NTNU
Norwegian University of
Science and Technology

Evaluating Passing Behaviour in Association Football

Else Marie Håland
Astrid Salte Wiig

Industrial Economics and Technology Management
Submission date: June 2018
Supervisor: Magnus Stålhane, IØT
Co-supervisor: Lars Magnus Hvattum, Molde University College

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

Problem Description

Passing is the most frequent event happening during a football match, and by successfully passing the ball forward on the pitch, the chance of creating goal-scoring opportunities increases. The purpose of this thesis is to find ways to evaluate players' passing behaviour. The results can provide coaches and players with valuable information, with the aim of increasing performance.

Preface

This master's thesis was written for the Department of Industrial Economics and Technology Management at the Norwegian University of Science and Technology during the spring of 2018. The thesis finalises the authors' Master of Science and is a continuation of the work done for a project thesis in the autumn of 2017. Both authors have technical background in Energy and Environmental Engineering and are specialising in Empirical and Quantitative Methods in Finance.

The work done in the thesis concerns the use of analytical methods in the context of association football, and it was initiated by Rosenborg Ballklub, a professional Norwegian association football team. Rosenborg Ballklub wanted to explore new approaches to assess their players in order to improve performance. Hence, the thesis is the result of a cooperative initiative between Rosenborg Ballklub and the Department of Industrial Economics and Technology Management.

Trondheim, 08.06.2018

Else Marie Håland
Astrid Salte Wiig

Else Marie Håland
Astrid Salte Wiig

Summary

Association football is the world's most popular sport, and there is a growing interest in the use of statistical methods to support decision-making within the sport to gain competitive advantages. In this thesis, the passing behaviour of football players is evaluated. The passing abilities of players, key players on teams and effectiveness of passing motifs in the Norwegian top division Eliteserien have been analysed through regression models and network analyses. The results from the different analyses build upon each other and are based around three aspects of a pass's success: accuracy, game overview and effectiveness. For the analyses, event-data from four consecutive seasons of Eliteserien (2014-2017) has been utilised.

Three generalised additive mixed models have been developed to assess players' passing abilities. Each model considers one of the three aspects of success, and the models were built on 565,720 observations in the data from the 2014-2016 seasons and tested on data from the 2017 season. The AIC criterion is used to determine whether a variable should be treated as a smooth term or a fixed effect, and Wald tests are performed for variable elimination. In general, the signs of the coefficients and the shapes of the smooth functions in the resulting models make sense. The models are used to identify the top ten pass makers in the 2017 season of Eliteserien, and the passing abilities of players over a period of time are determined by the ratio of the number of observed successful passes for a player over the models' expected number of successful passes for that player in the given time period.

The predicted probabilities obtained for passes in the passing ability models, together with the number of passes between players, were used as weights in network analyses. For each team and season in the data set, a network analysis was performed for each aspect of success to determine the key players in a team based on different network metrics. The closeness, betweenness and PageRank centrality measures and the Barrat clustering coefficient have been considered. Results show that the player positions of the key players vary among the measures. Offensive players tend to be ranked higher on the closeness centrality and the PageRank measure for recipients, while defenders, with many easy passes in between each other, tend to be ranked higher on the PageRank measure for passers.

Rosenborg Ballklub players show some tendencies that differ from the general results across the teams in Eliteserien as defenders seem to be more involved in the offensive play. Case studies give an indication that the team depends more upon players on the right-hand side in matches where they had a low ball possession.

To analyse what influences the effectiveness of four-sized passing motifs in Eliteserien, a generalised additive model was built using data from all four seasons. A total of 203,208 motifs were included in the analysis. Most of the explanatory variables in the model are based on the results from both the passing ability models and the network analyses. The findings indicate that the more compact motif types, i.e. where less unique players are involved, are less likely to lead to shots. However, there is no apparent relation between teams' internal usage of effective motif types and their end-of-season table position.

Sammendrag

Fotball er den mest populære sporten i verden, og det er en økende interesse for bruken av statistiske metoder innenfor sporten med formål om å skape et bedre beslutningsgrunnlag som kan gi lag konkurransefortrinn. I denne masteroppgaven blir fotballspilleres pasningsevner vurdert, de mest sentrale spillerne for lag funnet og effektiviteten til pasningsmotiv studert gjennom logistiske regresjonsmodeller og nettverksanalyser. Resultatene fra de ulike analysene bygger på hverandre og de er sentrert rundt tre aspekt av en pasnings suksess: nøyaktighet, overblikk og effektivitet. Event-data fra fire etterfølgende sesonger av den øverste norske divisjonen Eliteserien (2014-2017) blir brukt i analysene.

For å evaluere spilleres pasningsevner, er det blitt utviklet tre generaliserte additive miksede modeller. Hver av disse tar for seg et av de definerte aspektene av suksess, og modellene er bygget på 565,720 observasjoner i dataen fra 2014-2016 sesongene og testet på data fra 2017 sesongen. AIC kriteriet er brukt for å bestemme om variabler skal behandles som kategoriske eller som glatte funksjoner, og Wald tester er utført for å eliminere variabler fra modellene. Generelt sett er fortegnet på koeffisientene og formen på de glatte funksjonene i de resulterende modellene intuitive. Modellene blir brukt til å identifisere de ti beste pasningstakerne i Eliteserien 2017, og pasningsevnene til spillerne for en gitt periode blir bestemt ut fra raten mellom antall vellykkede pasninger spilleren har gjort i perioden og det antallet pasninger som modellene forventer at spilleren skal ha klart.

De predikerte sannsynlighetene for suksess fra modellene for pasningsevnene blir, sammen med antallet pasninger mellom to spillere, brukt videre som vekter i nettverksanalyser. For hvert lag og hver sesong er det satt opp tre nettverk, ett for hvert aspekt av suksess. Målet med disse er å finne de mest sentrale spillerne for lagene basert på ulike nettverksmål. *Closeness*, *betweenness*, *PageRank* og *clustering* er de nettverksmålene som er tatt i betraktning. Resultatene indikerer at spillerposisjonene til de øverst rangerte spillerne varierer blant målene. Offensive spillere har en tendens til å bli rangert høyere på closeness og PageRank for mottakere, mens forsvarsspillere, som ofte sentrer lette pasninger til hverandre, har en tendens til å bli rangert høyere på PageRank for pasningstakere.

Spillerne i Rosenborg Ballklub viser tendenser som avviker noe fra de generelle resultatene funnet for Eliteserien ettersom forsvarsspillere virker å være mer involvert i det offensive spillet. To case-studier gir indikasjoner på at laget later til å være mer avhengig av spillerne på høyre side i kamper hvor de har lav ballbesittelse.

Effektiviteten til pasningsmotiv av størrelse fire i Eliteserien har blitt analysert ved å utvikle en generalisert additiv modell. Data fra alle fire sesongene ble brukt, noe som utgjorde 203,208 observasjoner. Flesteparten av de uavhengige variablene i modellen er basert på resultatene fra modellene for de analyserte pasningsevnene og nettverksanalysene. Resultatene indikerer at mer kompakte motivtyper, hvor færre unike spillere er involvert, har lavere sannsynlighet for å føre til et skudd. Det er ingen klar sammenheng mellom et lags interne bruk av effektive motivtyper og lagets tabellposisjon.

Acknowledgement

First, we want to thank our supervisors, Associate Professor Magnus Stålhane and Professor Lars Magnus Hvattum. Their guidance and support throughout the semester have been helpful and important for the progress of the thesis. We would also like to thank Rosenborg Ballklub and in particular their representative, the assistant coach Hugo Pereira, for providing us with the opportunity to work with an emerging field of study that suits our interests well.

Further, we want to express our gratitude to two former students who explored a similar subject, Haakon Haave and Håkon Høiland, for handing us their programming code. This has eased the pre-processing of data considerably, giving us the opportunity to have time for a more in-depth analysis. At last, we would like to thank Opta Sports, Kenneth Wilsgård from Norsk Toppfotball, Molde University College and the Norwegian University of Science and Technology for providing us with the data material needed.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Association football, referred to as football in this thesis, attracts more than 4 billion followers worldwide and is the most popular sport measured by participation, media coverage and key economic figures (Total Sportek, 2007). Enormous amounts of money are circulating in the world of football and the three teams generating the highest revenue in the 2016/2017 season earned a combined sum of €2 billion (Boor et al., 2018). Moreover, €1.72 billion of the total gross commercial revenue from the 2017/2018 UEFA Champions League, the 2017/2018 UEFA Europa League and the 2017 UEFA Super Cup are estimated to be distributed to the participating teams (UEFA, 2017). Considering these numbers, there is no doubt that success is valuable in football. Teams are therefore constantly seeking new ways to improve their performance, with the aim of winning titles to receive prize money and increased attention.

Recently, the use of statistics in sports has gained popularity. By utilising analytical methods to evaluate player and team performance, competitive advantages may be obtained. This emerging field of research is known as sport analytics. The aim of sport analytics is to support decision-making by helping players and coaches to make better-informed choices. Applications include tactical analyses, player recruitment, injury prevention and business decisions. Today, extensive amounts of data are available, and the opportunities to make more sophisticated and targeted analyses are increasing along with technological advances. Sport analytics is widely recognised in sports such as basketball, baseball and football.

The use of analytical methods in football will be further explored in this thesis. Three models are developed to evaluate players' passing abilities in the Norwegian top division for men, Eliteserien, and each model looks into a different aspect of a pass's success defined as accuracy, game overview and effectiveness. A pass is accurate if it successfully reaches its target, while a player being able to make tactically good passes that the pass recipient is able to follow up has a good game overview. Effective passes are more likely to lead to shots. The results from the passing ability models are further used in network analyses to identify the key

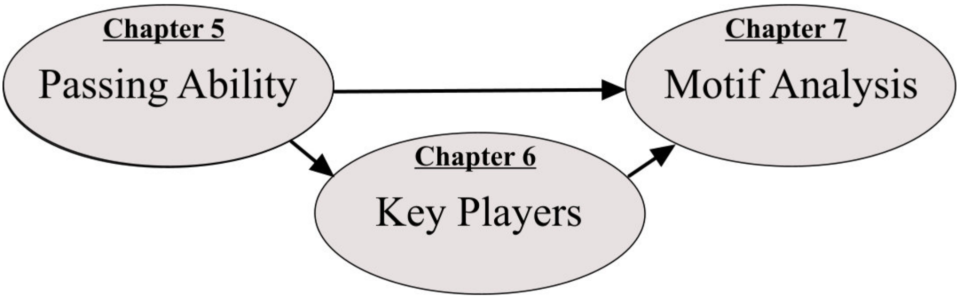


Figure 1.1: An overview of the analyses done in the thesis. The arrows show how the results in an analysis are connected to other analyses.

players in teams. At last, a motif analysis is performed by building a model to investigate the effectiveness of different types of passing motifs. Both the results from the passing ability models and the network analyses are included as variables in the model. Figure 1.1 shows how the results from the different analyses done in this thesis are connected.

1.1 Motivation and Research Questions

Rosenborg Ballklub (Rosenborg) has won 36 national titles and is historically Norway's most successful football team (rbk.no, 2018). With impressive winner statistics achieved in Norway, Rosenborg is looking to further develop in order to succeed both nationally and internationally. To achieve this, they want to make use of analytical methods to make better-informed decisions, both at the player level and the manager level.

The importance of passing in football is great. When successfully passing the ball between teammates, both in the attacking and defensive phases of a play, ball possession is kept with the purpose of creating goal-scoring opportunities and avoiding goals against. The analyses performed in this thesis aim to handle Rosenborg's request of utilising analytical tools for decision support as well as to be an addition to the academic field of sport analytics. By making use of past match data, Rosenborg will be provided with valuable information about their players' abilities to make accurate, tactically good and effective passes. Further, network analyses will reveal information about their players' role in the team and analyses of passing motifs will give an indication about which motif types that tend to be effective in terms of leading to shots. The approaches used build on and add to existing literature by examining similar problems and exploring both new and previously applied methods to deal with them.

For this thesis, three main research questions are defined and sub-questions are added for applications to Eliteserien. The questions are answered in three separate chapters, and they are as follows:

RQ 1: Which factors influence the success of a pass in Eliteserien?

- Who were the best passers in the 2017 season of Eliteserien?

RQ 2: How can the key players in a football team be identified?

- Who were the key Rosenborg players in the 2017 season of Eliteserien?
- How does Rosenborg depend upon certain key players when having differing ball possessions?

RQ 3: What determines the success of a passing motif in Eliteserien?

- Are the top performing teams in Eliteserien inclined to use the more effective motif types?

1.2 Report Structure

The remainder of the thesis is structured as follows. In Chapter 2, the basic theory on the statistical methods and the networks used are presented. Then, an introduction to sport analytics is given, followed by a review of research on passing both in football and in other sports in Chapter 3. The data used in the thesis is described in detail in Chapter 4 and the model set-up and the results of the passing ability models are presented in Chapter 5. Further, the set-up used for the network analyses performed to identify key players in teams is introduced in Chapter 6, including the outcome of the analyses. In Chapter 7, the set-up and the results of the regression model for the motif analysis are presented. At last, the concluding remarks and recommendations for further research are given in Chapter 8 and Chapter 9 respectively.

Chapter 2

Basic Theory

The basic theory relevant for the analyses performed in this thesis is presented in the following sections. The theory concerns the basics of logistic regression models, including their estimation, selection and validation methods, and network analysis.

