involved in more than one role in the same passing sequence. A new visual analysis of the results of applying this new approach seems to show more significant fingerprints of each player, as seen in Figures 5 and 6.

A much higher quantity of information can be extracted from these radar plots. When considering size 3 motifs, Ederson, a goalkeeper, tends to be in the starting part of the plays he participates in: he has a much higher participation in ABC_A (ABC_A means that the player occupies position A in the ABC motif) and ABC_B than in other motifs. This is confirmed again when considering size 4 motifs, with the high values of $ABCD\_A$ and $ABCD\_B$. On the other hand, Messi, a forward, seems to exhibit a very different behaviour on the pitch. He participates much more in the ending side of the plays, due to his play style and



(a) Ederson (b) Cédric Soares (c) Jorginho (d) Messi

Fig. 5: Radar plots of the four example players considering size 3 motifs and specific player roles (or orbit) in each motif



(a) Ederson (b) Cédric Soares (c) Jorginho (d) Messi

Fig. 6: Radar plots of the four example players considering size 4 motifs and specific player roles (or orbit) in each motif

position. This is confirmed by his higher percentage of *ABC_C* and *ABC_B* in size 3 motifs and *ABCD_C* and *ABCD_D* in size 4 motifs.

Other aspects of each of these players can be deduced from the radar plots characterising them. For instance, the fact that Cédric Soares, a full back, seems to appear more often either at the start of the plays or at the end of the plays, but less frequently in the middle of those plays. This can be justified with the fact that a full back in modern day football such as Cédric Soares has to be capable of being at the start of the plays due to the defensive nature of the role he has on the pitch, but also has to be capable of appearing in more offensive areas, usually to cross the ball to the teammates in the opponent's box.

This empirical analysis of the radar plots of each player seems to point that looking not only at the motifs each player participates in, but also looking at which position they occupy in each motif they participate in, causes the passing and playing style of each player to emerge in the visualisation of the results.

The most similar players, according to this approach, to each player in our study group are reported in Table 4 and Table 5. Empirically, the players reported as similar to the ones in our study group really seem to make more sense than the ones obtained with the previous approach. We can see, for example, that when considering either size 3 or size 4 motifs, Ederson is only similar to goalkeepers (even to himself). Alisson is the partner of Ederson in defending the goal for the Brazilian national team and it is widely known that they have similar play styles. This is clearly mirrored in the results, given the fact that

	1st	2nd	3rd	4th	5th
Ederson	Alisson	M. ter Stegen	S. Ulreich	S. Ruffier	H. Lloris
Cédric Soares	L. Venuti	C. Traoré	R. Bertrand	Jordi Alba	A. Masina
Jorginho	D. Demme	R. van La Parra	D. Liénard	Koke	Fabián
Messi	G. Bonaventura	S. Missiroli	Bernardo Silva	A. Mooy	T. Hazard

Table 4: Top-5 similar players to our 4 example players according to their participation and specific position in size 3 passing sequences. We compared the passing patterns of each player in the 1st half of the season to the patterns of the entire set of players (of all leagues) in the 2nd half of the season.

	1st	2nd	3rd	4th	5th
Ederson	Pepe Reina	D. de Gea	S. Ulreich	Ederson	Alisson
Cédric Soares	D. Suárez	M. Olsson	Y. Sabaly	L. Dubois	J. Korb
Jorginho	D. Liénard	N. Hoefler	X. Shaqiri	M. Antenucci	A. Knockaert
Messi	A. Sánchez	Koke	Kalou	E. Hazard	B. Cristante

Table 5: Top-5 similar players to our 4 example players according to their participation and specific position in size 4 passing sequences. We compared the passing patterns of each player in the 1st half of the season to the patterns of the entire set of players (of all leagues) in the 2nd half of the season.

	England	France	Germany	Italy	Spain	All
Size 3	38,62%	55,75%	42,59%	40,98%	49,07%	13,6%
Size 4	39,84%	52,30%	38,89%	44,36%	49,53%	17,59%

Table 6: Accuracy values obtained when analysing the flow motifs of sizes 3 and 4 considering the specific position of each player in each motif.

Alisson is reported to be similar to Ederson when considering both size 3 or size 4 motifs. Moreover, some other player similarities seem to be really accurate from an empirical point of view: Jorginho and Demme, Messi and Bernardo Silva or Hazard, Cédric and L. Dubois or L. Venuti, among others.

We wanted to see if the accuracy of the evaluation model supported our empirical analysis of the results. As seen in detail in Table 6, the accuracy when considering each league individually was $45,44\%$ for size 3 motifs and $44,98\%$ for size 4 motifs. The accuracy considering the top-5 Leagues was of $13,60\%$ for size 3 motifs and of $17,59\%$ for size 4 motifs.

These results show a clear improvement in the accuracy of the model when we take into consideration not only the motifs each players participates in, but also the specific position each player occupies in that passing sequence.

Nonetheless, there seem to be limitations in this approach. When considering all five leagues simultaneously, we can see that the current method still struggles with properly identifying the same player in different parts of the season.

