Research article

Else Marie Håland, Astrid Salte Wiig, Lars Magnus Hvattum* and Magnus Stålhane

Evaluating the effectiveness of different network flow motifs in association football

https://doi.org/10.1515/jqas-2019-0097
Received September 13, 2019; accepted July 1, 2020; published online November 3, 2020

Abstract: In association football, a network flow motif describes how distinct players from a team are involved in a passing sequence. The flow motif encodes whether the same players appear several times in a passing sequence, and in which order the players make passes. This information has previously been used to classify the passing style of different teams. In this work, flow motifs are analyzed in terms of their effectiveness in terms of generating shots. Data from four seasons of the Norwegian top division are analyzed, using flow motifs representing subsequences of three passes. The analysis is performed with a generalized additive model (GAM), with a range of explanatory variables included. Findings include that motifs with fewer distinct players are less effective, and that motifs are more likely to lead to shots if the passes in the motif utilize a bigger area of the pitch.

Keywords: generalized additive model; passing; regression; soccer.

1 Introduction

Decisions in sport have traditionally been made qualitatively by humans, based on gut feelings or adherence to team culture and tradition. Sport analytics offers new ways of assessing the skill of players and teams. By making use of data to assist in decision-making, players' and teams' strengths and weaknesses can be evaluated, and accordingly, changes can be made to training sessions with the aim of improving performance.

Most major professional sports teams use staff members dedicated to apply statistical methods to help players and managers making better decisions, both before and during matches. The complexity of the data used is increasing, and the latest technologies, including big data, machine learning, and artificial intelligence, have opened up for more sophisticated analyses (STATS LLC 2017). For association football, several aspects are of researchers' interest, including the choice of playing style, prediction of goal-scoring chances, and determination of players' market value. By analyzing the game better, teams can obtain a competitive advantage, and research that provide a better understanding of the dynamics of the game is therefore of great importance.

This paper advances the research frontier with respect to analyzing passing behavior in football. When successfully passing the ball between teammates, ball possession is kept with the purpose of creating goal-scoring opportunities and avoiding goals against. When a sequence of passes occurs, the players on a team become connected in a passing network, where nodes represent players and arcs represent successful passes between the players. The resulting mathematical objects can be analyzed to obtain novel insights. This paper focuses on the use of network flow motifs to derive such insights.

Gyarmati et al. (2014) introduced flow motifs, being inspired by the concept of network motifs proposed by Milo et al. (2002). Considering a passing sequence, a flow motif is a subsequence of the passes where labels represent distinct players without identity. Figure 1 shows all possible combinations of flow motifs with $k = 4$ nodes. With four nodes, there are five different types of motifs: *ABAB*, *ABCA*, *ABAC*, *ABCB*, and *ABCD*. Duplicate nodes within a sequence imply that a single player is involved several times in the motif.

Applied to association football, network flow motifs provide a method for discovering patterns in teams' passing behavior. Gyarmati et al. (2014) analyzed flow motifs consisting of three consecutive passes to study teams' style of play. Using publicly available data from the top European leagues, teams' motif characteristics were investigated by comparing the prevalence of the flow

***Corresponding author: Lars Magnus Hvattum**, Faculty of Logistics, Molde University College, P.O. Box 2110, N-6402 Molde, Norway, e-mail: hvattum@himolde.no
Else Marie Håland, Astrid Salte Wiig and Magnus Stålhane: Industrial Economics and Technology Management, NTNU, Trondheim, Norway

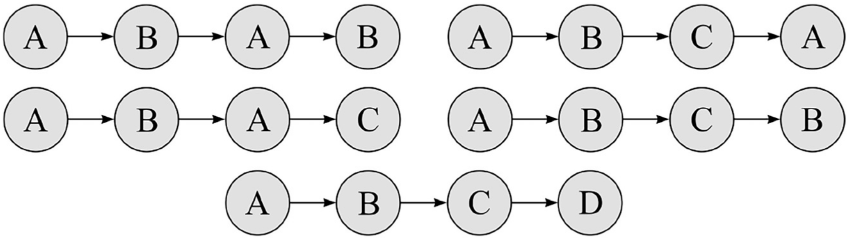

Figure 1: Motifs of size four.

motifs in the passing networks to the expected occurrence in randomly generated networks with identical properties. Furthermore, cluster analyses were performed to examine similarities and differences in teams' passing patterns. Peña and Navarro (2015) investigated whether flow motifs could be extended to the player level. By calculating the average number of a player's occurrences in each possible flow motif, the player's style of play was identified. Cluster analyses and similarity measures were used to identify which players have the most similar playing styles.

Bekkers and Dabadghao (2019) used network motifs to study teams' playing styles in both regular play and attacking phases of the game by distinguishing between possession flow motifs and flow motifs leading to goalscoring opportunities. Machine learning techniques were used to identify unique players, while radar graphs made comparisons between players' and teams' performance in the motifs available. The authors concluded that the Euclidean distance between a specific player and all other players' motif intensities can be used for scouting purposes to find players with similar characteristics.

Perdomo Meza (2017) considered the use of motifs to develop passing network profiles of teams, allowing teams to be clustered and compared. The author argued for the importance of repeatability of metrics developed, and found that the optimal size of motifs is four that is, consisting of four players and three passes. Perdomo Meza (2017) also extended previous work on flow motifs by incorporating spatial information when categorizing motifs. Malqui et al. (2019) focused on the visualization of motifs. In addition to considering the flow motifs themselves, they also took into account ball trajectories, as represented by the six endpoints of the three passes involved. Statistics about the use of motifs and trajectory clusters could then be visualized for teams, players, or matches.

Most work on flow motifs has focused on classifying playing styles. The existing literature has to a very limited degree discussed the efficiency of different motifs in terms of generating shots or goals. The contribution of this paper is to develop a generalized additive model (GAM) to investigate the influence of different explanatory variables on the effectiveness of motifs in terms of generating shots, using data from the Norwegian top division, Eliteserien.

To achieve this, previous work analyzing passes in the Norwegian top division is extended. Håland et al. (2020) evaluated the passing abilities of football players in Eliteserien through the development of three generalized additive mixed models (GAMMs). The first model evaluated pass difficulty, similarly to a model by Szczepański and McHale (2016). The second model dealt with the probability that a pass can be followed up, that is, whether the player receiving the pass is able to perform a successful follow-up action. This is referred to as the pass risk. The third model investigated pass potential, meaning the probability that the pass is part of a sequence of play that leads to a shot being taken.

Wiig et al. (2019) applied the results from the passing ability models of Håland et al. (2020) in a network analysis, similarly to the approach by McHale and Relton (2018), where nodes correspond to players, and arcs are related to the quality of passes made between the players. Three separate networks were considered, one for each of the aspects considered for passing ability: difficulty, risk, and potential. Then, to quantify each player's influence and importance in a team, centrality measures (Estrada 2011) were calculated, including closeness centrality, betweenness centrality, PageRank, and clustering coefficients.