2.1 Binary Logistic Regression

Binary logistic regression is a statistical technique used to model the relationship between dependent and independent variables when the dependent variable is binary, i.e. it takes the value of zero or one (Hosmer Jr et al., 2013). Consequently, the modelling approach is appropriate when dealing with problems where the outcome of observations can be classified as either a success or a failure. If i denotes the ith out of N observations, the dependent variable, y_i, is defined as:

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

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

$$\text{logit}(P_i) = \ln\left(\frac{P_i}{1 - P_i}\right) = \eta_i, \tag{2}$$

where P_i is the conditional mean and η_i is a function of the independent variables and their corresponding coefficient estimates. The conditional distribution is a Bernoulli distribution where an observation's probability of success is represented by the inverse logit function given by:

$$P_i = Pr(y_\text{i} = 1|\eta_i) = \frac{\exp^{\eta_i}}{1 + \exp^{\eta_i}} = \frac{1}{1 + \exp^{-\eta_i}}. \tag{3}$$

2.2 Generalised Linear Mixed Models

In a generalised linear mixed model (GLMM), both fixed and random effects are allowed to be included in a linear predictor. By incorporating random effects, one can build a model that accommodates correlation and one has a greater flexibility when making inferences about the data distribution (McCulloch et al., 2011). As of now, there is no consensus about the definitions of fixed and random effects, and several definitions are thus used by researchers. In Gelman et al. (2005), the terms *fixed* and *random* are referred to as *constant* and *varying*. Effects are constant if they are identical for all groups in a population, while they are varying if they are allowed to differ across the groups.

For a GLMM, η_i is a function of fixed and random effects given by:

$$\eta_i = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{\alpha}, \tag{4}$$

$$\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \Sigma_\sigma), \tag{5}$$

where $\boldsymbol{X_i}$ is a vector of fixed effects, $\boldsymbol{\beta}$ is a vector of fixed-effect coefficients, $\boldsymbol{Z_i}$ is a vector of random effects and $\boldsymbol{\alpha}$ is a vector of random-effect coefficients (Wood, 2006). The random effects are assumed to be normally distributed with Σ denoting the covariance matrix, parameterised by the coefficient vector σ.

2.3 Generalised Additive Mixed Models

A generalised additive mixed model (GAMM) is an extension to the previously mentioned GLMM for which the linear predictors are replaced by additive predictors, also known as smooth functions (Lin and Zhang, 1999). If some of the smooth functions are linear, they effectively become fixed effects. In the case of a GAMM, η_i can be written as:

$$\eta_i = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{\alpha} + f_1(x_{1i}) + \cdots + f_j(x_{ji}), \tag{6}$$

where $\boldsymbol{X_i}$ is a vector of fixed effects, $\boldsymbol{\beta}$ is a vector of fixed-effect coefficients, $\boldsymbol{Z_i}$ is a vector of random effects, $\boldsymbol{\alpha}$ is a vector of random-effect coefficients and f_1-f_j are smooth functions of variables x_{1i}-x_{ji} (Wood, 2006). The random effects are assumed to be normally distributed in accordance with equation (5). If the random-effect term in equation (6) is excluded, the model becomes a generalised additive model (GAM) (Hastie and Tibshirani, 1986).

Smooth functions are used to nonparametrically describe the dependency between the dependent variable and the independent variables (Hastie and Tibshirani, 1986); they allow the contribution of the independent variables to vary with a function. The function is unknown and is to be estimated during the regression through a scatterplot smoother such as a spline.

2.3.1 Smoothing Splines

A spline joins two or more polynomial curves, with knots representing the locations of the joins. Smoothing splines are useful for fitting smooth curves to a set of noisy observations through spline functions. The splines reduce a model's risk of becoming overfitted by penalising the coefficients of the estimated basis function. Overfitting refers to the problem that more terms than necessary are included in a model or that too sophisticated approaches are utilised in the model building approach, which can result in unreliable predictions and bad decisions (Hawkins, 2004). The trade-off between the degree of smoothness and the model's fit is given by a smoothing parameter, λ. The degree of penalisation of the coefficients and the degree of smoothness increase with λ (Wood, 2006). Illustrations of two smooth curves with different smoothing parameters are shown in Figure 2.1.

In the case of a full spline, a knot is placed at each data point. This is computational expensive, and penalised regression splines are often preferred for efficiency reasons due to fewer basis functions being used, with knots evenly distributed across the data. Thin plate regression splines, cubic regression splines and P-splines are examples of smoothing splines that estimate smooth functions of one variable. Cubic regression splines are often the preferred option when dealing with large data sets. For smooth functions of several variables, isotropic smooths or tensor product smooths are commonly used. In the case of isotropic smooths, the same degree of smoothness is assumed in all dimensions, while differences in smoothness across the variables are allowed for tensor product smooths. Overall, the choice of basis function is a question of acceptable mean squared error, desired computation time and the properties displayed by the particular data distribution (Wood, 2017).

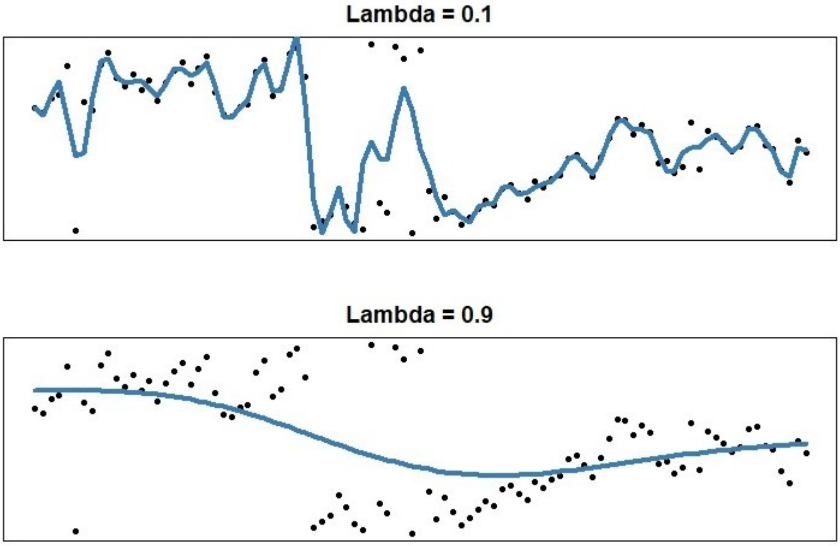


Figure 2.1: Illustrations of smooth curves with different smoothing parameters.

2.4 Estimation Methods for GLMMs and GAMMs

As random effects can be represented as penalised regression terms, which are similar to the representation of smooth functions, the same estimation methods can be applied to GLMMs and GAMMs as they effectively become GAMs (Wood, 2018). During estimation, such models are considered to be over-parameterised generalised linear models, which are estimated using penalised likelihood maximisation solved by penalised iteratively reweighted least squares (P-IRLS). When using the method of P-IRLS, weighted penalised least squares problems are solved and smoothing parameters are selected. Two basic algorithms exist for the selection of the appropriate smoothing parameter. The first method is single iteration, and it estimates the smoothing parameter in each iteration of the P-IRLS procedure, with no guarantee of convergence. In the second method, a selection criterion for the smoothing parameter based on the model deviance is defined and optimised. This algorithm is referred to as nested iteration, and convergence is guaranteed. For large data sets, the single iteration algorithm is considered to be more efficient (Wood et al., 2015).

To estimate the smoothing parameter, restricted maximum likelihood (REML), an idea first proposed by Patterson and Thompson (1971), can be used. The logic behind this approach is that maximum likelihood estimation is performed using a modified likelihood function calculated from a transformed set of data, allowing variance components to be estimated independently of the fixed-effect estimates. This is achieved by partitioning the likelihood function proposed by Patterson and Thompson (1971) into two parts where one of the parts excludes the fixed effects (Corbeil and Searle, 1976). REML tends to be the preferred option among researchers due to the method being less prone to local minima and undersmoothing compared to other available estimation methods such as generalised cross validation or an un-biased risk estimator (Wood, 2011).

2.5 Model Selection

In order to obtain the best fitted model, the set of independent variables needs to be carefully selected within the constraints of the available data. The aim of the selection process is to build a statistically stable model that accurately reflects the true outcome of the data by minimising the model's dependency on the observed data and by reducing its estimated standard errors. Also, by removing variables not favourable for the model fit, the problem of overfitting can be dealt with (Hosmer Jr et al., 2013). The Akaike Information Criterion (AIC) and the Wald statistic are examples of statistical measures that can be used for model selection purposes. Moreover, the AIC score is useful when comparing models with differing numbers of explanatory variables.

2.5.1 AIC

Following Hosmer Jr et al. (2013), the AIC value of a model is given by:

$$AIC = -2 \times L + 2 \times (p + 1), \tag{7}$$

where L is the log-likelihood of the fitted model and p represents the number of non-constant explanatory variables to be estimated. The model achieving the lowest AIC score has the best fit.

2.5.2 Wald Statistic

The Wald statistic of a fixed-effect variable is defined as:

$$W = \frac{\hat{\beta}}{\widehat{SE}(\hat{\beta})}, \tag{8}$$

where $\hat{\beta}$ is the estimated variable coefficient and $\widehat{SE}(\hat{\beta})$ represents the estimate of the standard error (Hosmer Jr et al., 2013). The best fitted model consists of those variables with Wald statistics that prove to be statistically significant at a predetermined significance level.

2.5.3 Stepwise Regression

The AIC and the Wald statistic can be used as model fit criteria in a stepwise regression to iteratively eliminate or add variables from or to an initially developed model. Stepwise regression can for instance be performed through forward selection, backward elimination or stepwise selection (Statistics Solutions, 2018). Backward elimination is used on a fully developed model. In each step, the variable that, when removed from the model, improves the model fit the most is excluded. The procedure is continued until no more variables can be removed in order to give a better model fit. When using the AIC as criterion, the variable that lowers the AIC the most when removed from the model is excluded, while in the case of the Wald statistic, the variable with the highest Wald statistic above a chosen significance level is removed.

Although being very effective in terms of choosing variables, stepwise regression is faced with some challenges. When adding or removing a single variable at a time, a suboptimal model may be the outcome as not all combinations of variables in a model are considered. Also, removing combinations of variables could be a better solution compared to considering only one variable in each step. As there always is a predetermined set of variables to choose from, some variables must be chosen even if they are insignificant, meaning that insignificant variables could be present in the final model (Goodenough et al., 2012). In addition to the mentioned challenges, Judd et al. (2011) state that stepwise regression can result in a suboptimal model when highly correlated explanatory variables are incorporated in the model.

2.6 Validation Methods for Binary Classifier Systems

To test how well a model describes the data, it is crucial to evaluate the model's goodness of fit, which is defined as the accuracy of the true outcome in the data based on the probabilities produced by the model (Hosmer Jr et al., 2013). Validation methods used to test the performance of a binary classifier, such as in the case of a binary logistic regression model, include the area under a receiver operating characteristic (ROC) curve, the area under a precision-recall (PR) curve and the Hosmer-Lemeshow (HL) test.

2.6.1 Area Under an ROC Curve

To understand the theory of ROC curves, the terms sensitivity and specificity are introduced. A model's sensitivity is given by the number of true positives over the total number of positive outcomes in the data, and the measure is commonly referred to as a model's true positive rate. On the other hand, specificity is a model's true negative rate. The formal definitions of sensitivity and specificity are:

$$Sensitivity = \frac{TP}{TP+FN}, \tag{9}$$

$$Specificity = \frac{TN}{TN+FP}, \tag{10}$$

where TP, FN, TN and FP are the numbers of true positives, false negatives, true negatives and false positives respectively. True or false depicts whether an observation is truly or wrongly predicted to be successful or unsuccessful (Fawcett, 2006).

A threshold must be set in order to classify the predicted outcome of a model. Predictions above this threshold are positive, while those below are classified as negative. In an ROC curve, the entire range of thresholds, limited between zero and one, is considered. For each threshold value, the sensitivity of a model is plotted against one minus the specificity. The latter is often referred to as a model's false positive rate. In Figure 2.2, illustrations of two ROC curves with different areas

Table 2.1: Guidelines for assessing a model's discrimination ability using the area under an ROC curve.

Value	*Description*
0.5	No discrimination
0.5 < AUC < 0.7	Poor discrimination
0.7 ≤ AUC < 0.8	Acceptable discrimination
0.8 ≤ AUC < 0.9	Excellent discrimination
AUC ≥ 0.9	Outstanding discrimination

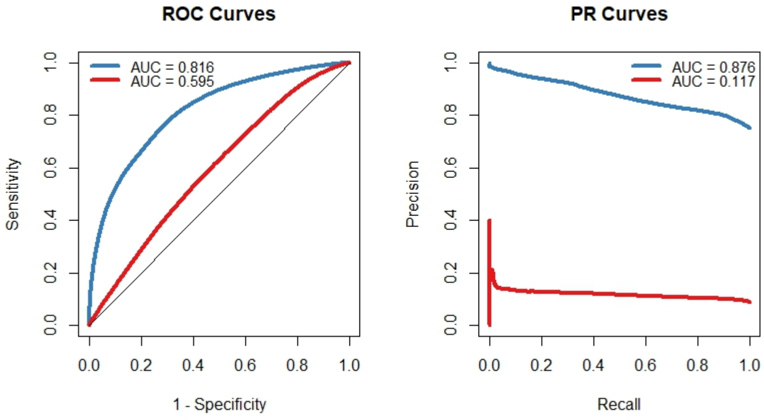


Figure 2.2: Illustrations of ROC and PR curves with their respective AUCs.

under the curves (AUCs) are shown. At the leftmost point on the graph, where the specificity is one and the sensitivity is zero, all predictions are negative, while all observations are predicted as positives at the rightmost point on the graph where the threshold is zero.

The AUC for an ROC curve is a measure of a model's ability to discriminate between positives and negatives. If being on the diagonal, the model functions as a random classifier and has an AUC of 0.5. The aim is to obtain a model with a perfect classification ability, which corresponds to an AUC of 1.0. When evaluating a model's discrimination ability based on the AUCs of ROC curves, Hosmer Jr et al. (2013) suggested to use the guidelines presented in Table 2.1. As seen from the given AUCs in Figure 2.2, the associated models have an excellent and a poor discrimination ability respectively.