Among other possible explanations to this decrease in performance, we believe that since we only worked with data from a single season, meaning that when dividing the data in half in order to build our evaluation model we only had data from 16 to 18 games at most to build a fingerprint of a player in two distinct halves of the season, the model is really sensitive to small fluctuations that may occur in the performance of the player in those games.

5 Conclusions

In this work we presented a study on the similarity of association football players according to their passing behaviours during the course of one season in 5 different European football leagues.

Studying similarity measures for football players can be very useful for football teams, since they can help the scout department discover new players with potential to join the team they work form.

When taking into consideration only the frequency of participation of the player in different passing motifs, the results showed that little information was actually being extracted from that data, especially when only considering size 3 motifs.

With the addition of the specific position a player occupies in each motif he participates in, we could see a great improvement in player characterisation and similarity both empirically and objectively.

However, even though the results show a great improvement and seem to yield good practical results, some limitations are evident. The fact that the addition of more players to the data set really decreases performance seems to indicate that the algorithm has some difficulties in identifying the uniqueness in player's play style when the amount of data is increased. This could be due to the fact that the amount of data available is not enough to the algorithm to behave better with noisier data set or it could be the fact that more aspects of the game have to be taken into account in order to uniquely identify the play style of a player.

6 Future Work

The results achieved in this work show real promise and, as such, some future work can be done to complement and enhance this approach.

The introduction of the spatio-temporal dimension of the passing data into the model can help in calculating player similarity.

The availability of tracking data of the players can help understand player movement when he does not have the ball. During most of the game, a player does not have the ball in his control, so a huge part of the role of a player in football match is being discarded.

Study team behaviour and similarity using a similar methodology may be possible and useful to improve the knowledge a team has of their opponent, which can help in designing a unique strategy to beat an opponent.

Acknowledgements. This research was funded by FCT and INESC-TEC under the grant SFRH/BD/136525/2018, Ref CRM:0067161.

References

1. Baboota, R., Kaur, H.: Predictive analysis and modelling football results using machine learning approach for english premier league. International Journal of Forecasting **35**, 741–755 (2019)
2. Bekkers, J., Dabadghao, S.: Flow motifs in soccer: What can passing behavior tell us? Journal of Sports Analytics **5**(4), 299–311 (2019)
3. Fűrész, D.I., Rappai, G.: Information leakage in the football transfer market. European Sport Management Quarterly pp. 1–21 (2020)
4. Gyarmati, L., Kwak, H., Rodriguez, P.: Searching for a unique style in soccer. arXiv preprint arXiv:1409.0308 (2014)
5. Haave, H.S., Høiland, H.: Evaluating association football player performances using markov models (2017)
6. Håland, E.M., Wiig, A.S., Hvattum, L.M., Stålhane, M.: Evaluating the effectiveness of different network flow motifs in association football. Journal of Quantitative Analysis in Sports **16**, 311 – 323 (2020)
7. Kroken, C., Hashi, G.: Market efficiency in the european football transfer market (2017)
8. Matesanz, D., Holzmayer, F., Torgler, B., Schmidt, S.L., Ortega, G.J.: Transfer market activities and sportive performance in european first football leagues: A dynamic network approach. PLoS ONE **13** (2018)
9. McLean, S., Salmon, P., Gorman, A.D., Wickham, J., Berber, E., Solomon, C.: The effect of playing formation on the passing network characteristics of a professional football team. Human Movement **2018**, 14–22 (2018)
10. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science **298**(5594), 824–827 (2002)
11. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., Giannotti, F.: A public data set of spatio-temporal match events in soccer competitions. Scientific data **6**(1), 1–15 (2019)
12. Peña, J.L., Navarro, R.S.: Who can replace xavi? a passing motif analysis of football players. arXiv preprint arXiv:1506.07768 (2015)
13. Reinders, H.: Talent identification in girls soccer: A process-oriented approach using small-sided games (2018)
14. Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernández, J., Medina, D.: Effective injury forecasting in soccer with gps training data and machine learning. PLoS ONE **13** (2018)
15. Tovar, J., Clavijo, A., Cardenas, J.: A strategy to predict association football players' passing skills. Universidad de los Andes Department of Economics Research Paper Series (2017)
16. Wiig, A.S., Håland, E.M., Stålhane, M., Hvattum, L.M.: Analyzing passing networks in association football based on the difficulty, risk, and potential of passes. International Journal of Computer Science in Sport **18**, 44 – 68 (2019)
17. Wu, Y., Xie, X., Wang, J., Deng, D., Liang, H., Zhang, H., Cheng, S., Chen, W.: Forvizor: Visualizing spatio-temporal team formations in soccer. IEEE Transactions on Visualization and Computer Graphics **25**, 65–75 (2019)

18. Yu, Q., Gai, Y., Gong, B., Gómez, M.Á., Cui, Y.: Using passing network measures to determine the performance difference between foreign and domestic outfielder players in chinese football super league. International Journal of Sports Science & Coaching **15**, 398 – 404 (2020)