The relationship between the work of (Håland et al. 2020, Wiig et al. 2019), and this paper can thus be summarized as follows. Håland et al. (2020) developed models to assess individual passes along three dimensions: the difficulty of the pass (how likely it is to reach a teammate), the risk of the pass (how likely it is that the recipient does not lose control of the ball), and the potential of the pass (how likely it is to lead to a shot). Being able to categorize these aspects of passes, it is possible to find players that are able to perform better than expected in their passing game. Wiig et al. (2019) used the same dataset and the outputs of the models of (Håland et al. 2020) to create passing networks for each team, where players are nodes and arcs are weighted based on the quality and quantity of the passes made between the players. Centrality measures for nodes in graphs can then be used to categorize how important each player is in the passing game of a team.

In the following, this work performs additional analysis on the level of short passing sequences. Based on inputs from past work on the evaluation of passes and the importance of players in the passing game of a team, flow motifs are analyzed. The goal is to understand what determines the success of a passing motif, whether different teams rely on passing patterns that result in different distributions of passing motifs, and whether the best teams are more likely to use the more effective motif types.

2 Experimental setup

The analysis presented in Section 2 is performed using the *mgcv* package in RStudio (Wood 2011), and the data used are from all matches in four seasons of the Norwegian top division, from 2014 to 2017. The data are in the form of events recorded in the Opta24Feed (Opta Sports 2018). A single file for each of 960 matches describes on-ball involvements as well as some off-ball events such as bookings and substitutions. The events are recorded manually, and include information about pitch locations, time stamps, and the players involved. Events identified by the Opta24Feed as passes are used in this study, including passes from open play, headed passes, long passes, and crosses.

The motifs considered consist of three passes, which must be sequential and from the same passing sequence. Moreover, the recipient of a pass must be the passer of the next pass, the players have to be from the same team, and each pass must be performed within 5 s after the previous event. The data cover 749,859 passes by 831 different players, and parsing these as passing motifs results in a total of 203,313 motifs of three passes each. Although motifs of different lengths could be studied, three passes is the length used in all previous studies on motifs in association football, and was found to be the most informative length by Perdomo Meza (2017).

2.1 Generalized additive models

Binary logistic regression is a statistical technique used to model the relationship between dependent and independent variables when the dependent variable is binary (Hosmer Jr et al. 2013). If i denotes the ith out of N observations, the dependent variable, y_i, is defined as:

$$y_i = \begin{cases} 1, & \text{if observation } i \text{ is successful } (i = 1, \ldots, N) \\ 0, & \text{if observation } i \text{ is unsuccessful } (i = 1, \ldots, N). \end{cases}$$

In general, the independent variables in a logistic regression model are related to the dependent variable via the logit link function given by:

$$\operatorname{logit}(P_i) = \ln\left(\frac{P_i}{1 - P_i}\right) = \eta_i,$$

where Pi is the conditional mean and $\eta_i = \boldsymbol{X_i\beta}$ is a function of the independent variables $\boldsymbol{X_i}$ and their corresponding coefficient estimates $\boldsymbol{\beta}$. The conditional distribution is a Bernoulli distribution where an observation's probability of success is represented by the inverse logit function given by:

$$P_i = Pr(y_{\mathrm{i}} = 1|\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{1}{1 + e^{-\eta_i}}.$$

Logistic regression is a special case of a generalized linear model (GLM). A GAM arises when allowing additive predictors, also known as smooth functions (Lin and Zhang 1999), in addition to the linear predictors of a GLM (Hastie and Tibshirani 1986). In the case of a GAM, η_i can be written as:

$$\eta_i = \boldsymbol{X_i\beta} + f_1(x_{1i}) + \ldots + f_j(x_{ji}), \tag{1}$$

where $\boldsymbol{X_i}$ is a vector of fixed effects, $\boldsymbol{\beta}$ is a vector of fixed-effect coefficients, and $f_1, \ldots, f_j$ are smooth functions of variables $x_{1i}, \ldots, x_{ji}$ (Wood 2006).

2.2 Model description

Section 2.2 describes a GAM used to evaluate passing motifs occurring in Eliteserien. The observations considered are subsequences of three uninterrupted passes, that is, motifs of length four. The dependent variable, Y^M, is binary and takes the value one if the motif leads to a shot, and zero otherwise. The shot does not necessarily have to come immediately after the observed passing sequence, but must be made within 15 s of the last pass of the sequence. Own goals are considered as shots by the team being awarded the goal.

Most of the explanatory variables used in the GAM are based on the results from passing ability models in (Håland et al. 2020) and network metrics to analyze key players in (Wiig et al. 2019). Explanations of these models and metrics are therefore given first. Håland et al. (2020) developed three GAMMs, which extends a GAM by also including random effects in the predictor. The purpose of these GAMMs was to evaluate passes, and to provide a score between zero and one for each executed pass, indicating whether the pass is likely to 1) successfully reach a player on the same team, 2) reach a player on the same team and then be followed up by any successful event, and 3) allow the team performing the pass to make a shot within 15 s of the last pass in the current passing sequence. The dependent variables of these three models are referred to as Y_1^P, Y_2^P, and Y_3^P, respectively, with values closer to 1 indicating a higher probability of success.

A wide range of explanatory variables were used in (Håland et al. 2020) to model the success of a pass. For completeness, these are listed in Table 1. For the pass number category, $s = 1$ is used for the first pass in a sequence, $s = 2$ is used for the second pass, $s = 3$ is used for the third, fourth, or fifth pass of the sequence, and $s = 4$ is used for all subsequent passes. The variables $Z_{1,k}$, $Z_{2,t}$, and $Z_{3,o}$ are normally distributed random effects.

The output of the GAMMs in (Håland et al. 2020) allows any pass to be evaluated with respect to the probability that it reaches a teammate (Y_1^P), the probability that the team performs at least one additional successful action after the pass (Y_2^P), and the probability that the team is able to make a shot shortly following the pass (Y_3^P). To generate explanatory variables for the motif model, the predicted probabilities of success for the three passes involved in the motif are added up. This is done separately for each of the three success criteria, leading to an overall evaluation of the difficulty, risk, and potential associated with the passing motif in question. Table 2 lists the three resulting explanatory variables under the name $SumY_i^P$, for $i = 1, 2, 3$. These variables are continuous, and take on values from the interval [0, 3].

Following Wiig et al. (2019), passing networks are built using outputs from the three GAMMs of (Håland et al. 2020), where connections are formed between players if they have made passes between each other, and where the weights of these connections depend on the associated values of Y_1^P, Y_2^P, and Y_3^P for the passes made. These networks are occasionally referred to as network 1, 2, and 3, to indicate that they are calculated based on weights derived

Table 1: Summary of explanatory variables used in models by Håland et al. (2020). Fixed-effect variables are denoted by X, random-effects variables by Z, and smooth terms by $f(\cdot)$. The types of variables of fixed and random effects are continuous (C), categorical/factor (F), binary (B), and interaction (I).