2.6.2 Area Under a PR Curve

Another validation method used to assess the fit of a model is the area under a PR curve. For this curve, the precision, which is a model's positive predictive value, is plotted against the recall, which is identical to the sensitivity of an ROC curve, for the same range of thresholds as for the ROC curve. With the same notation as in equations (9) and (10), the recall and precision are defined as (Fawcett, 2006):

$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

$$Precision = \frac{TP}{TP + FP}. \tag{12}$$

The fit of a model increases with the size of the area under the PR curve. A perfect model consists of only true positives and true negatives, and it is obtained when the AUC is 1.0. Two PR curves are illustrated in Figure 2.2, and as seen from the figure, a higher AUC is associated with a concave-shaped curve, which is achieved when the recall increases faster than what the precision decreases.

Davis and Goadrich (2006) studied the relationship between ROC and PR curves, and they proved that PR curves better reflect the true performance of a model when the data distribution is highly skewed. For instance, if there is an excess of true negatives in the data set, the precision parameter is able to capture the skewness by comparing the false positives to the true positives rather than to the true negatives.

2.6.3 Hosmer-Lemeshow Test

The HL test is a validation tool which appropriately tests a model's calibration ability, i.e. its ability to generate true predictions. In this test, the data is sorted in ascending order by the predicted probabilities and then divided into g equally sized groups. Within each group, the observed and expected frequencies of positives and negatives are obtained and compared to test whether poor predictions are included in the estimated model. The test statistic, $\widehat{C}$, is given by:

$$\widehat{C} = \sum_{k=1}^{g} \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right], \tag{13}$$

where o_{1k}, o_{0k}, $\hat{e}_{1k}$ and $\hat{e}_{0k}$ are the observed true positives, the observed true negatives, the predicted true positives and the predicted true negatives for group k respectively. The test statistic is chi-squared distributed with $g - 2$ degrees of freedom. The null hypothesis of a good model fit is tested against the alternative hypothesis of a poor model fit. Hence, low values of the test statistic and high p-values indicate a good calibration ability (Hosmer Jr et al., 2013).

When the sample size gets big, the results of HL tests are proven to be sensitive to the chosen number of groups (Allison, 2014). This is due to the fact that the power of the HL test increases with the sample size, which can make even small differences between the observed and expected frequencies statistically large. Moreover, there is a lack of consensus of how to choose the optimal g, although the standard choice has commonly been ten.

Paul et al. (2013) do not recommend to use the HL test for sample sizes larger than 25,000, but have proposed a strategy for such cases. The idea is to choose random samples of equal sizes from the full data set, and for each sample, perform the HL test with $g = 10$. A similar idea was also suggested by Kramer and Zimmerman (2007). While Paul et al. (2013) and Kramer and Zimmerman (2007) proposed the use of 1000 and 5000 observations in each sample respectively, they did not suggest how to choose the number of samples to test. Bartley (2014) elaborated on the previous work, and proposed the use of 100 random samples with 5000 observations in each sample. Further, it is suggested that if more than ten of the HL tests result in rejections of the null hypothesis at a significance level of 5%, the validity of a model should be carefully examined.

2.7 *k*-fold Cross-Validation

Cross-validation is a statistical technique used to assess the predictive power of a model. The key idea is to divide the data into two subsets where one subset consists of *known* data used to train the model whereas the other subset includes *unknown* data used to validate the model. By testing how a model generalises to an independent data set, the problem of overfitting might be reduced.

k-fold cross-validation represents one type of cross-validation where the data initially is divided into k randomly chosen and equally sized subsets, referred to as folds. Further, k iterations of the training and validation folds are performed. For each iteration, one of the k folds is used as the validation set, while the remaining $k-1$ folds constitute the training set. Although the number of folds is not a fixed parameter, ten folds are commonly used in statistical analyses (Refaeilzadeh et al., 2009).

2.8 Brier Score

The Brier score is a measure used to evaluate probability forecasts. Originally introduced by Brier (1950), the score for binary events in its most common formulation is given by:

$$BS = \frac{1}{N}\sum_{k=1}^{N}(y_k - o_k)^2, \tag{14}$$

where N is the number of observations in the forecast data, y_k is the probability that event k occurs and o_k equals one if event i occurs and zero otherwise. Thus, the Brier score is basically the mean squared error of the probability forecasts, and the score takes a value within the range of zero and one. A Brier score of zero means perfect forecasting, while high scores indicate poor forecasting (Wilks, 2011).

2.9 Network Theory

Network theory is a part of graph theory, and it has several applications, including social networks, computer science, biology and medicine. Social network analysis is a useful method for examining the relationships and patterns among social entities; it allows researchers to investigate both social structures and individual attributes simultaneously. Different measures of centrality can be computed to identify the importance of an entity, while clustering coefficients can be calculated to study the extent to which the entities in a network cluster together (Boccaletti et al., 2006).

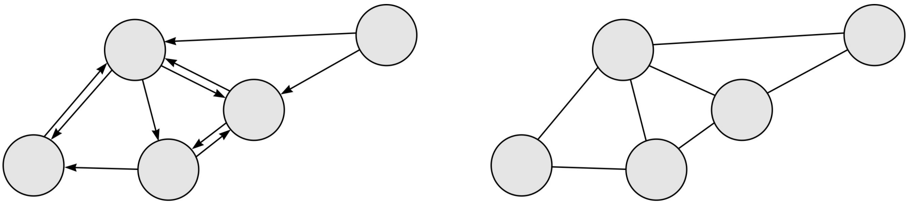
2.9.1 Network Structure

A network consists of nodes and edges. Following the notation in Boccaletti et al. (2006), an undirected (directed) and unweighted graph $G = (\mathcal{N}, \mathcal{L})$ consists of two sets, $\mathcal{N}$ and $\mathcal{L}$, such that $\mathcal{N} \neq \emptyset$ and $\mathcal{L}$ is a set of unordered (ordered) pairs of elements in $\mathcal{N}$. The nodes of G are the elements of $\mathcal{N} \equiv \{n_1, n_2, ..., n_N\}$, while its edges are the elements of $\mathcal{L} \equiv \{l_1, l_2, ..., l_K\}$, where the number of elements in $\mathcal{N}$ and $\mathcal{L}$ are given by N and K respectively. Two nodes are adjacent if they are connected together by an edge. An edge is defined by two nodes, i and j, and is denoted as l_{ij}. For a directed graph, direction matters, and the two directed edges l_{ij} and l_{ji} are not necessarily equal as opposed to for an undirected graph. Illustrations of a directed and an undirected graph are given in Figure 2.3.

An adjacency matrix, A, can be used to represent a graph. For a directed network, it is an $N \times N$ matrix with elements a_{ij} $(i, j = 1, ..., N)$. If there is a directed edge from node i to node j, a_{ij} equals one, otherwise the value is zero. Further, the directed edges can be weighted according to the strength of connection between the nodes. A weighted graph G^W adds to an unweighted graph by including an additional set $\mathcal{W} \equiv \{w_1, w_2, ..., w_K\}$, where $\mathcal{W}$ is a set of edge weights. Sets $\mathcal{L}$ and $\mathcal{W}$ have an equal size of K. A weighted graph can be described by a weights matrix, W. The elements of an $N \times N$ weights matrix are given by w_{ij}, where w_{ij} represents the strength of the connection between nodes i and j. If two nodes are not directly connected, w_{ij} equals zero. The adjacency matrix for a weighted graph consists of entries such that $a_{ij} = 1$ if $w_{ij} \neq 0$ and $a_{ij} = 0$ if $w_{ij} = 0$.

A sequence of distinct adjacent nodes is defined as a path, and the shortest path refers to the path of minimal distance between two nodes. The length of the shortest path differs in the case of an unweighted and a weighted network. In an unweighted network, the shortest distance is equivalent to the lowest number of directed edges needed to be traversed. For a weighted network however, the distance between two nodes is the summed weights on the chosen path. Hence, the shortest path in this case is the path that has the lowest sum of edge weights, and it might turn out to not be the one with the fewest directed edges to traverse.

To interpret the edge weight between two nodes in a weighted graph as a strength of a connection instead of a cost, it is proposed by Opsahl et al. (2010)



Panel A: Directed graph

Panel B: Undirected graph

Figure 2.3: An illustration of a directed graph in Panel A and an undirected graph in Panel B.

to inverse the weights before applying the shortest path algorithm. Then, when following Opsahl et al. (2010), the length of the shortest path between nodes i and j can be found by solving:

$$d_{ij}^{w\alpha} = \min\left(\frac{1}{(w_{ih})^{\alpha}} + \cdots + \frac{1}{(w_{hj})^{\alpha}}\right), \tag{15}$$

where α is a positive tuning factor, giving the equivalent to an unweighted network if it has the value of zero and giving the outcome as from the use of Dijkstra's algorithm if the value is one. The Dijkstra's algorithm, which is thoroughly explained in Dijkstra (1959), is commonly used to compute the shortest path between nodes in a directed and weighted network.

A directed graph is said to be strongly connected if there is a directed path from node i to node j for every pair of distinct nodes in the network. If there is only an undirected path between the nodes however, the graph is weakly connected (Weisstein, 2018). Otherwise it is unconnected.

2.9.2 Closeness Centrality

The closeness centrality of a node depends on the length of the paths from the node to all other nodes in the network. It provides a measure of a node's independence from other nodes, with higher scores being associated with greater independence (Freeman, 1978). For a weighted network where the weights are considered to be strengths, and both the directed edges going in and out of a node are considered, the closeness centrality of node i is given by (Pena and Touchette, 2012):

$$C_{\mathrm{C}}(i) = \frac{2N_1}{\sum_{j\neq i} d_{ij}^{w\alpha} + \sum_{j\neq i} d_{ji}^{w\alpha}}. \tag{16}$$

.

$$N_1 = N - 1, \tag{17}$$

where $d_{ij}^{w\alpha}$ represents the shortest path between nodes i and j given by equation (15) and N is the total number of nodes in the graph. N_1 is a normalisation factor letting the closeness measure being comparable across networks of different sizes (Freeman, 1978). Additionally, it lets the measure being interpreted as the inverse of a node's average distance of the shortest paths to the other nodes in the graph.

2.9.3 Betweenness Centrality

The betweenness centrality of a node is based on the number of shortest paths between two other nodes passing through the node. The idea behind the measure is that a node in a network is central if it is placed on the shortest path between two connecting nodes. Also, it can viewed as a measure of a node's potential to control the flow of information in a graph (Freeman, 1977). The betweenness centrality of node i can be defined as:

$$C_{\mathrm{B}}(i) = N_2 \times \sum_{\substack{j,k \in \mathcal{N} \\ j \neq k}} \frac{g_{jk}(i)}{g_{jk}}, \tag{18}$$

$$N_2 = \frac{2}{N^2 - 3N + 2}, \tag{19}$$

where g_{jk} is the number of shortest paths from node j to node k, while $g_{jk}(i)$ is the number of shortest paths between nodes j and k passing through node i (Boccaletti et al., 2006). N_2 is a normalisation factor that enables comparisons between networks with differing number of nodes, N (Freeman, 1977). The shortest paths can be calculated by a method proposed by Brandes (2001), which is modification of Dijkstra's algorithm. Brandes (2001) proves that his method is more efficient and less computational burdensome.

2.9.4 PageRank Centrality

The PageRank algorithm was introduced by Brin and Page (1998) to provide a method of measuring the importance of web pages. The intuition behind the algorithm is that a web page achieves a high PageRank either if many web pages are pointing to it or if some of the web pages pointing to it have a high PageRank themselves. Hence, all web pages' PageRank scores must be calculated simultaneously as the scores depend on each other.

For a network, the PageRank of node i is given by:

$$PR(i) = \frac{1-d}{N} + d \times \sum_{j \in \mathcal{M}(i)} \frac{PR(j)}{C(j)}, \tag{20}$$

where d is a damping factor representing the probability that a node will join together with other nodes, N is the total number of nodes in the network, $PR(j)$ is the PageRank of node j, $C(j)$ is the number of directed edges departing from node j and $\mathcal{M}(i)$ is a set of nodes that are connected to node i (Fu et al., 2006; Page et al., 1999). This definition deviates from the original definition proposed by Brin and Page (1998) as the first term on the right-hand side in the original equation is divided by the total number of nodes. By doing so, the PageRank scores in a network sum to one.

For a weighted network, the PageRank centrality can be estimated as:

$$PR^w(i) = \frac{1-d}{N} + d \times \sum_{j \in \mathcal{M}(i)} \frac{w_{ji}}{L_j} PR^w(j) \tag{21}$$

where w_{ji} are elements of the weights matrix, $L_j = \sum_k w_{jk}$ is the sum of the weights on the edges with direction out from node j and the other parameters are as given in equation (20) (Pena and Touchette, 2012). Higher weights on the incoming edges to a node correspond to a higher PageRank for that node.

2.9.5 Clustering

Clustering coefficients are computed to get a quantification of nodes' tendency to cluster together, and they account for the transitivity of a graph, that is, the probability that adjacent nodes of a node are connected. A high clustering coefficient of node i means that if node i is connected to node j, and node j is connected to node h, then the probability of node i also being connected to node h is high (Barrat et al., 2007). Following the method of Barrat et al. (2007), the weighted local clustering coefficient of node i in an undirected graph is given by:

$$c_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}, \tag{22}$$

where s_i is the strength of node i, that is, the sum of the node's edge weights of adjacent nodes, k_i is the node degree, which is the number of undirected edges incident to the node, w_{ij} and w_{ih} are weights and a_{ij}, a_{ih} and a_{jh} are elements of the adjacency matrix. $s_i(k_i - 1)$ is a normalisation factor and ensures that $0 \leq c_i^w \leq 1$. The equation is undefined for nodes with only one connecting node.