Variable	Description	Type
$X_{1,s}$	Pass number category in the current sequence of passes ($s = 1, 2, 3, 4$)	F
X_2	Tackle in the previous event	B
X_3	Aerial duel in the previous event	B
X_4	Interception in the previous event	B
$X_{5,i}$	The player making the pass also took part in a tackle ($i = 1$),aerial duel ($i = 2$), or interception ($i = 3$)	B
X_6	Ball recovery due to a loose ball in the previous event	B
X_7	Previous pass was a header	B
X_8	Current pass is a header	B
X_9	Player performing the pass plays for the home team	B
X_{10}	Previous event was a free kick	B
X_{11}	Previous event was a throw-in	B
X_{12}	Corner taken within the past five events	B
$X_{13,i}$	The team attempting a pass just executed a corner ($i = 1$), free kick ($i = 2$), or throw-in ($i = 3$)	B
X_{14}	The match is played on artificial grass	B
$X_{1,2}X_{10}$	Pass sequence number 2 interacting with free kick in previous event	I
$X_{1,2}X_{11}$	Pass sequence number 2 interacting with throw-in in previous event	I
$Z_{1,k}$	Player k passing the ball ($k = 1, \ldots, 689$)	F
$Z_{2,t}$	Team t the player is representing ($t = 1, \ldots, 19$)	F
$Z_{3,o}$	Opponent team o to the player passing the ball ($o = 1, \ldots, 19$)	F
$f_1(x_0, y_0, x_1, y_1)$	4-D smooth for the start (x_0, y_0) and end coordinates (x_1, y_1) of a pass	C
$f_2(x_2, y_2)$	2-D smooth representing the average position of the player given by coordinates (x_2, y_2)	C
$f_3(x_3)$	1-D smooth representing game time, x_3, in minutes	C
$f_4(x_4)$	1-D smooth for time played by the player passing the ball, x_4	C
$f_5(x_5)$	1-D smooth representing the goal difference, x_5	C
$f_6(x_3, x_5)$	2-D smooth representing the interaction between game time and goal difference	I
$f_7(x_7)$	1-D smooth for the time passed since last occurred event, x_7	C
$f_8(x_8)$	1-D smooth for the Elo rating, x_8, of the opponent team	C
$f_9(x_9)X_{14}$	1-D smooth functions representing month of play, x_9, interacting with type of grass	I

Table 2: Explanatory variables for the motif analysis. Some variables are replicated for the three passing ability models and their respective networks. These variables have an indicator of $i = (1, 2, 3)$. Players involved more than once in a motif are added accordingly to the sums considered, and the player scores are considered for the season the motif happens. Types of variables are continuous (C) and factor/categorical (F).

Variable	Description	Type
$SumY_i^P$	The sum of the predicted probabilities of success for the three passes in the motif	C
Closeness	The sum of the closeness scores for all four players involved in the motif	C
Betweenness	The sum of the betweenness scores for all four players involved in the motif	C
Clustering	The sum of the Barrat clustering coefficients of all four players	C
$PageRankRecipient_i$	The sum of the three recipients' PageRank Recipient scores	C
$PageRankPasser_i$	The sum of the three passers' PageRank passer scores, $i \neq 2$	C
MotifType	Indication of the motif type	F
Zones	The number of unique zones in which the motif takes place	C

from Y_1^P (pass difficulty), Y_2^P (pass risk), and Y_3^P (pass potential). Several centrality measures from network theory were then computed for each of the three networks. The centrality measures are calculated per player, relative to their team, and indicate how influential the player is in terms of making or receiving passes. That is, each player is scored from 0 (for the least central player on the team) to 1 (for the most central player on the team), with separate scores for each centrality measure considered.

Table 2 shows eight explanatory variables derived from centrality measures. Each of these involves adding up the centrality scores of the players involved in a motif, based on one specific centrality measure and network type. The closeness centrality of a network node is a function of the length of the paths from the node to all other nodes in the network (Freeman 1978), and can be interpreted as a measure of the easiness of reaching a particular player within a team. That is, players with higher closeness scores tend to reach more players in fewer passes (Clemente et al. 2016). For closeness, only the difficulty of passes is considered when analyzing motifs, and the values given by network 1 and Y_1^P are used to calculate a single explanatory variable by summing the closeness scores for the players involved in the motif.

The betweenness centrality of a node in a network is based on the number of shortest paths between two other nodes passing through the node (Freeman 1977). A node is considered to be central if it appears on the shortest path between many pairs of nodes. The betweenness score of a player gives an indication of how the ball-flow between teammates depends on that player, and players with high scores play are important in connecting other teammates in the passing game (Gonçalves et al. 2017). As for closeness, a single explanatory variable is calculated based on the betweenness centrality of the four players involved in a motif.

A third centrality measure is the Barrat clustering coefficient (Barrat et al. 2007). It indicates the tendency of a node to cluster together with other nodes. A high clustering coefficient of node i indicates that if nodes i and j are connected, and nodes j and h are connected, then node i is likely to be connected directly to h as well (Barrat et al. 2007). A single explanatory variable is derived from the clustering coefficients of the players involved in a motif, with higher values indicating that the players involved in the motif have a tendency to form clusters with other players.

Five of the explanatory variables are based on centrality measures derived using the PageRank algorithm (Brin and Page 1998). For PageRank, separate networks are built for evaluating players making passes and for players receiving passes (Wiig et al. 2019). PageRank centrality becomes a recursive notion of popularity or importance, where a player can be important either when receiving passes from other important players, or when passing to ball to other important players. To calculate explanatory variables based on PageRank centrality, either the three players making passes or the three players receiving passes in the motif are considered. Thus, these variables take on values in the interval [0, 3], as opposed to the variables based on betweenness, closeness, and clustering, which take on values in [0, 4]. For recipients of passes, all three networks are used to derive separate explanatory variables, whereas for players making passes, only the networks for pass difficulty and pass potential are used.

The explanatory variables derived from (Håland et al. 2020; Wiig et al. 2019) are all continuous variables and are treated as smooth terms in the GAM. The aims of including the smooth terms for $SumY_1^P$, $SumY_2^P$, and $SumY_3^P$ are to test how the effectiveness of motifs is influenced by the difficulty of the passes made, the risk taken by the players involved, and the potential of the passes made. The

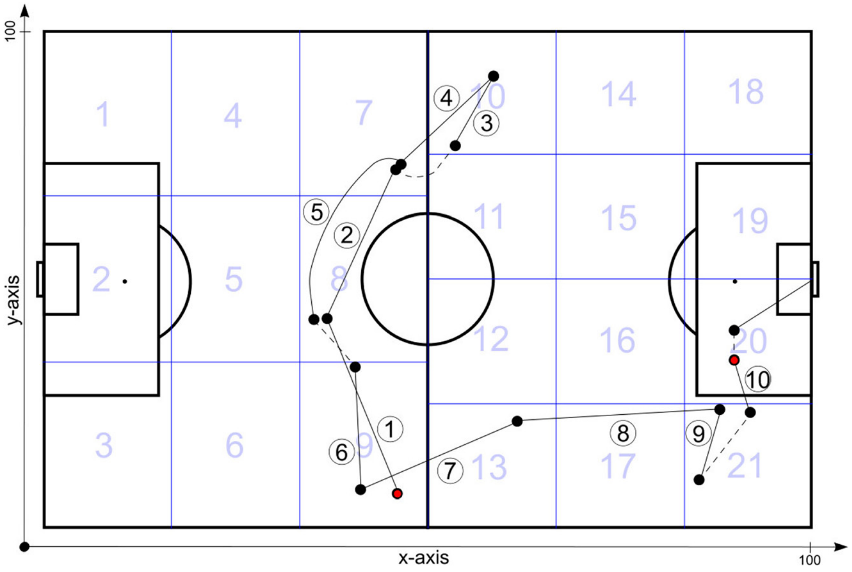

Figure 2: The pitch coordinate system divided into 21 zones. Direction of play is left to right, always relative to the team in possession of the ball. Also shown is a sequence of 10 passes from a match in Eliteserien.

network metrics obtained for players in (Wiig et al. 2019) are used to test whether the involvement of key players in a motif has an effect on its outcome.

Two additional explanatory variables are included. First, a categorical variable is added to identify the motif type. With a motif size of four, five different motif types are possible as illustrated in Figure 1. The reference is chosen to be the ABCD motif, i.e., with four unique players being involved. Second, to test whether the area covered by the players involved in a motif has an impact on its effectiveness, a variable counting the number of unique zones on the pitch in which the players have been active during the motif is added. The pitch is divided into 21 unique zones, as shown in Figure 2, and a total of six zones may be involved in a flow motif: the start and end zone of each of the three passes. Both the start and end zones are included as players might receive the pass in one zone, perform a ball touch or ball carry, and then pass the ball further from another zone within the 5-s time frame. Although the number of zones for a motif is discrete, the variable for zones is interpreted as a continuous variable, to allow its use in a GAM smooth function. Fixed-effect variables are selected through the use of a Wald test (Hosmer Jr et al. 2013), while the smooth terms are not tested as fixed effects.

Figure 2 also illustrates a sequence of 10 passes made by Rosenborg against Viking in a match played in August 2016. The sequence started with a free kick and ended with a shot on goal, after involving seven unique players. A total of eight flow motifs can be recorded from the sequence, the first of which involves four unique players in the pattern ABCD and four distinct zones. Passes labeled 3 and 5 were made by the same player, so the flow motif starting with the second pass is encoded as ABCB, spanning three distinct zones. The details of each motif from this passing sequence are given in Table 3.

3 Results and discussion

Before presenting the main results, the model is first validated using the area under the curve (AUC) for the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve (Hosmer Jr et al. 2013). From Figure 3, the AUC for the ROC curve indicates an acceptable fit according to the guidelines suggested in (Hosmer Jr et al. 2013). The

Table 3: Details of players and motifs in the passing sequence illustrated in Figure 2.

No.	Player	Action	Motif							
			1	2	3	4	5	6	7	8
1	Svensson	Free kick	A							
2	Reginiussen	Pass	B	A						
3	Eyjólfsson	Pass	C	B	A					
4	Gersbach	Pass	D	C	B	A				
5	Eyjólfsson	Pass		B	A	B	A			
6	Reginiussen	Pass			C	C	B	A		
7	Svensson	Pass				D	C	B	A	
8	Jensen	Pass					D	C	B	A
9	Gytkjær	Pass						D	C	B
10	Helland	Pass							D	C
—	Gytkjær	Shot								B
No. of zones:			4	3	3	4	4	3	3	3

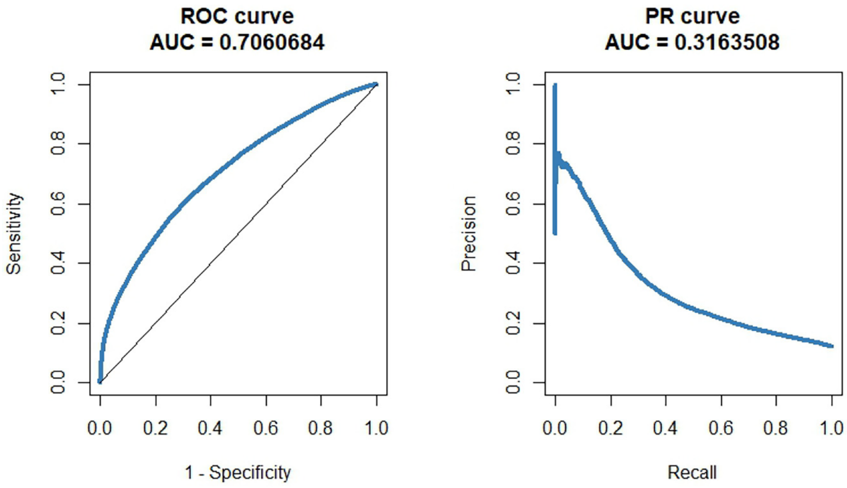

Figure 3: ROC and PR curves for the passing motif model.

PR curve gives a low AUC due to a highly skewed data distribution. However, the points at which the PR curve has to start and end are not giving much room for a high AUC, which makes the value obtained seem appropriate. Model calibration refers to the property that the number of predicted positives and the number of actual positives should be approximately equal for all ranges of probabilities. A Hosmer–-Lemeshow test (Hosmer Jr et al. 2013) has been performed to test the model calibration, considering different groups of observations, under the null hypothesis that the model provides a good fit. Following Bartley (2014), 100 random samples are drawn, and the test indicates a good fit of the model if less than 10 of the random samples result in a rejection of the null hypothesis at a significance level of 5%. This was the case both when using sample sizes of 1000 and 5000 observations, thus indicating an acceptable calibration.

Table 4: The regression results from the motif analysis. Significance level is indicated by *.

Fixed effects	
Variable	**Coefficient(SE)**
MotifTypeABAB	−0.111*(0.046)
MotifTypeABCB	−0.095***(0.022)
Intercept	−0.503***(0.067)
Smooth terms	
Variable	**Sign.**
$SumY_1^P$	***
$SumY_2^P$	***
$SumY_3^P$	***
Closeness	*
Betweenness	*
Clustering	***
$PageRankRecipient_1$	***
$PageRankRecipient_2$	***
$PageRankRecipient_3$	***
$PageRankPasser_1$	***
$PageRankPasser_3$	***
Zones	***

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

3.1 Regression results

The resulting GAM is built on 203,208 observations, which is a slight reduction from the initially observed number of motifs of size four in the data. The reduction is due to some passes not having defined probabilities of success in accordance with the GAMMs explored in (Håland et al. 2020). The regression results are shown in Table 4. A significance level of 10% is used, and negative values of the fixed-effect coefficients imply a reduced probability of success, that is, the motif leading to a shot.