2.9.6 Network Motifs

The concept of network motifs was introduced by Milo et al. (2002), and it provides a method for discovering patterns of interconnections. The idea is to study the functional properties of subgraphs consisting of k nodes in a network. A subgraph $G' = (\mathcal{N}', \mathcal{L}')$ of $G = (\mathcal{N}, \mathcal{L})$ is a graph where $\mathcal{N}' \subseteq \mathcal{N}$ and $\mathcal{L}' \subseteq \mathcal{L}$ (Boccaletti et al., 2006). In Figure 2.4, all possible combinations of motifs with $k = 4$ are shown. Hence, with four nodes, there are five different types of motifs: *ABAB*, *ABCA*, *ABAC*, *ABCB* and *ABCD*. Duplicate nodes within a subgraph imply that a single node is involved several times in the motif.

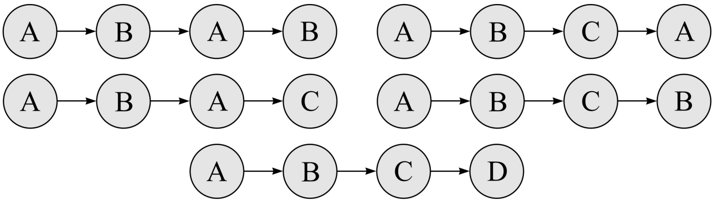


Figure 2.4: Motifs of size four.

Chapter 3

Literature Review

In this chapter, an introduction to sport analytics, with emphasis on its usage in football, is given. Further, a review of existing literature related to passing, both in the context of football and other sports, is presented. A separate section is devoted to explain how this thesis compares to other studies in terms of the modelling approaches used.

3.1 Introduction to Sport Analytics

Decisions in sport have traditionally been made qualitatively by humans, being based on gut feelings or adherence to previous choices (Steinberg, 2015). Sport analytics offers new ways of assessing the skill of players and teams. By making use of data material 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 increasing performance. Additionally, sport analytics has become a valuable tool used by scouts to identify transfer targets (The Guardian, 2017).

The idea of sport analytics was introduced by Mottley (1954), while Machol and Ladany (1976), Maher (1978) and Coleman (2012) boosted the attention with their thorough reviews of studies applying analytical methods in sport. Today, every major professional sports team has people dedicated to apply statistical methods to help players and managers making better pre-game and match decisions (Steinberg, 2015). From the introduction of sport analytics, the complexity of the data material used has increased and the latest technologies, including big data, machine learning and artificial intelligence, have opened up for more sophisticated analyses (STATS LLC, 2017).

In football, the use of analytical methods has evolved since it was first used by Charles Reep in the 1950s (Sammonds, 2017). Several aspects of the sport are of researchers' interest, including the choice of playing style, prediction of goal-scoring chances and determination of players' market value. The popularity of the latter field of study has grown recently in line with the growing economic impact in football (Frick, 2007). Players' market value can act as an estimate of transfer fees, and Sæbø and Hvattum (2015), He et al. (2015) and Müller et al. (2017), among others, have developed valuation models for football players. Considering the style of play, several studies have been looking into whether direct play or possession play is more effective in terms of creating goal-scoring opportunities (Reep and Benjamin, 1968; Bate, 1988; Hughes and Franks, 2005). The expected goals theory, which measures a player's probability of scoring a goal given some input parameters, is a popular metric among researchers today. Different methods are used to describe this metric, including those introduced by Sáez Castillo et al. (2013), McHale and Szczepański (2014) and Deb and Dey (2017).

3.2 Performance Related to Passing in Football

Evaluation of performance is crucial to improve decision-making across all levels in a sports organisation. For instance, in-depth analyses of players' technical, tactical and physical capabilities provide coaches and players with additional information that can be used to make adjustments to training sessions by identifying areas of improvements. Typically, performance assessment of players is done by giving players a score depending upon their performance through a rating system.

Several metrics and methods used to evaluate performance in football, both skill-specific and overall, have been proposed by researchers. These include the EA Sports Players Performance Index (McHale et al., 2012), the QPass metric (Gyarmati and Stanojevic, 2016), the plus-minus rating (Sæbø and Hvattum, 2015; Kharrat et al., 2017), Markov modelling (Szczepanski, 2015; Haave and Høiland, 2017), regression analysis (Szczepański and McHale, 2016; Rein et al., 2017; McHale and Relton, 2018), machine learning techniques (Brooks et al., 2016; Bransen, 2017) and network analysis (Pena and Touchette, 2012; Pina et al., 2017; Bekkers and Dabadghao, 2017).

Passing is the most frequent event happening during a football match, and its importance for achieving success is great. Success in football is obtained by scoring more goals than the opponent team, and by making effective passes, more goal-scoring opportunities arise. Lately, studies related to passing have become more popular, both at the player and team level. In the following, existing literature on passing behaviour is presented, followed by a review of studies performing network and motif analyses.

3.2.1 Evaluating Passing Behaviour

At the player level, Szczepański and McHale (2016) evaluated football players' passing ability by building a GAMM to estimate the probability of a pass's success. A set of independent variables reflecting the environment in which a pass is made is included to account for the overall difficulty of the pass. Random effects of players, the players' team and the opponent teams are included in the model to investigate individual passing skills and teams' abilities to facilitate and hamper passes. The model was developed using event-data from one season of the English Premier League, and predictions were made for the consecutive season. When evaluating players' passing ability, predictions for the average difficult pass were used to ensure fair comparisons between the players by filtering out the pass difficulty. The resulting model recognises both the difficulty and frequency of passes made by a player. Hence, players having a higher completion rate, but fewer passes attempted, may be ranked below players with a slightly lower completion rate and more completed passes. Tovar et al. (2017) developed a similar GAMM and demonstrated the model's ability to predict the performance of any player transferring from the Colombian to the Spanish league.

Whereas Szczepański and McHale (2016) considered only the difficulty of a pass, Power et al. (2017) introduced two measures to cover both the difficulty and effectiveness of passes. An effective pass has a greater chance of resulting in a goal-scoring opportunity. By using a supervised learning approach, Power et al. (2017) estimated the *risk* and *reward* of passes through the use of event-data and player tracking data from two seasons of the English Premier League. The risk of a pass is defined as the probability that a player executes the pass taking its difficulty into account, while the reward of a pass is defined as the probability that the pass made ends with a shot within the next ten seconds, i.e. its effectiveness. Different metrics were proposed by the authors to identify the players who make riskier passes than the average player, the best players to make and receive difficult passes and the best players to make and receive dangerous passes. Difficult and dangerous passes are passes that fall into the 75th percentile of the riskiest and highest rewarded passes respectively. Results indicate that players get rewarded for making critical passes in a ball possession, that is, passes that unlock a defence.

Pass effectiveness was also explored by Brooks et al. (2016) who used machine learning techniques to measure the importance of passes by examining the relationship between pass location in a possession and shot opportunities. The field was divided into 18 zones, and the distances from the origin and destination of a pass to the centre of all zones were measured. These distances and pairs of origin-destination distances represent the pass location. Each pass in the data set was given a value according to its probability of leading to a shot, and players were ranked based on the value of their passes.

Other studies related to pass effectiveness include Rein et al. (2017) and Gyarmati and Stanojevic (2016). Rein et al. (2017) used a quite different notion of pass effectiveness, and they introduced two metrics for it represented by the change in a team's space control and the change in the number of defenders between the ball carrier and the opponent's goal from the initiation to the completion of a pass. Voronoi-diagrams were used to assess the space controlled by each team. The two metrics were used as fixed effects in three mixed models to test their influence on the number of shots, the number of goals and the outcome of a match. Results show that both metrics positively affect the number of goals scored and the probability of winning a match.

Gyarmati and Stanojevic (2016) used event-data from the 2015/2016 season of the Spanish La Liga to evaluate football players' contributions to creating goal-scoring opportunities by ranking them according to the intrinsic values of their passes using *QPass*, a metric introduced by the authors. The idea behind the metric is to derive the value of having or not having the ball at a specific point on the field, where the field is partitioned according to a team's style of play. Further, the *merit* of a pass, which is defined as the change in its field values, is evaluated by the QPass metric. Surprisingly, the results indicate that unsuccessful passes sometimes can benefit the team through increased field values.

Mackay (2017) did not precisely consider pass effectiveness in terms of creating goal-scoring chances, but rather looked at the probability of different actions on the field resulting in a goal. This was done by developing a model using ridge logistic regression. Sliding windows and recurrent sliding windows were used to test whether the inclusion of variables concerning previous actions in a team's ball possession would increase the performance of the model. The dependent variable takes the value of one if an event is part of a ball possession ending with a goal, and five explanatory variables entered the model: a categorical variable telling the nature of the event, an indicator variable telling whether the current action ended on the head of a teammate, the speed of action, the speed of the direct distance covered and the predicted goal probability by location. The latter variable was obtained by fitting a GAM.

The model developed by Mackay (2017) was built using event-data from five seasons of the English Premier League, and the predicted goal probability by location in the current event turned out to be the most influential variable. In general, more recent actions tend to have a greater influence than their previous events. Individual player performance was determined by looking at the differences between the model's generated probability of success at the moment a player received the ball and the moment after the player no longer had possession of the ball. The differences for each of the player's ball involvements were summed, meaning that players being able to move the ball into situations where goals are more likely to be the outcome achieve higher scores. The top 15 players in the 2016/2017 season were identified, all of whom played in offensive positions.

3.2.2 Network Analysis

Several studies have analysed sequences of passes, playing styles and individual contributions to teams in football. For such purposes, teams are commonly viewed as networks where nodes represent players with edges connecting them. The edges can be weighted according to a chosen criterion, and the number of successful passes between two players has commonly been used. This modelling approach can be categorised as a type of social network analysis.

Duch et al. (2010) used social network analysis to determine individual player contributions to teams, while Pena and Touchette (2012) used it to examine team strategies. In the latter paper, the playing style of a team was observed by fixing each node according to the team's tactical strategy. Further, the team's game-play robustness, characterised by the lowest number of intercepted passes required to disturb its natural flow and to isolate a subgroup of its players, was evaluated. Additionally, different centrality measures were calculated to study players' individual contribution to their team in terms of importance and connectedness.

Centrality measures were also explored by Arriaza-Ardiles et al. (2018). In this paper, players' closeness and betweenness centrality scores were computed in order to study players' capability to connect with teammates and their ability to make contributions in the play between other players. Further, clustering coefficients were calculated to measure relations between players and to identify the importance of the players who most frequently interacted with each other when ball possession was kept. Another centrality measure, the PageRank centrality, was investigated by Rojas-Mora et al. (2017) to find the most important players of a team in matches from the group stage of the Copa America 2015. A combined network for both teams in a match was used, meaning that players on different teams were connected by edges representing miskicks and interceptions.

Gama et al. (2014) used a network-based approach to identify the key players in the attacking phases of play, while Pina et al. (2017) investigated the relationship between specific network metrics and teams' outcome of offensive plays, after controlling for the effect of total passes. In the latter paper, a hierarchical logistic regression model was developed using data from the group stage of the UEFA Champions League 2015/2016, and the extent to which network density, clustering coefficients and centralisation could predict a team's performance on offensive plays were examined. Results suggest a negative relationship between the success of offensive plays and network density, which is the only significant predictor variable in the model. This result was supported by Peixoto et al. (2017) who applied social network analysis to study differences across centrality measures in successful and unsuccessful offensive plays.

McHale and Relton (2018) elaborated on the work done by Szczepański and McHale (2016) on assessing players' passing ability and developed a GAMM to estimate the likelihood of a pass's success with the purpose of using the results from the model to identify the key passers in a team through a network analysis. A major difference from the model developed by Szczepański and McHale (2016) is that player tracking data was utilised, enabling better proxies for the fixed-effect variables. The edges in the network were weighted according to the difficulty of the passes between players, given by the estimated likelihoods from the fitted GAMM. By taking pass difficulty into account when weighting the passes, the players' involvements in the passing of the ball can more fairly be compared as it is not only the number of completed passes that is considered. The key players in a team were determined by calculating the exponential centrality measure of each player. Of the five teams for which the three most important players were presented, only one player was a defender.

3.2.3 Network Motif Analysis

Recently, the study of network motifs, which was introduced by Milo et al. (2002), has aroused interest among researchers who apply network analysis to football. Gyarmati et al. (2014) analysed flow motifs consisting of three consecutive passes in order 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 motifs in the passing networks to the expected occurrence in randomly generated networks with identical properties. Further, cluster analyses were performed to examine similarities and differences in teams' passing patterns. Elaborating on this previous work, 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 (2017) 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 goal-scoring 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 in order to find players with similar characteristics.

3.3 Passing in Other Sports

Sport analytics has been and is still in widespread use in several team sports other than football. However, the focus is rarely on passing or other specific skills of a player as opposed to modelling the game outcome or determining the total rating or contribution of players. Related to passing, teams' possession of the ball, or in general the object to be played with, is mostly considered in the past.

3.3.1 Regression Models

Vicente-Vila and Lago-Peñas (2016) performed a performance analysis looking into the effectiveness of goalkeepers and other predictors on ball possessions in futsal. In futsal, the goalkeeper can play as an outfield player while attacking, which was incorporated in a binary logistic regression model as a dummy variable. The main findings from the study are that possessions that result in a goal are more likely to occur when the goalkeeper plays as an outfield player, when the sequence lasts less than ten seconds and when the pressure on players is lower. Gómez et al. (2015) explored a similar model within futsal to determine which characteristics of a possession that are more or less effective, while Gómez et al. (2013) and Conte et al. (2017) used similar approaches in basketball on possessions and fast break actions respectively.

The inside pass used in basketball is investigated by Courel et al. (2013). Descriptive statistics were provided to induce whether inside passes have an impact on teams' offensive success, and a multinomial regression model was developed to test the significance of chosen predictors on the success of an inside pass. Offensive possessions involving the inside pass were found to be more effective, and the positioning of the passer and the action immediately chosen by the recipient were found to significantly affect the likelihood of a successfully made inside pass. Further exploration of the inside pass has been conducted by Courel Ibáñez (2017) with the aims of deciding how inside passes affect possession effectiveness and finding predictors for the success of the inside pass. Additionally, the influences of individually made actions and combined interactions from passers and receivers on performance in the case of inside passes in both offensive and defensive plays were tested. Conforming with the former research, the inside pass is found to increase possession effectiveness. Other sports in which player or team performance has been investigated through regression models include baseball (Piette et al., 2010), ice hockey (Macdonald et al., 2013) and rugby (Higham et al., 2014).

3.3.2 Network Analysis

Although network analysis within sport seems to be more commonly used in the context of football, it has been applied to other sports as well. In basketball, Fewell et al. (2012) examined the degree centrality, clustering, entropy and flow centrality for teams playing in the first NBA play-off round through both the teams' individual and combined weighted transition graphs. In the graphs, nodes represent fixed player positions, start-of-play actions and possession outcomes, while directed edges represent ball movements between positions or indicate which positioned player who is responsible for the start or outcome of a possession. Only offensive plays that involved at least three of the players in the starting lineup were considered. The aim was to identify differences in offensive strategies. The findings from the overall network suggested, as expected, that the point guard is essential in a leadership role, with the teams having a centralised ball distribution, while for some individual teams, this tend to vary when comparing the clustering and the degree centrality.

Piette et al. (2011) started out with a bipartite graph with the two sets of nodes consisting of basketball players and units of five players. The units represent the players playing together on the ground for a team when a certain event takes place. Edges were only allowed between a player and a player unit, and they were weighted based on the efficiency of the unit. A transformation to a unimodal graph was made such that the weights would only be between players, now indicating the efficiency of the units for which the players had played in together. To find the most central players, the eigenvector centrality with random restart was calculated. Additionally, the p-value of the centrality was found by bootstrapping the initial centrality measure in order to avoid that the centrality score of a player became inflated due to many adjacent nodes, i.e. making it possible to tell whether a player's performance was subject to chance. Using three different data sets to create three graphs looking at defensive efficiency, offensive efficiency and total efficiency, the players who are under- or over-performing were identified, and the players' importance in relation to their team units was established. Network analyses have also been performed in other sports including volleyball (Clemente et al., 2015; Kang et al., 2015) and cricket (Dey et al., 2017).

3.4 The Thesis's Supplementation to Existing Literature

In this thesis, three chapters are devoted to separately answer the three main research questions defined in Chapter 1. Firstly, the passing abilities of football players in the Norwegian top division Eliteserien are evaluated in Chapter 5 through the development of three GAMMs. The first model considers the traditional approach taken to passing, which is pass accuracy, and it is similar to the model suggested by Szczepański and McHale (2016). However, two more sources of data are used, several new binary variables are proposed and some new smooth terms are added. The second model is seemingly unique in the way the dependent variable is defined, and it handles the tactical aspect of a pass by examining the probability that a pass can be followed up, i.e. whether the play easily can be continued. The third model investigates pass effectiveness, which has been considered in many different ways by researchers in the past. Many of the definitions used for pass effectiveness are however restricted such that offensive players tend to be favoured. By considering all passes in a sequence that leads to a shot as effective in this thesis, the thought is that a more fair assessment of players can be made.

Secondly, the results from the passing ability models are used in network analyses similar to the approach used by McHale and Relton (2018). What is differing is that three networks are considered, one for each of the aspects considered for passing ability, which makes the work done in this thesis unique. Some chosen network metrics, which are based on the ones used in Pena and Touchette (2012), are calculated to get a quantification of players' influence and importance in a team. The complete network set-up is presented in Chapter 6.

Finally, a motif analysis is performed in Chapter 7 by building a GAM to investigate the effectiveness of different types of passing motifs. Both the results from the passing ability models and the network analyses are utilised as explanatory variables. The idea is to investigate the effects of accurate, tactically good and effective passes on motif effectiveness, and to test whether key players are important in passing motifs that lead to shots. Similar approaches for investigating passing motifs as done in this thesis have not been encountered elsewhere in the literature, although the idea of using network metrics in a regression model is taken from Pina et al. (2017).

Chapter 4

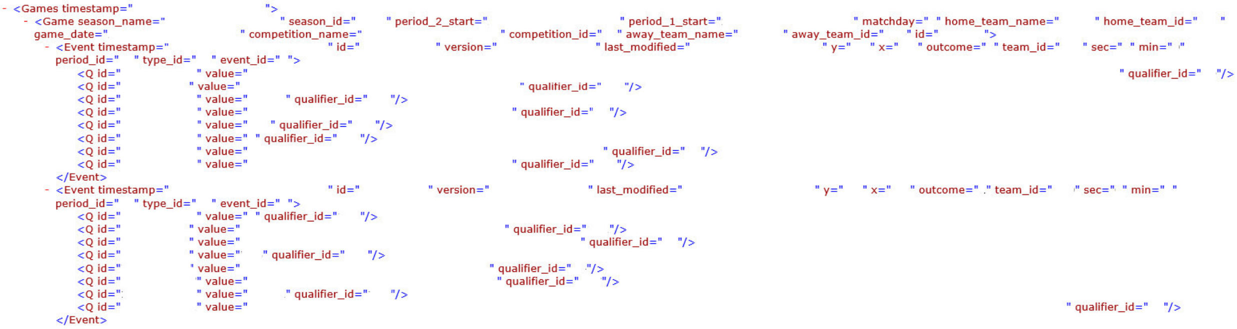
Data

The data used to analyse football players' passing behaviour is introduced in this chapter. Past match data, team ratings and information about the type of ground surface were obtained for the teams playing in Eliteserien during the 2014-2017 seasons.

4.1 Opta Data

Past match data, in the form of event-data, and personal player information were obtained from Opta Sports's (Opta) Opta24Feed. Opta is a credible provider of sports data; they have three analysts collecting data in each match (Greig, 2017). However, the collection process may induce human errors or inaccuracies. Wiig and Håland (2017) discovered a series of errors in an identical data set and took action to deal with some of them. The same measures are taken to handle the errors during data processing here. These measures involve calculating the angle of the ball movement and length of a pass when they are wrongly set to zero, switching the order of shot events and out-of-play events when they occur in the wrong order, manually inserting player positions in cases where they are wrong or incomplete and limiting the ball possession of a player to be less than 30 seconds. For the latter modification, this implies that the time between events is limited to this amount as anything else would be unnatural mid-play.

For each match played in Eliteserien, Opta delivers a separate XML-file containing the corresponding event-data, with every event occurring in the game being represented by a node. Events cover all on-ball involvements in the match, with inclusions of some off-ball situations such as bookings and substitutions. With the data following the movement of the ball, the positioning of players not in possession of the ball is left unknown. Consequently, proxies for opponent pressure are considered in the model set-up when building models to determine players' passing abilities in Section 5.1.2.



```
- <Games timestamp="                    ">
  - <Game season_name="                    " season_id="      " period_2_start="                    " period_1_start="                    " matchday="  " home_team_name="        " home_team_id="     "
    game_date="                    " competition_name="                    " competition_id="   " away_team_name="         " away_team_id="     " id="         ">
    - <Event timestamp="                    " id="            " version="                    " last_modified="                    " y="    " x="    " outcome="  " team_id="     " sec="  " min="  "
      period_id="   " type_id="   " event_id="  ">
        <Q id="            " value="                                                                                    " qualifier_id="    "/>
        <Q id="          " value="                                   " qualifier_id="     "/>
        <Q id="            " value="        " qualifier_id="     "/>
        <Q id="            " value="                                  " qualifier_id="    "/>
        <Q id="            " value="     " qualifier_id="     "/>
        <Q id="            " value="  " qualifier_id="     "/>
        <Q id="            " value="                                         " qualifier_id="    "/>
        <Q id="            " value="                                  " qualifier_id="     "/>
      </Event>
    - <Event timestamp="                    " id="            " version="                    " last_modified="                    " y="    " x="    " outcome=" ." team_id="     " sec="  " min="  "
      period_id="   " type_id="   " event_id="  ">
        <Q id="            " value="  " qualifier_id="     "/>
        <Q id="          " value="                                   " qualifier_id="     "/>
        <Q id="          " value="                                         " qualifier_id="    "/>
        <Q id="          " value="    " qualifier_id="     "/>
        <Q id="          " value="                                " qualifier_id="   "/>
        <Q id="           " value="                               " qualifier_id="     "/>
        <Q id="            " value="        " qualifier_id="     "/>
        <Q id="            " value="                                                                            " qualifier_id="    "/>
      </Event>
```

Figure 4.1: Opta F24 data structure.

As illustrated in the example structure of an event from an Opta data file in Figure 4.1, different identifiers are used to describe the events happening during a game. For each event, the identifiers provide information about the characteristics, the whereabouts and the outcomes of the events. The coordinates used by Opta to locate the happenings on the pitch are given relative to the team in possession of the ball and range from 0 to 100 on both the x-axis (touchlines) and y-axis (goal lines) as seen in Figure 5.1. The playing direction is thus always left to right. Although the pitches in Eliteserien are differing in size, Opta uses the same size for all pitches. The size is identified as being the standard dimensions, 105×68 meters, set by the International Football Association Board, for which Haave and Høiland (2016) provide evidence of coinciding well to the average pitch size in Eliteserien. These dimensions are also utilised when doing calculations to correct errors.

Data processing was conducted in Java, and the XML-files were read using the Java Document Model Interface. Further, the data was restructured and stored in a database with tables containing information about all events, passes and four-sized motifs identified in the data set. In total, 960 matches, 749,859 passes performed by 831 different players and 203,313 motifs are included in the data set. Note that this is the total number, and not necessarily also the number used in the analyses. Passes include passes from open play, headed passes, long passes and crosses.

4.2 Elo Rating

To provide a measure of the strength of the teams playing in Eliteserien, the Elo ratings of the teams, obtained from ClubElo (2018), are included in the analyses. For each team, a text file containing the history of the team's ratings can be downloaded. The ratings are processed in Java, where the team's average score for each month in each season is calculated before it is stored together with the Opta data in a database.

4.3 Ground Surface Type

Information about the type of ground surfaces on football pitches in Eliteserien is obtained from Eliteserien (2018). The ground surface type for each team, either natural or artificial grass, is given in Appendix A. Note that Vålerenga is ticked off on both types of surfaces. The team inaugurated their new stadium with artificial grass in September 2017 after previously having played on natural grass. This is accounted for in the analyses.

Chapter 5

Evaluating Passing Ability

In this chapter, the model set-up to determine the passing abilities of players is introduced and explained, including all variables, the model selection method, the method used to do predictions and the data tools utilised in the analyses. In the end, the results are presented and compared to other relevant studies and the first research question is addressed.

5.1 Model Set-Up

In this thesis, three binary logistic regression models are developed to analyse football players' passing abilities. The models are built on data from the 2014-2016 seasons of Eliteserien. Expanding upon the generalised linear models investigated in Wiig and Håland (2017), which initially were inspired by Szczepański and McHale (2016), the final models will include smooth functions and random effects. Hence, they become GAMMs. Starting from either a plain GLMM or from a GAMM with all continuous variables modelled as smooth terms, the number of smooth terms is determined by comparing the model fit when fixed-effect variables are used to when smooth functions are used instead. The three models look into different aspects of a pass's success defined as pass accuracy, game overview and pass effectiveness, and they are referred to as Model 1, Model 2 and Model 3. A summary of the models can be found in Table 5.1.

Table 5.1: Summary of the passing ability models.

Model	*Dependent variable*	*Description*
1	Y_1	Pass accuracy
2	Y_2	Game overview
3	Y_3	Pass effectiveness

5.1.1 Dependent Variables

The most traditional way of determining the success of a pass is to look at its accuracy, i.e. whether the pass reached its intended target. Model 1 is used to analyse this aspect of a pass, with the dependent variable Y_1 indicating whether the pass successfully reached a player on the same team. To model a more complex aspect of a good pass, namely a player's ability to spot opportunities and take advantage of them in the match, Model 2 considers a player's ability to read the game by having a good game overview. This is incorporated into the model through the dependent variable Y_2, which tells whether the event after a pass was successful or not. Success may only be indicated if the pass itself was successful. Model 3 measures the offensive contribution of a pass in terms of its effectiveness. If the pass was part of a sequence of passes that eventually led to a shot, the dependent variable Y_3 indicates success. If a passing sequence is interrupted by events that do not initiate a new sequence, success is also indicated if a shot was made by the same team within 15 seconds of the last pass in the sequence. Own goals conceded by a team are considered as shots by the team awarded with the goal.

5.1.2 Explanatory Variables

In this section, the explanatory variables are presented in detail. These variables are initially the same for all three models and are chosen based on their presumed influence on the dependent variable of Model 1. As the positioning of the players not in possession of the ball is unknown, several of the variables are proxies for opponent pressure. To determine whether an explanatory variable should be modelled as a smooth function, Model 1 is tested with the applicable variables considered as both smooth terms and fixed effects to find out which type of variable that gives the best fit for each case. Thus, different notations are needed for the variables in question for the case of them being modelled as fixed effects or smooths. Interaction terms will also differ. Table 5.2 presents a summary of the explanatory variables used in the case of a GLMM or a GAMM.

Proxies for Implications of a Player's Position on the Pitch

The positions of the player passing the ball and the player receiving the ball should have an impact on the success of a pass as the players' whereabouts could imply something about the pressure from opponents on both players. To account for this, two categorical variables, $X_{1.z}$ and $X_{2.z}$, are included in the models. The variables indicate where the pass is initiated from and received on the pitch respectively, with z being a specified zone on the field. The pitch is divided into 21 zones as illustrated in Figure 5.1. An interaction between the two zone variables is also tested as Wiig and Håland (2017) found this to be a possible solution to an observed underestimation of offensive passes. By combining zones in such a way, it is natural to assume that the relationship between the zones and the direction of a pass is also somewhat incorporated in the models.