3.1.1 Fixed terms

In the final model, two of the motif types have been removed after applying the Wald test. Hence, three motif types form the reference for the variable. The initially chosen reference (*ABCD*) is characterized by having four distinct players involved in the motif. This pattern is more likely to cover a larger area of the pitch. This is

Table 5: The average number of zones covered by each type of motif.

	ABAB	ABAC	ABCA	ABCB	ABCD
Count	6152	32518	18335	31038	115270
Avg. zones	2.29	2.88	2.89	2.94	3.52

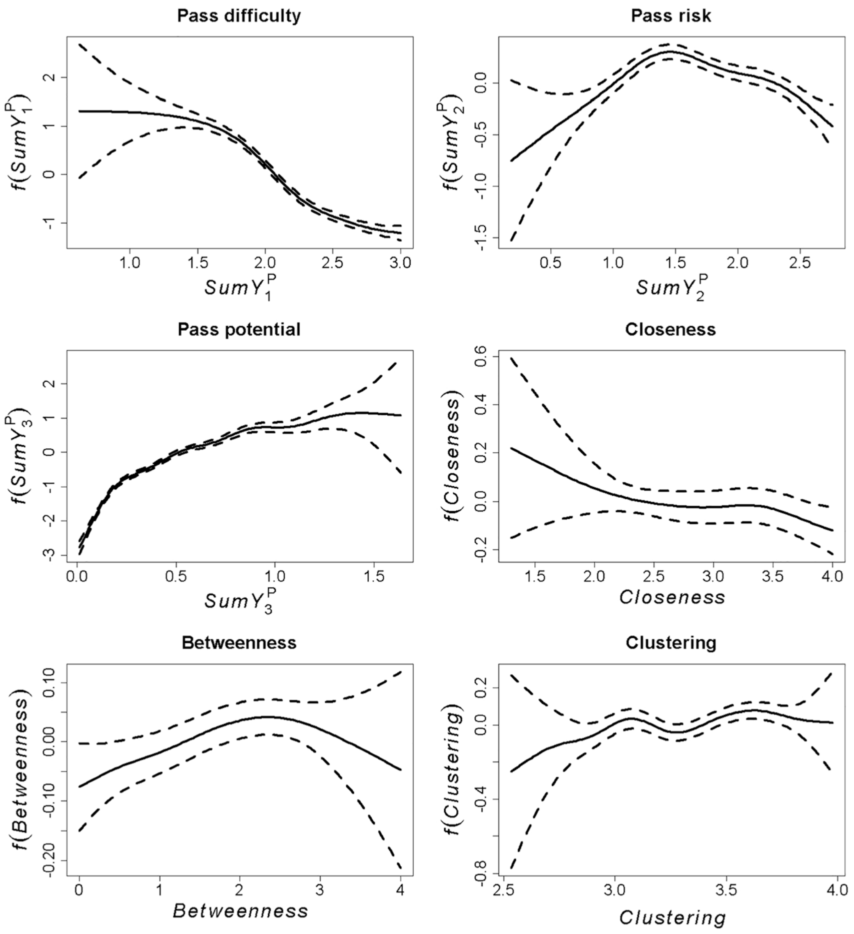

Figure 4: The resulting smooth functions from the motif analysis. The dotted lines indicate the 95% confidence intervals of the functions.

corroborated by Table 5, which shows that *ABCD* covers more zones on average than the other motifs. For all the motifs, the minimum observed number of zones covered is one, with all three passes starting and ending within the same zone, and the maximum is six, with the three passes starting and ending in all different zones.

Motif type *ABCA*, which is added to the reference, is similar to *ABCD* as only the first and last involved player is the same. For the second addition to the reference, (*ABAC*), it is more difficult to understand how this motif type is indistinguishable from the two others in the reference as its pattern is seemingly more similar to the categories that remain in the final model. Both the *ABAB* and *ABCB* motif types are left in the model and have negative signs on their corresponding coefficients. Thus, the more compact motif type, *ABAB*, seems to be less effective in terms of resulting in shots.

3.1.2 Smooth terms

The resulting smooth functions from the motif analysis are displayed in Figure 4 and Figure 5. Higher outcomes from the functions imply an increased probability for a motif to lead to a shot. Considering the difficulty of passes in a motif, the shape of the corresponding smooth function is intuitive. As higher sums indicate more easy passes, the probability of success increases when more difficult passes are made. Passes are in general easier to make further away from the opponent's goal. Thus, having several passes with high probabilities of success might indicate that the motif takes place far away from the opponent's goal, and the shape of the smooth function is thus intuitive as shots are more likely to be attempted near the opponent's goal.

Similarly reasonable results are also present for the pass risk measures when looking at the region with a tight confidence interval. As Y_2^P measures how likely a pass is to be followed up by a successful event, high values of $SumY_2$ indicates that the passes involved in the motif were unlikely to lead to a loss of possession. Compared to an average motif, Figure 4 shows that motifs consisting of passes that are easier to follow up are less likely to lead to shot opportunities. This is plausible due to the same reasoning as used for pass difficulty, as passes that are easier to make also tend to be easier to follow up. However,