Table 5.2: Summary of the explanatory variables initially used in the passing ability models. Fixed-effect variables are denoted by X, random-effect variables by Z and smooth terms by $f(\cdot)$. Sections GLMM and GAMM give the notation used for a variable when considered as a fixed effect and smooth term respectively. Empty entries in the GAMM section indicate that the same variable as in the GLMM section is used or that the variable is not applicable for this case. Possible types of variables are continuous (C), categorical/factor (F), binary (B) and interaction (I).

GLMM			*GAMM*		
Variable	*Description*	*Type*	*Variable*	*Description*	*Type*
$X_{1.z}$	Start zone z of a pass ($z = 1, \ldots, 21$)	F	$f_1(x_{start},$	4-D smooth handling the starting (x_{start}, y_{start}) and	C
$X_{2.z}$	End zone z of a pass ($z = 1, \ldots, 21$)	F	$y_{start},$	ending coordinates (x_{end}, y_{end}) of a pass	”
$X_{3.d}$	Direction d of a pass ($d = 1, 2, 3, 4$)	F	$x_{end},$	”	”
X_4	Length of a pass	C	$y_{end})$	”	”
X_5	Difference in increased distance from the opponent's goal	C	”	”	”
X_6	Ball changes side of the field	B	”	”	”
$X_{7.s}$	Pass number category s in the current sequence ($s = 1, 2, 3, 4$)	F	$f_2(X_7)$	1-D smooth representing pass number in current sequence	C
X_8	Time passed since the last occurred event	C	$f_3(X_8)$	1-D smooth representing time passed since last occurred event	C
$X_{9.p}$	Game period p ($p = 1, 2, 3, 4$)	F	$f_4(X_9)$	1-D smooth representing game time in minutes	C
$X_{10.gd}$	Goal difference category gd ($gd = 1, 2, 3, 4, 5$)	F	$f_5(X_{10})$	1-D smooth representing goal difference	C
X_{11}	Tackle in the previous event	B			
X_{12}	Aerial duel in the previous event	B			
X_{13}	Interception in the previous event	B			
$X_{14.i}$	The same player who took part in a tackle ($i = 1$)/aerial duel ($i = 2$)/interception ($i = 3$) also made the pass	B			
X_{15}	Previous pass was a header	B			
X_{16}	Current pass is a header	B			
X_{17}	Player performing the pass plays for the home team	B			
X_{18}	Previous event was a free kick	B			
X_{19}	Previous event was a throw-in	B			
X_{20}	Corner taken within the past five events	B			
X_{21}	Elo rating of the opponent team	C	$f_6(X_{21})$	1-D smooth representing the Elo rating of the opponent team	C
X_{22}	Ball recovery due to a loose ball in the previous event	B			
X_{23}	The match is played on artificial grass	B			
$X_{24.m}$	Month m the match is played ($m = 3, \ldots, 11$)	F	$f_7(X_{24})$	1-D smooth representing month of play	C
$X_{25.i}$	The same team executing a corner ($i = 1$)/ free kick ($i = 2$)/ throw-in ($i = 3$) attempted the pass	B			
$X_{26.z}$	The average position of the player by zone z ($z = 1, \ldots, 21$)	F	$f_8(\bar{x}, \bar{y})$	2-D smooth for a player's average position with coordinates ($\bar{x}$, $\bar{y}$)	C
X_{27}	Time played by the player passing the ball	C	$f_9(X_{27})$	1-D smooth representing time played by player passing the ball	C
$X_{1.z1}$*$X_{2.z2}$	Interaction between the start zone $z1$ and end zone $z2$	I	Incorporated in f_1		
$X_{9.p}$*$X_{10.gd}$	Game period p interacting with goal difference gd	I	$f_{10}(X_9, X_{10})$	2-D smooth for the interaction between game time and goal difference	I
X_{23}*$X_{24.m}$	Type of ground surface interacting with month m of play	I	Included in $f_7(X_{24})$ through a *by* argument		
$X_{7.2}$*X_{18}	Pass number 2 interacting with free kick in previous event	I			
$X_{7.2}$*X_{19}	Pass 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 of the player passing the ball ($o = 1, \ldots, 19$)	F			

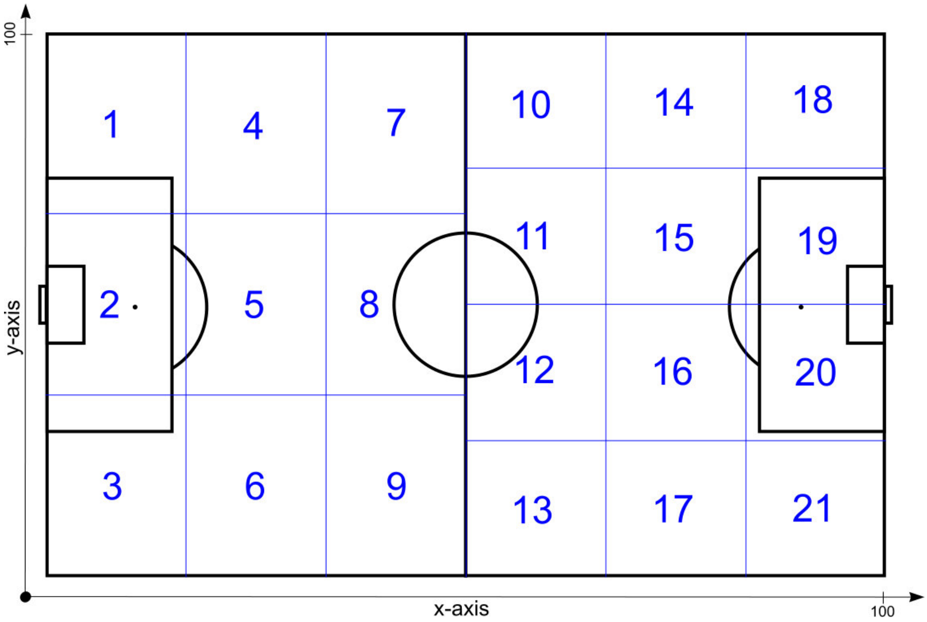


Figure 5.1: The Opta pitch coordinate system divided into 21 zones. Direction of play is left to right, always relative to the team in possession of the ball.

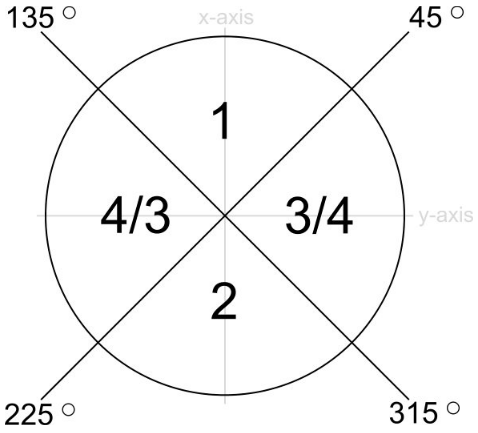


Figure 5.2: Direction of a pass with the player facing forward in the direction of play at 90°.

With a consistent overestimation of the passing abilities of attacking players, Wiig and Håland (2017) proposed adding a variable to deal with the player positions of all players. Rather than adding a categorical variable with categories for each defined player position, an approach similar to the ones taken by Szczepański and McHale (2016) and Bransen (2017) is considered. For each match, the average position occupied by a player is calculated by averaging the x- and y-coordinates of all the events the player is involved in. As players change their player positions both within and between matches, and due to the same player position being occupied differently depending upon a team's playing strategy, the approach seems to be appropriate. From the average values, the corresponding zone is determined and incorporated by a zone average variable, $X_{26.z}$. Thus, each pass observed has a predictor telling the average position of the player in a specific match. This is different from Szczepański and McHale (2016), where the anticipated player position based on previously played matches with exponential weighting is considered to avoid endogeneity, whilst Bransen (2017) also based the average position on the current match.

A categorical variable, $X_{3.d}$, handles the direction of a pass with the direction being one of four possibles: forward, backward, inward or outward. The direction is taken to be relative to the passer's position on the field when facing forward. Hence, inward passes made from either side of the pitch belong to the same category. Figure 5.2 describes how the direction of a pass is categorised. The length of an attempted pass (X_4) serves as a proxy for pass difficulty, while the difference in increased distance from the opponent's goal (X_5) serves as a proxy for both pass difficulty and opponent pressure. Positive values of the latter variable indicate that the ball moves further away from the opponent's goal, and the values are calculated using an x-coordinate of 100 and a y-coordinate of 50 at the goal. Crosses are added to the model with a slightly loose definition that is very much dependent on the location of the players involved in the pass. All passes made where the ball changes side of the pitch are considered as crosses. A dummy variable, X_6, is used for this purpose.

When modelling with smooth functions, the variables $X_{1.z}$ - X_6 and the zone interactions are replaced with a 4-D smooth function consisting of the real starting and ending coordinates of a pass, $f_1(x_{start}, y_{start}, x_{end}, y_{end})$. A 2-D smooth function, $f_8(\bar{x}, \bar{y})$, is used for the coordinates for the average position of a player in a match rather than using the average zone. The average coordinates are calculated similar to the when finding the average zone. However, as zones are no longer considered, the y-coordinates are calculated as absolute distances from the centre of the axis to avoid cancellation of terms for players playing on both sides of the pitch.

Proxies for Opponent Pressure on the Initiating Pass

There are several ways of recovering the ball in open play. Tackles, aerial duels and interceptions are examples of ball recoveries in which an opponent player is likely to be nearby when a pass is made afterwards. As open-play ball recoveries can be used to construct proxies for opponent pressure, binary variables are created for each of the three mentioned recovery methods to indicate whether one of these events happened in the event prior to a pass. The variables are denoted X_{11}, X_{12} and X_{13} for a tackle, an aerial duel and an interception respectively. An extra dummy variable, $X_{14.i}$, is added for each of the three situations to tell whether the player attempting the pass also was involved in the prior event. This is an adjustment to the models developed by Wiig and Håland (2017) and Szczepański and McHale (2016), where only one variable was used for two or more of the situations combined. It is perceived that there might be a difference in the signs of the coefficients for the different cases however. If the same player is involved, it is reasonable to assume for certain that an opponent is close, especially for the cases involving tackles or aerial duels. Opta defines an interception to be happening when a player prevents the ball from reaching its target by intercepting any pass between opponent players. Hence, for interceptions, opponent pressure might not be high.

Ball recoveries are defined by Opta as events in which a team wins possession of the ball and successfully keeps possession for at least two passes or an attacking play. Opta logs ball recoveries both separately and together with tackles, aerial duels and interceptions. All separately logged ball recoveries are assumed to be loose balls, and X_{22} is an indicator of whether this occurred in the previous event. There is no indication of opponent pressure in the case of loose balls. Thus, a separate variable indicating whether the same player is involved is not considered for this case.

Proxies for Opponent Pressure in case of Set Play

Set pieces induce a pause in the match and the players usually cluster before the set play is performed and the play is continued. Thus, players are surrounded by their opponents and are under pressure, which a set of proxy variables handles in the models. A free kick or throw-in in the previous event is indicated by variables X_{18} and X_{19} respectively. In the event of a corner, the effect of the set piece is often lasting longer as both teams move most of their players in front of the goal on the half where the corner is to be taken. Hence, the binary variable X_{20} indicates whether a corner was taken within the past five events. This count is no longer valid if the ball goes out of play, if there is a foul or if the goalkeeper is in control of the ball. Another variable, $X_{25.i}$, is added to tell whether the team that made the set play action also attempted the pass. The effect of this variable is assumed to be more pronounced in the event of a corner as there are huge differences between defence and attack for this case. If the defending team wins the ball, huge areas are open, seemingly making passing easier as opposed to for the attacking team.

Other Proxies for Opponent Pressure

After a ball recovery, the first pass in a sequence is often challenging due to opponent players being close by. The categorical variable $X_{7.s}$ is included to investigate whether the initiating pass is more difficult to attempt than the other passes in a sequence, and the variable can take on four possible values, which are outlined in Table 5.3. Both open-play ball recoveries and set play actions are allowed to be

Table 5.3: The categories into which the sequence number variable $X_{7.s}$ and the goal difference variable $X_{10.gd}$ are divided. The categories in the first column correspond to the index of s and gd for the respective cases.

Cat.	$X_{7.s}$	$X_{10.gd}$
1	First pass in a sequence	Behind with more than two goals
2	Second pass in a sequence	Behind with one or two goals
3	Third, fourth or fifth pass in a sequence	Draw score
4	Sequence number higher than five	Leading by one or two goals
5		Leading by more than two goals

starting points of sequences. It should be noted however that the first pass after the event of a set play has sequence number two. To investigate the impact of the first pass after a set play, two interaction terms are added to separate this pass from other passes with a sequence number of two: the interaction between $X_{7.2}$ and the binary variable telling whether the previous event was a free kick (X_{18}) and the interaction between $X_{7.2}$ and the indicator telling whether the previous event was a throw-in (X_{19}). An interaction term including the corner variable is not added as this variable considers the past five events. When modelled as a smooth function, the sequence numbers are used directly as opposed to the categories defined.

Sequences of passes are handled in the following way. If possession of the ball is kept after an attempted tackle by the opponent team, sequences are terminated and restarted. Moreover, sequences are terminated if the ball goes out of play, the opponent team takes control of the ball, a shot is attempted or a foul is committed. However, if a ball carry is successfully performed by players during the passing sequence, the sequence is continued.

Pass Difficulty

It is perceived that the difficulty of executing a successful pass increases with fewer seconds passed since the last occurred event. To investigate this belief, X_8 is added to represent the time passed in seconds since the last occurred event in the match. Headed passes are seen as more difficult to make accurate than kicked passes, and headed passes in the prior event presumably make it more difficult for the pass recipient to successfully perform the next pass. Indicator variables X_{16} and X_{15} are incorporated in the models to test whether the perceived effects are true for the respective cases. Another aspect which has an impact on the difficulty of making a pass is the overall strength of the opponent team given by the team's Elo rating. The rating, represented by X_{21}, is provided for the month and season a match is played in. In the case of smooth function modelling, X_8 and X_{21} are represented by smooth functions to better capture an assumed non-linear relationship with the dependent variable.