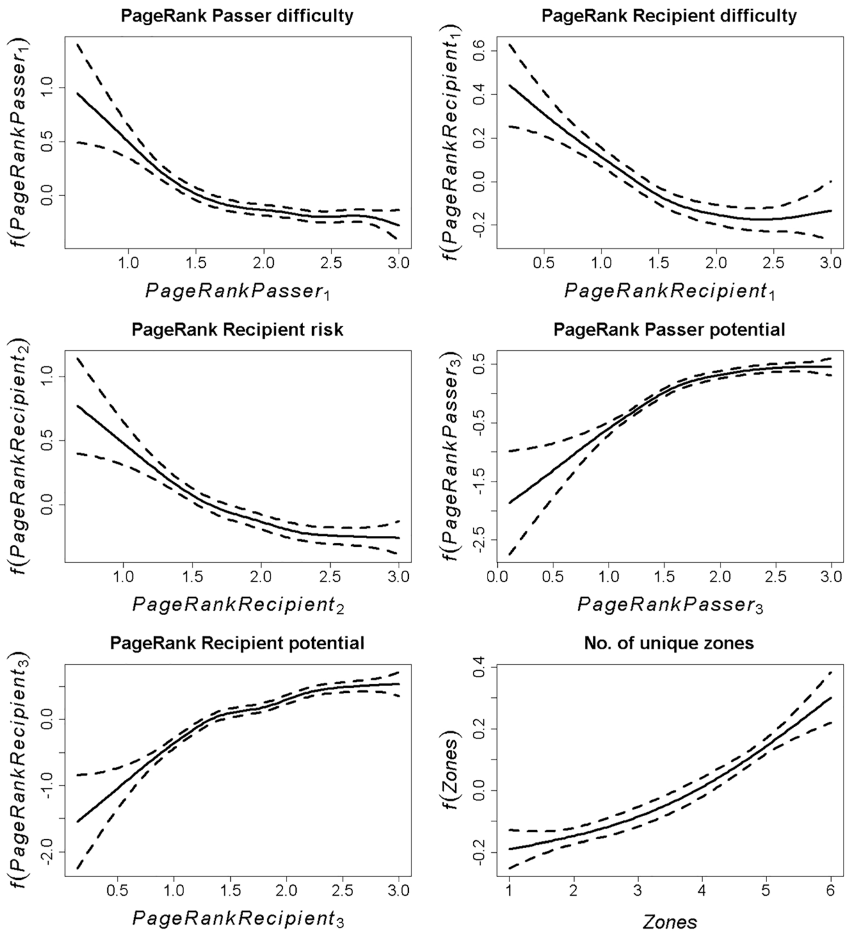

Figure 5: The resulting smooth functions from the motif analysis. The dotted lines indicate the 95% confidence intervals of the functions.

this trend is not evident for low values of $SumY_2$, which also give a negative contribution to the likelihood of a motif ending with a shot. There are two explanations for this. One is that this region has much larger confidence intervals, indicating fewer observations and more uncertainty in the effect. The second is that too many risky passes means that the motif is likely to end with a loss of possession, rather than a shot. Thus, the best motifs may be those involving a medium level of risk. Intuitively, with many high-potential passes in a motif, their potential should be positively related to the motif's potential for leading to a shot. This belief is captured by the model as evidenced by the positive slope for the smooth function.

For the closeness centrality, the slope of the function is negative, indicating that higher values of this measure for the players involved are associated with lower probabilities of success, although the slope is gentle within the region of narrower confidence interval. Wiig et al. (2019) found that players in more offensive player positions received higher scores on this measure, thus the negative slope may indicate that defensive or central players are more often involved in successful motifs. Also, the highest sum possibly obtained for the measure is four, which is hard to obtain as mostly only one player on each team has received the maximum score of one. Hence, all players involved in the motif must have received similarly high closeness scores to reach a sum close to four, or fewer distinct players must have been involved in the motif. In fact, most of the highest sums for this measure are found for the *ABAB* motif type, which is the motif type identified as giving the lowest probability of success.

Considering the betweenness centrality, the corresponding smooth function has a concave shape where the likelihood of success increases for lower values and decreases for higher values. Hence, the chance of a motif to be effective is highest when the sum of the betweenness scores for the players involved is moderate. For the betweenness scores calculated for the Norwegian top division, Wiig et al. (2019) observed that the scores varied among player positions and that players could receive a score of zero even if they had a high number of pass involvements compared to other players on their team. Some of the players who received low scores, however, did receive high scores for the PageRank effectiveness measures. Hence, players that are effective do not necessarily

Table 6: The distribution of motif types and the total number of four-sized motifs for each team in the 2017 season. The three highest percentages of each motif type and shot rates per game are highlighted in bold text, and the teams are ordered by their end-of-season table positions. Shots per game (SpG) for all teams are obtained from WhoScored.com (2018).

	Team	*ABAB*	*ABAC*	*ABCA*	*ABCB*	*ABCD*	*Obs*	*SpG*
1	Rosenborg	0.025	0.153	0.080	0.154	0.588	4930	13.3
2	Molde	0.024	0.158	0.088	0.156	0.574	3469	13.0
3	Sarpsborg 08	**0.043**	0.165	0.088	**0.161**	0.542	3539	13.1
4	Strømsgodset	0.026	0.162	0.090	0.149	0.574	3489	13.0
5	Brann	0.036	**0.174**	0.089	0.159	0.543	2945	13.2
6	Odd	0.034	**0.177**	0.095	0.156	0.537	3791	10.9
7	Kristiansund	0.035	0.161	0.092	0.160	0.552	1898	11.4
8	Vålerenga	0.029	0.158	0.096	0.147	0.570	5156	13.0
9	Stabæk	**0.041**	**0.177**	0.089	**0.166**	0.526	3845	**14.5**
10	Haugesund	**0.042**	0.161	**0.111**	0.158	0.528	2212	12.9
11	Tromsø	0.035	0.172	**0.100**	**0.161**	0.533	3255	**13.8**
12	Lillestrøm	0.039	0.158	**0.098**	0.152	0.552	1117	**15.2**
13	Sandefjord	0.022	0.151	0.090	0.143	**0.594**	2420	10.1
14	Sogndal	0.025	0.145	0.094	0.135	**0.602**	1723	11.9
15	Aalesund	0.031	0.162	0.069	0.147	**0.592**	2619	12.7
16	Viking	0.030	0.171	0.084	0.160	0.554	3185	11.2

also have high betweenness scores, which could explain the shape of the function.

The slope of the smooth function for the clustering coefficient is more or less positive, implying that higher values of this measure correspond to an increased likelihood for a motif to be effective. This is a reasonable result as higher values of the clustering coefficients are associated with a well-balanced team where the teammates have strong connections to each other.

For the PageRank passer and recipient measures for network 1 (pass difficulty) and the PageRank recipient centrality for network 2 (pass risk), the probability of success decreases with higher sums of the measures. The involvement of key players in terms of these PageRank measures reduces motif effectiveness. As defenders tend to receive higher scores on the PageRank passer measure, while offensive players tend to obtain higher PageRank recipient scores for network 1 (pass difficulty), the two graphs for the pass difficulty network are a bit contradictory. The negative slopes thus imply that the optimal strategy would be to have offensive players passing the ball and defenders receiving the ball in the offensive play. Hence, only the function for the PageRank passer measure can be seen as intuitive. Higher values of the PageRank passer and recipient measures for network 3 (pass potential), however, contribute to an increased probability of success as seen from the positive slopes. This is intuitive as the PageRank scores for network 3 are based on the frequency of players' involvements in effective passes.

The likelihood of success increases with the number of unique zones on the pitch in which the players have been active during the motif. By utilizing a bigger area, a team might take advantage of open areas by making tactically better passes as the players tend to be more mobile in between passes for such cases. The shape of the function supports the findings from the fixed-effect variables as larger areas covered seem to be beneficial for success. As there is a time limit of 5 s between each pass in a motif, large movements could indicate that counter-attacks have been performed. If so, the model suggests that counter-attacks are effective, which is intuitive.

3.2 Distribution of motif types in Eliteserien

The distribution of four-sized motifs for each team playing in Eliteserien 2017 is shown in Table 6. Additionally, the number of shots attempted per game is given for the teams. The aim of presenting these statistics is to investigate whether the top performing teams, seen in light of the regression results, tend to use some specific motifs to a higher extent compared to the other teams in the league.

Clearly, for all teams the proportion of the *ABCD* motif is the highest. In fact, more than half of the performed motifs by each team in the season involve four distinct players. Among the top teams of the season, Rosenborg, Molde, and Strømsgodset have quite similar distributions of motif types used, whereas Sarpsborg 08 stand out by having the highest internal percentage of the *ABAB* motif. Interestingly, Stabæk has the second best rate of shots per game, but has some of the highest internal proportions of motif types where less distinct players are involved, indicating that they have a more compact playing style. This observation contradicts the results from the regression where these motif types are found to be less likely to lead to shots.

The three teams with the highest internal proportion of using the *ABCD* motif are situated in the bottom of the table, and they have relatively low rates of shots per game. Moreover, some of the teams with the lowest internal percentages have the highest number of shots per game. Considering that this motif type is one of those that are most likely to be effective, these numbers are surprising. One would expect that the teams being more effective in terms of generating shots would be inclined to use the most effective motif types. However, which motif types the teams actually succeed with in terms of scoring goals are not considered. Nevertheless, two out of the three teams with the highest shot rates do use the *ABCA* motif more frequently, while the third team has the highest internal ratio of the *ABAC* motif. Both of these motif types are included in the reference of the model developed and have thus the highest probabilities of leading to a shot.

3.3 Comparison with existing literature

Pina et al. (2017) used network metrics as fixed terms in a logistic regression model to test how they affect the success of offensive plays in football. Such plays cover entire passing sequences and not parts of them like motifs do. A limited number of network metrics was utilized, and only the density score of the team performing the sequence was found to be significant. With a negative coefficient, a higher density for the team, or interconnectedness between the players, implies less chance of succeeding with the offensive play. Although not being the same types of centrality measures, the closeness and the PageRank measures for network 1 (pass difficulty) and network 2 (pass risk) also turned out to have negative slopes.

By using motifs of size four, Gyarmati et al. (2014) studied teams' playing styles in the Spanish La Liga. FC Barcelona, the league winners, stand out by using the three compact motif types more often than the other teams in the league. Interestingly, the team use the motifs *ABCA* and *ABCD*, which were found to be more likely to be effective in the Norwegian top division, less than the other teams. However, even though being the league's top scorers, the team is ranked in sixth place in terms of the number of shots per game for the season considered (WhoScored.com 2018). Thus, the team's compact playing style does not seem to generate more shot attempts, an observation that is supported by the results from the motif analysis in this paper.

Bekkers and Dabadghao (2019) investigated teams' playing styles by studying possession motifs and motifs resulting in a shot immediately after the last pass. The results are not directly comparable, as Bekkers and Dabadghao (2019) looked at the probability of scoring given that a shot is taken after a sequence of up to three passes. Their results indicate that shots taken following motifs with fewer players involved are of worse quality. This adds to the finding in this paper that such compact sequences are also less likely to lead to shots.

Other than the observation that the more compact motif types tend to be less effective in terms of leading to a shot, the findings from the analysis in Section 3.3 are seemingly new to the literature on passing motifs. For instance, it is found that the more space utilized on the pitch during a motif, the more likely it is of being effective. Hence, counter-attacks seem to be proven more likely to lead to shot attempts. Also, difficult passes and passes that are more challenging to follow up appear to be more effective in a motif. The results suggest that teams should look for smart passing alternatives such that they can attempt combinations of passes that enable them to advance fast on the pitch.

3.4 Final evaluation

Three goals of this research were stated in Section 1. The first goal is to understand what determines the success of a passing motif. The second goal is to see whether different teams rely on passing patterns that result in different distributions of passing motifs. Finally, the third goal is to see whether the best teams are more likely to use the more effective motif types. To find which factors affect the outcome of a passing motif, a GAM was built on all passing motifs of size four in the data set.

Regarding the factors determining the success of a passing motif, the results indicate that the pattern of the motif does influence its outcome in terms of generating shots. More compact motif types with fewer distinct players involved tend to be less effective. This is also supported by the smooth function for the number of unique zones covered in the motif.

All smooth functions based on the predicted probabilities of success have reasonable shapes. Passes that are easier to make (higher probability of reaching a teammate) and that are easier to follow up (higher probability of a successful event following the pass) are associated with a decrease in the probability of success. That is, motifs with such passes are less likely to lead to a shot. Having passes with a higher potential of leading to a shot in a motif naturally increases the likelihood for the motif to be effective as well.

The participation of key players in a motif has differing results for the different network metrics considered. However, the tendency is that more intuitive effects on the effectiveness of motifs are present for the metrics where offensive contribution already is taken into account, such as the PageRank metrics for the networks related to risk and potential. Some of the network metrics are highly correlated, as pointed out by Wiig et al. (2019), which might be the reason why some counter-intuitive results are present.

For the second goal, whether teams have different distributions of passing motifs, Table 6 shows that there is some variation between the teams. However, the differences are relatively minor, in particular when compared to the variance in the number of motifs observed for each team. All teams use the *ABCD* motif the most, and the *ABAB* motif the least.

Turning to the third goal and the question regarding whether top performing teams in Eliteserien are inclined to use the more effective motif types, the distribution of motif types used for each team in the 2017 season reveals that there is ostensibly no connection between teams' end-of-season table positions and their distribution of motifs types. The teams with the highest shot rates have lower internal proportions of the *ABCD* motif compared to most of the other teams, but they do have higher proportions of the two other reference types, *ABCA* and *ABAC*, which are equally affecting the outcome of a motif.

4 Concluding remarks

Passing in association football can be modeled using networks. Sequences of passes, where the same player may appear in the sequence several times, are considered as a network motif, following (Milo et al. 2002). To analyze what influences the effectiveness of four-sized passing motifs in the Norwegian top division, Eliteserien, a GAM was built using data from four seasons, spanning the years 2014–2017. A total of 203,208 motifs were included in the analysis. Most of the explanatory variables in the model were based on the results from the passing ability models of (Håland et al. 2020) and the network analyses of (Wiig et al. 2019).

The main finding of the analysis was that the more compact motif types, where fewer distinct players are involved, have a lower likelihood of being effective in terms of leading to shots. However, there was no clear connection between teams' table rankings and their distribution of motif types. Overall, the teams with higher shot rates had a higher internal proportion of some of the more effective motif types, although these teams did not score the most goals. Hence, a team's ability to convert a chance into a goal is important so that the team actually is able to take advantage of effective motif types. Nevertheless, teams may look into which motif types are the most effective, and consider how their passing system can be made more efficient. This may provide the top-scoring teams with more goal-scoring opportunities of which they can take advantage.