Situational Variables

As players spend more time on the pitch in a match, it is expected that they become more exhausted, which in turn will affect the success of their passes. A continuous variable, X_{27}, represents the time in minutes a player has spent on the pitch when a particular pass is attempted. Game time is also an influencing factor for the outcome of a pass by being a stressing factor for the players; when the time is running out, the chance of getting the desired game outcome is decreasing. Thus, depending on the current score and the time left of the match, teams will play differently to chase victory. Each match in the data set is divided into four equal time periods, represented by the categorical variable $X_{9.p}$. The score at the time a pass is being made is included in the models with a five-level categorical variable, $X_{10.gd}$, with categories that are explained in Table 5.3. This variable is relative to the team in possession of the ball. An interaction between the game time and goal difference variables is added to test the effect of their combination relative to the variables' own contributions.

Other variables describing the external circumstances of a football match which may influence the outcome of a pass include seasonal effects, ground surface type and home team advantage. There should be a fatigue effect when a season is close to an end, while the performance of players is expected to increase in the beginning of the season as their match shape is improved. Further, the climate is changing throughout the year, which also affects the condition of a field.

The month a match is played in is represented by a categorical variable, $X_{24.m}$, only containing months March to November as the matches in Eliteserien are played during these months. It is perceived that maintenance of natural grass is most challenging in the near-winter months, which might influence players' passing abilities. The month of play variable is thus tested in an interaction with an indicator variable for the type of grass on the pitch to investigate the effect of playing on artificial grass compared to natural grass throughout the season. Dummy variable X_{23} indicates whether a match is played on artificial grass or natural grass. To test whether players benefit from playing on their home ground, variable X_{17} is included to indicate if a pass is attempted by the home team.

The categorical variables $X_{9.p}$, $X_{10.gd}$ and $X_{24.m}$ are treated as continuous variables when they are represented by smooth functions. Minutes played in the match is used as opposed to dividing the match into periods, and for the goal difference variable, a positive score indicates that the team is in the lead. The interaction between minutes played and goal difference is considered as a tensor product, while the interaction between type of grass and month of play is modelled as a 1-D smooth function as only one of the variables is continuous. Time played by the player passing the ball is used directly in a smooth function as it already is a continuous variable.

Random Effects

Inspired by Szczepański and McHale (2016), players' passing abilities and teams' abilities to facilitate and hamper passes are included as random effects by variables

$Z_{1.k}$, $Z_{2.t}$ and $Z_{3.o}$ respectively. It is expected that players with high coefficient estimates are among those with the best passing skills in the league. Similarly, teams that achieve high facilitation and hampering scores are perceived to be good at these specific abilities. By assuming normally distributed random effects, noise in the models resulting from players or teams exhibiting exceptionally good or bad skills can be reduced as the coefficients will shrink towards the mean.

5.1.3 Model Selection

To start with, a GLMM with only fixed-effect variables and a GAMM with all the smooth terms presented in Table 5.2, both including a full set of variables, are estimated for Model 1. As the REML scores, which are the outputs from the chosen estimation method, are not comparable across models with differing fixed-effect structures, the AIC scores are computed and used for comparison (Miller, 2017). The model with the lowest AIC score functions as the basis in a stepwise selection to determine which smooth terms should be incorporated in the model. With the GLMM as basis, the smooth term yielding the lowest AIC score after replacing it with its corresponding fixed-effect variable is added to the model in favour of the fixed effect. This is repeated in several steps until no more switches improve the model fit. Similarly, with the GAMM as the basis, smooth terms are removed in favour of their fixed-effect counterparts if appropriate. The procedure of determining smooth terms is assumed to generate similar results for all models. Hence, it is not performed separately for Model 2 and Model 3. After deciding upon which smooth terms to include in the models, the remaining fixed-effect variables are, for each model, selected through the use of a Wald test.

Although stepwise regression might cause some problems in identifying the best fitted model, it is used due to its effectiveness and low computational burden. Additionally, with a data set consisting of a considerably high number of observations, the problem of having highly correlated variables is believed to be minimal.

When having the start and end zone variables for a pass in the models, the interactions between them are also considered. The number of zone interactions to be included in Model 1 is determined by comparing the Brier scores obtained from 10-fold cross validations of the plain GLMM with differing numbers of interactions. The resulting number of interactions is taken to be the same for the two other models, and the number will be used regardless of the combination of fixed effects and smooth terms. As long as the interactions are included in a model, they are not subject to removal when performing the Wald test for variable elimination. The same applies for the stand-alone start and end zone variables.

5.1.4 Predictions and Player Ratings

The three models are estimated based on data covering the 2014-2016 seasons of Eliteserien, and their predictive capability is tested for the 2017 season. Following Szczepański and McHale (2016), predictions must be made differently depending on the usage. Full predictions are used to find the models' expectation of the outcome of a pass. This is a probability that, when averaged out over an entire season for each player, could be compared to the actual observed success rate the player exhibits. Then it can be determined how a player performs compared to the models' expectation. The prediction is based on all predictors in the models. In cases where passes for new players or teams are predicted, they are assumed to be average. This is done by giving them a corresponding random-effect coefficient of zero. However, the prediction method is not suitable to rank players' performance relative to each other. This is better done by looking only at the coefficients of the random-effect variables corresponding to each player.

As the random-effect coefficients for the models in this thesis cover three seasons in one, they will indicate who have been the best and worst passers over three seasons, making it difficult to establish the best players per season. Thus, the likelihood of a successful pass, given that it is performed by the average player, is used instead to rate players. This is equivalent to setting all random effects in the full predictions to zero and is thus the same approach for rating players as the one used by Wiig and Håland (2017). The predicted average likelihood of success can then be compared to the actual outcome for a player, and the ratio between them, over a period of time, determines the rating of the player. To test whether the models are able to predict the same top-ranked players as the training data identifies when being run in a separate model, the player ratings based on predictions are compared to the random-effect coefficients for the 2017 season.

5.1.5 Data Tools and Choice of Reference

All models were run in RStudio, an open-source integrated development environment for the statistical programming language R (R Core Team, 2017). From the previously constructed database, data on all passes was retrieved and stored in RStudio. As the built-in procedure for handling categorical variables only allows the full set of categories, or none of them, to be included in a model, dummy variables were created for each category of the categorical variables used. This will allow for changes in the reference of the categorical variables as omissions of single categories may happen. An initially chosen reference for the categorical variables is needed to avoid singularities, and the references used are presented in Table 5.4. Those only apply when the fixed-effect variable is used in the model rather than its corresponding smooth term. No reference is needed for the random effects as all levels will receive their own coefficient.

The logistic regressions were run using the *bam* function from the *mgcv* package (Wood, 2011). This function is constructed to deal with large data sets and fits a GAM consisting of fixed effects and smooth functions. The parameters are estimated in a single P-IRLS iteration, and the smoothing parameters are estimated by fast REML. To further increase efficiency, discretisation of variables was chosen. Random effects may be added to build a GAMM, and they are then treated as smooth terms. Using this approach, the random effects are seen as penalised regression terms being penalised by a ridge penalty. In the case of running a GLMM, no smooths are added, leaving only the fixed and random effects.

For all 1-D smooth functions, the cubic regression spline was chosen as basis function. This is due to the computational gain implied by this choice rather than using the default thin plate regression splines. Tensor product smooths were chosen for all the multidimensional smooth functions. Although some of the multidimensional smooths include variables that all are representing coordinates, it is assumed that these can not be considered as isotropic due to the difference in what the x- and y-axis are representing on a football pitch. The axes represent a different length per unit move in coordinates and quite different perceptions of the level of pressure on players. For interactions between two continuous variables, tensor smooths were used, while for interactions between a continuous variable and a binary or multi-level categorical variable, a 1-D smooth with a *by* inclusion to the categorical variable was used. With a *by* inclusion, smooth functions for all levels in the categorical variable are constructed. Thus, with a binary variable, two smooths are included in the models.

5.2 Results

In this section, the results of the model selection procedure are presented. Then, a validation of the models is performed and the regression results are discussed. At last, the results are compared to similar studies and the first research question is evaluated.

5.2.1 Model Selection

The number of zone interactions to be included in the models if the f_1 smooth function turned out to be unfavourable was decided by performing several 10-fold cross validations with the Brier score as a performance measure. This concluded the use of 300 interactions. The Brier scores for the different numbers of combinations of interactions tried are tabulated in Table B.1.

Further, a smooth term selection was performed in a stepwise manner starting from a full GAMM and moving backwards replacing single smooth terms with their fixed-effect counterparts. This procedure, presented in Table B.2, was chosen due to the lower AIC from the full GAMM compared to a GLMM. After the first step, it

Table 5.4: The initially chosen references for the possible categorical variables in the passing ability models.

Variable	*Reference*
$X_{1.z}$	Zone 2 ($z = 2$)
$X_{2.z}$	Zone 2 ($z = 2$)
$X_{3.d}$	Direction forward ($d = 1$)
$X_{7.s}$	First pass of sequence ($s = 1$)
$X_{9.p}$	First period ($p = 1$)
$X_{10.gd}$	Draw score ($gd = 3$)
$X_{24.m}$	Month of March ($m = 3$)
$X_{26.z}$	Zone 2 ($z = 2$)

was apparent that the sequence number variable should be treated as a categorical variable. Continuing replacements beyond this variable did not improve the model fit. The smooth function for game time, $f_4(X_9)$, was insignificant for Model 1, but it was decided to keep it in the model. This decision is based on the fact that the model was not improved by instead using the fixed-effect equivalent, and because the variable is significant in an interaction with the goal difference variable. Whether or not this or other smooth terms are insignificant in the two other models are not further discussed as it was presumed that the same smooth functions would be applicable for all models. From the AIC scores, it is evident that the inclusion of the 4-D smooth function most substantially improved the model fit.

Moving on with the above resulting combinations of smooth terms, insignificant fixed-effect variables were removed subject to Wald tests. As opposed to the smooth term selection, a Wald test was performed on the three models separately, giving different final models for the pass aspects considered. The fixed-effect variables remaining in the final models are summarised in Table 5.5, and the complete regression results are presented and discussed in Section 5.2.3.

Table 5.5: Summary of the Wald tests performed on the models after smooth term selection. The remaining fixed-effect variables in each model are marked with a tick.

Variable	*Short description*	*Model 1*	*Model 2*	*Model 3*
$X_{7.1}$	Pass no. 1	*ref*	*ref*	*ref*
$X_{7.2}$	Pass no. 2	✓	✓	✓
$X_{7.3}$	Pass no. 3, 4 or 5	✓	✓	✓
$X_{7.4}$	Pass no. 6+	✓	✓	✓
X_{11}	Tackle	✓	✓	✓
X_{12}	Aerial duel	✓	✓	✓
X_{13}	Interception	✓	✓	✓
$X_{14.1}$	Same player: tackle	✓	✓	
$X_{14.2}$	Same player: aerial duel		✓	
$X_{14.3}$	Same player: interception	✓	✓	✓
X_{15}	Previous pass was header	✓	✓	✓
X_{16}	Headed pass	✓	✓	✓
X_{17}	Home team advantage	✓	✓	✓
X_{18}	Free kick		✓	
X_{19}	Throw-in	✓		✓
X_{20}	Corner		✓	✓
X_{22}	Loose ball	✓	✓	✓
X_{23}	Ground surface type	✓	✓	✓
$X_{25.1}$	Same team: corner	✓	✓	✓
$X_{25.2}$	Same team: free kick			
$X_{25.3}$	Same team: throw-in	✓	✓	
$X_{7.2}{*}X_{18}$	Interaction	✓	✓	✓
$X_{7.2}{*}X_{19}$	Interaction	✓	✓	

5.2.2 Model Validation

The techniques described in Section 2.6 are used to validate the resulting models. Additionally, the fit of the models is tested by evaluating a sequence of passes from a match in the data set.

ROC and PR Curves

In Figure 5.3, the ROC and PR curves for each model are shown with their resulting AUCs. Considering the guidelines presented in Table 2.1, Model 1 exhibits an excellent discrimination ability. The high AUC from the model's PR curve supports this finding. Both Model 2 and Model 3 fall into the category of acceptable discrimination ability, and their fits are thus good according to the ROC curves. The two models however, have quite different PR curves. Model 2 has an AUC that is reasonable, indicating a good fit. Moreover, the curve is concave with a consistently negative slope when disregarding the strange behaviour at recall values of zero. The shape is similar to the one for Model 1 and is beneficial as concavity implies a better model fit due the precision decreasing more slowly than the recall.

For Model 3, one has to take a closer look at the PR curve to better understand its shape. In a PR curve, the first plotted point after the zero recall region has a precision value equal to the highest estimated probability of success given by the model, and the last plotted point has a precision value equal to the percentage of observed successes in the data. As Model 3 has a skewed distribution for both the observed outcomes and the estimated probabilities for the model's dependent variable, the points at which the curve has to start and stop have quite low precision values. Hence, the curve itself must be concave and extremely curved to produce a high AUC, which is very unlikely given the data set. The PR curve is in fact convex, but the curvature is not very extreme. A low AUC could thus plausibly be considered as acceptable, actually indicating a good model fit considering the skewness in the data.

The strange behaviour around the zero recall value seen in all PR plots is due to the threshold values used. When covering the range between zero and one in small intervals, the precision for the first positive values of the recall varies a lot with respect to the truly predicted outcomes in the range above the threshold as few observations are considered in the first intervals. Thus, it is unstable at first, but stabilises when more observations are added with less strict thresholds.