An important limitation of this work is that only one league is studied, and that this league is not considered to be among the best leagues in Europe. That is, passing behavior may be different in better leagues, and the relative effectiveness of network motifs may differ as a result. This is also the first study that aims to evaluate the effectiveness of passing motifs, and additional studies on other data sets may be necessary to increase the confidence in the results.

In general, adding more variables that potentially could influence the effectiveness of passing motifs should be considered to improve the model fit. For instance, it would be interesting to explicitly investigate whether passing motifs that are parts of counter-attacks are more likely to be effective. However, this would call for the need of additional types or sources of data. By adding mixed effects to account for the teams involved in a motif, the effectiveness of teams could be explored. If combining this in an interaction with motif types, it would be revealed which teams are more effective in using the different motifs. However, as some of the motifs are used far less than others, the amount of data needed for the analysis should be higher than only four seasons.

A different type of motifs that might be considered is zone motifs. Rather than having patterns of players, the patterns could consist of zones on the pitch. Some preliminary calculations showed that the vast number of potential combinations when having 21 zones and a motif size of four made the analysis too extensive. By finding a way of dealing with the combinations, for instance by selecting a range of them or using fewer zones, such an analysis could provide insight into where on the pitch the most effective motifs take place, regardless of which players are involved.

As the more compact motif types were found to less effective, it could be interesting to perform a similar analysis using goals rather than shots as a criterion for the dependent variable and compare the analyses. Perhaps the compact motif types turn out to be more effective in terms of leading to a goal. If so, it could be tested whether teams with high shot effectiveness, i.e., with many goals scored compared to the number of attempted shots, are correct to not use the motif types that were found to be more effective in this work.

Acknowledgments: The authors are grateful for the helpful comments and suggestions made by the associate editor and the reviewers on an initial version of this manuscript. Opta Sports are thanked for providing access to relevant data from Eliteserien, without which this work could not have been performed. Football coach Hugo Pereira is thanked for his input and feedback during the research. Jørgen Skålnes is thanked for his assistance in executing additional calculations during revisions of the work.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

Barrat, A., M. Barthelemy, and A. Vespignani. 2007. “The Architecture of Complex Weighted Networks: Measurements and Models.” In *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science*, 67–92. World Scientific.

Bartley, A. 2014. “Evaluating Goodness-Of-Fit for a Logistic Regression Model Using the Hosmer-Lemeshow Test on Samples from a Large Data Set.” MSc thesis. Ohio, USA, The Ohio State University.

Bekkers, J., and S. Dabadghao. 2019. “Flow Motifs in Soccer: What Can Passing Behavior Tell Us?” *Journal of Sports Analytics* 5: 299–311.

Brin, S., and L. Page. 1998. “The Anatomy of a Large-Scale Hypertextual Web Search Engine.” *Computer Networks and ISDN Systems* 30(1–7): 107–17.

Clemente, F. M., F. M. L. Martins, and R. S. Mendes. 2016. *Social Network Analysis Applied to Team Sports Analysis*. Netherlands: Springer International Publishing.

Estrada, E. 2011. *The Structure of Complex Networks: Theory and Applications*. New York, USA: Oxford University Press.

Freeman, L. C. 1977. “A Set of Measures of Centrality Based on Betweenness.” *Sociometry* 40: 35–41.

Freeman, L. C. 1978. “Centrality in Social Networks Conceptual Clarification.” *Social Networks* 1(3): 215–39.

Gonçalves, B., D. Coutinho, S. Santos, C. Lago-Penas, S. Jiménez, and J. Sampaio. 2017. “Exploring Team Passing Networks and Player Movement Dynamics in Youth Association Football.” *PloS One* 12(1): e0171156.

Gyarmati, L., H. Kwak, and P. Rodriguez. 2014. “Searching for a Unique Style in Soccer.” In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD) Workshop on Large-Scale Sports Analytics*, August. arXiv preprint arXiv:1409.0308.

Håland, E. M., A. S. Wiig, M. Stålhane, and L. M. Hvattum. 2020. “Evaluating Passing Ability in Association Football.” *IMA Journal of Management Mathematics* 31: 91–116.

Hastie, T., and R. Tibshirani. 1986. “Generalized Additive Models.” *Statistical Science* 1: 297–318.

Hosmer, D. W., Jr, S. Lemeshow, and R. X Sturdivant. 2013. *Applied Logistic Regression*, Vol. 398. Hoboken, New Jersey: John Wiley & Sons.

Lin, X., and D. Zhang. 1999. “Inference in Generalized Additive Mixed Modelsby Using Smoothing Splines.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(2): 381–400.

Malqui, J. L. S., N. M. L. Romero, R. Garcia, H. Alendar, and J. L. D. Comba. 2019. “How Do Soccer Teams Coordinate Consecutive Passes? A Visual Analytics System for Analysing the Complexity of Passing Sequences Using Soccer Flow Motifs.” *Computers & Graphics* 84: 122–33.

McHale, I. G., and S. D. Relton. 2018. “Identifying Key Players in Soccer Teams Using Network Analysis and Pass Difficulty.” *European Journal of Operational Research* 268(1): 339–47.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. “Network Motifs: Simple Building Blocks of Complex Networks.” *Science* 298(5594): 824–7.

Opta Sports. 2018. *World Leaders in Sports Data*. https://www.optasports.com/ (accessed April 13 2018).

Peña, J. L., and R. S. Navarro. 2015. “Who Can Replace Xavi? a Passing Motif Analysis of Football Players.” arXiv preprint arXiv:1506.07768.

Perdomo Meza, D. A. 2017. “Flow Network Motifs Applied to Soccer Passing Data.” In *Proceedings of MathSport International 2017 Conference*, edited by C. De Francesco, L. De Giovanni, M. Ferrante, G. Fonseca, F. Lisi, and S. Pontarollo, 305–19. Padova, Italy: Padova University Press.

Pina, T. J., A. Paulo, and D. Araújo. 2017. “Network Characteristics of Successful Performance in Association Football. A Study on the UEFA Champions League.” *Frontiers in Psychology* 8: 1173.

STATS LLC. 2017. “AI and the Growing Use of Technology in Sport.” https://www.stats.com/industry-analysis-articles/ai-growing-use-technology-sport/ (accessed April 11 2018).

Szczepański, Ł., and I. McHale. 2016. “Beyond Completion Rate: Evaluating the Passing Ability of Footballers.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(2): 513–33.

WhoScored.com. 2018. Whoscored.com. Also available at https://www.whoscored.com/.

Wiig, A. S., E. M. Håland, M. Stålhane, and L. M. Hvattum. 2019. “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.

Wood, S. N. 2006. *Generalized Additive Models: An Introduction with R*. Boca Raton, Florida: Chapman and Hall/CRC.

Wood, S. N. 2011. “Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(1): 3–36.