Hosmer-Lemeshow Test Results

The results of the HL tests are given in Table 5.6. The tests were performed on 100 samples from the data, each containing 1000 and 5000 observations, and each sample was divided into ten groups. The number of samples and the significance level used are as proposed by Bartley (2014), who also recommended using a threshold level of ten samples being significant for a model to determine whether the model had a lack of good fit. Following this, it is apparent that Model 2 and Model 3 can be categorised as having good fits when considering only the sample size of 1000 observations, while Model 1 is just outside the range. With the higher

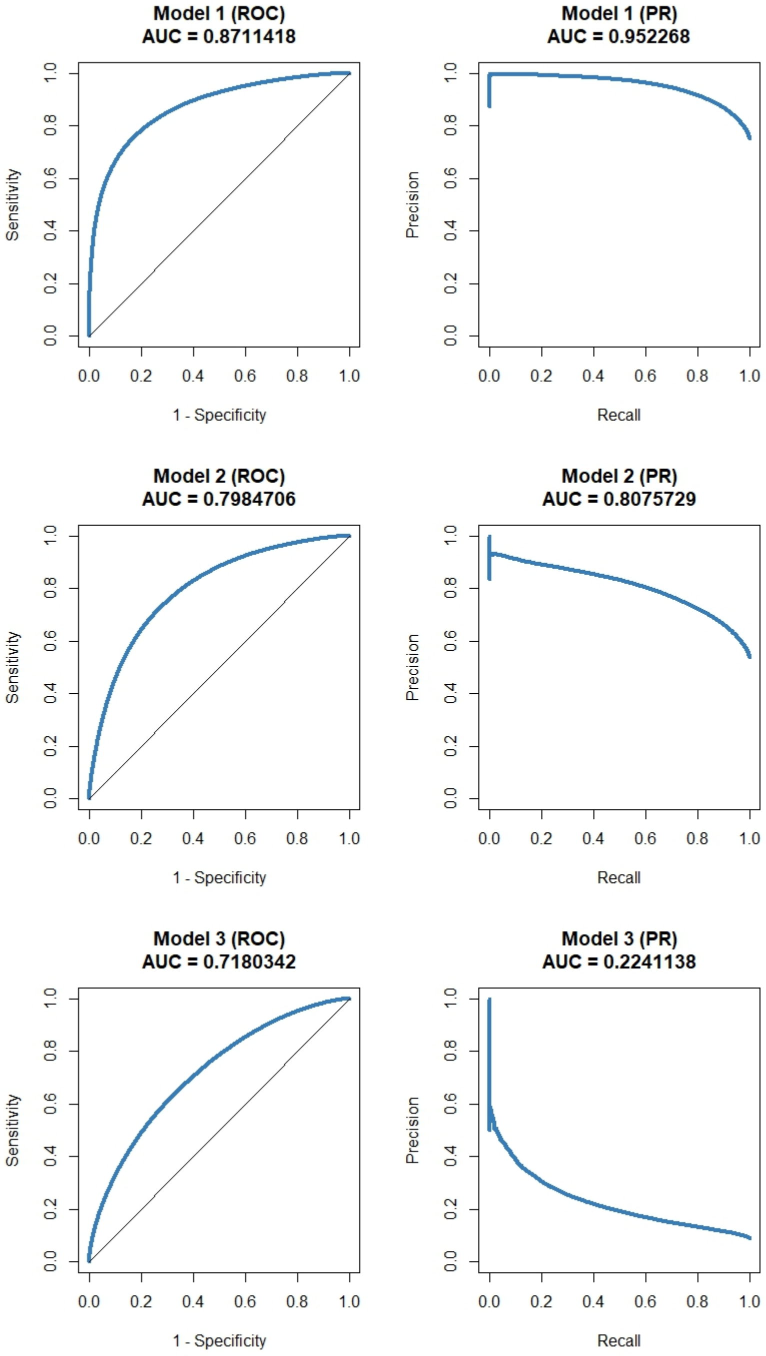


Figure 5.3: ROC and PR curves for the passing ability models.

Table 5.6: HL test results for each model with different sample sizes, n. The given results are the number of samples (out of 100) that has a p-value of less than 0.05.

Model	*n = 1000*	*n = 5000*
1	11	87
2	10	47
3	4	8

sample size of 5000, only Model 3 can be seen as having a good fit. However, as previously mentioned, the results from HL tests must be interpreted with care when dealing with large data sets, which is quite obvious from the differing results given in the table, proving that the HL test is very sensitive to the size of the data set. The results from the lower-sized samples seem to be reasonable in light of the results from the AUCs.

Sequence of Passes for Validation

A sequence of ten passes obtained from a match between Rosenborg and Stabæk is analysed to test the validity of the developed models. The course of the sequence, which is performed by Rosenborg players, is tabulated in Table 5.7 and a graphical representation is shown in Figure 5.4. Note that the variables for which the values are constant between the events are not presented. The match is played in October 2017 on Rosenborg's home ground where natural grass is the surface type. The sequence takes place during the second half of the game and the score at the time of the happening is draw. A failed tackle attempted by an opponent player initiates the sequence, and the sequence ends with a shot. Thus, all the observed values of the dependent variables are one.

For Model 1, all but the last pass in the sequence are predicted to be accurate with a probability above 80%, with the probabilities being highest for passes directed backwards. As most of the passes are made on the sides or on Rosenborg's own half, these values are reasonable. Surprisingly, pass number nine has a quite high predicted probability of success although this pass is made relatively close to Stabæk's goal. This is probably due to the fact that there was a time gap of eight seconds between the pass and the previous event. The last pass in the sequence is directed forward and the ball moves into the penalty area of Stabæk, which intuitively makes this pass more difficult to attempt. This perceived difficulty is captured by the model as seen from the lower predicted probability of success.

In the case of Model 2, all passes made backwards have a considerably higher predicted probability of being followed up compared to the passes made forward. Similar to Model 1, the last pass in the sequence is the most difficult to follow up. For Model 3, the probabilities of success increase for passes made forward or towards the centre of the field. Moreover, the probabilities are higher closer to the opponent's goal. The probabilities for all passes in terms of both Model 2 and Model 3 are reasonable as the pressure from the opponents usually is lower on a team's own half and as shots usually are attempted close to the opponent's goal.

Table 5.7: Values of the variables in a sequence of passes made by Rosenborg players that is used for validation. Rosenborg played against Stabæk in October 2017 on their own home ground covered with natural grass. The score was draw, and Stabæk's Elo rating at the time of play was 1284.13. Refer to Figure 5.4 for a graphical representation of the starting and ending coordinates. All other variables do not change between events and are thus not tabulated below. There are no headed passes, and the sequence is initiated right after a tackle in which the first passer is not partaking.

Failed tackle by Stabæk player Ronald Hernández											
No.	*Player*	*Pos*	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	X_7	X_8	X_9	$\bar{x}$	$\bar{y}$	X_{27}
1	Samuel Adegbenro	WI	0.871	0.658	0.086	1	3	67	63	36	67
2	Morten Konradsen	CM	0.989	0.895	0.094	2	2	67	48	39	6
3	Tore Reginiussen	CD	0.998	0.894	0.084	3	3	67	33	19	67
4	Vegar Hedenstad	FB	0.854	0.558	0.091	3	4	68	43	39	68
5	Nicklas Bendtner	ST	0.851	0.625	0.104	3	4	68	64	26	68
6	Vegar Hedenstad	FB	0.899	0.549	0.115	4	3	68	43	39	68
7	Pål André Helland	WI	0.984	0.818	0.094	4	4	68	77	35	68
8	Nicklas Bendtner	ST	0.980	0.839	0.132	4	2	68	64	26	68
9	Tore Reginiussen	CD	0.833	0.635	0.192	4	8	68	33	19	68
10	Mike Jensen	CM	0.576	0.315	0.260	4	1	68	59	26	68
Missed shot by Pål André Helland											

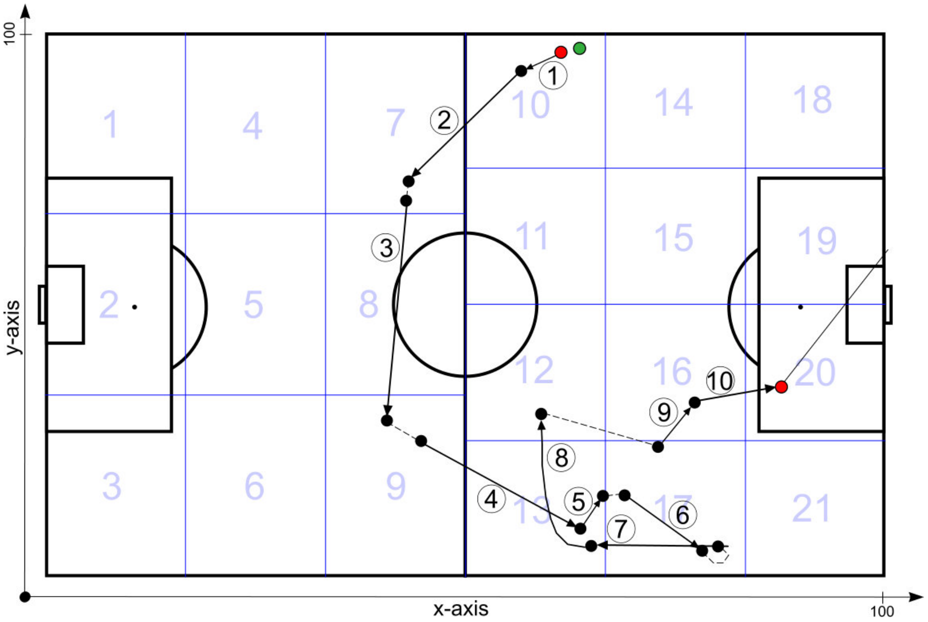


Figure 5.4: A graphical representation of the sequence of passes used for validation. The start and end of the sequence are represented by red dots, and a green dot illustrates where the preceding tackle occurred. Dotted lines are either ball touches or ball carries.

5.2.2.1 Overall Assessment of Validity

According to the AUC values for the ROC and PR curves, all models can be classified to have good model fits. Variations in the validity across the models are reasonable as each model considers a different aspect of a pass's success. Model 1 seems to have the best fit, which is also within reasoning as the explanatory variables initially were developed based on their presumed impact on pass accuracy. Access to more detailed data, e.g. player tracking data, would probably have contributed to better model fits. The evaluation of the models' predicted probabilities in a chosen passing sequence reveals intuitive results for all models. Considering the HL test, the results need to be interpreted with care. Under some circumstances however, the results of the HL tests coincide with the results from the other validation methods used. Overall, there exists evidence to conclude that all three models have reasonably good fits.

5.2.3 Regression Results

The results of the final GAMMs are presented and discussed in this section. Tables of the complete regression results, including both the fixed effects and the smooth terms, can be found in Appendix C. A significance level of 10% is used and positive signs of the coefficients correspond to an increased probability of success. It should be noted that the sensitivity of an explanatory variable in a logistic regression model depends on all variables in the model. Consequently, one can not directly extract from the results the extent to which each respective coefficient influences the overall change in likelihood due to a unit change in the corresponding explanatory variable.

Model 1 is built on a total of 565,720 observations of passes. In Model 2, three passes are omitted due to these having no defined next event in the game, while one observation is missing for Model 3. This single pass is the last event of the last match processed in the data set, which seems to be the reason why it is not recorded properly. However, it is unlikely that the omission has had any impact on the regression results.

Fixed Effects

The coefficients of all fixed-effect variables included in the final models are presented in Table 5.8. Other than the intercepts, no coefficients appear to be very high or low in terms of magnitude. It is apparent that the first pass in a sequence of passes (the reference) is the most difficult pass to successfully execute. If the first pass is initiated after a tackle, the probability of success increases, while it is opposite in the case of an aerial duel. This is true also when the same player is involved. Surprisingly, the expectation of high opponent pressure on the initiating pass due to a tackle preceding it is not supported by the results. However, different types of tackles are not considered here, and many of them, such as sliding tackles, might leave the opponent on the ground, meaning that the opponent pressure is not very high. For interceptions, the effect is differing between the models, although it is

Table 5.8: Fixed-effect coefficients of the estimated GAMMs. All tabulated entries are significant at a level of 10%.

Variable	*Short description*	*Model 1*	*Model 2*	*Model 3*
$X_{7.2}$	Pass no. 2	0.351	0.311	0.240
$X_{7.3}$	Pass no. 3, 4 or 5	0.498	0.471	0.223
$X_{7.4}$	Pass no. 6+	0.580	0.515	0.223
X_{11}	Tackle	0.343	0.189	0.325
X_{12}	Aerial duel	−0.100	−0.596	−0.134
X_{13}	Interception	−0.327	−0.148	0.091
$X_{14.1}$	Same player: tackle	0.154	0.097	
$X_{14.2}$	Same player: aerial duel		0.532	
$X_{14.3}$	Same player: interception	0.757	0.387	0.338
X_{15}	Previous pass was header	−0.311	−0.294	−0.110
X_{16}	Headed pass	−1.227	−0.968	−0.785
X_{17}	Home team advantage	0.138	0.117	0.092
X_{18}	Free kick		−0.181	
X_{19}	Throw-in	0.184		−0.106
X_{20}	Corner		−0.172	0.363
X_{22}	Loose ball	0.515	0.277	0.289
X_{23}	Ground surface type	0.123	0.187	0.127
$X_{25.1}$	Same team: corner	0.226	0.120	−0.156
$X_{25.3}$	Same team: throw-in	−0.215	−0.367	
$X_{7.2}$*X_{18}	Interaction	0.306	0.543	−0.093
$X_{7.2}$*X_{19}	Interaction	0.225	0.447	
Intercept		1.299	−0.517	−3.248

throughout positive when the same player who is involved in the interception also makes the pass. The perception that the pressure on the passer could be lower in such circumstances is thus supported.

As expected, aerial duels in the previous event complicate the task of executing a successful pass. This negative effect is further increased with the inclusion of the contribution from headed passes. A headed pass will always be the outcome of an aerial duel, and it will always be made by one of the players involved in the duel. Thus, it makes sense that the variable indicating whether the same player being involved is insignificant for some of the models. Additionally, the positive effect in Model 2 is offset by the more negative effect of the header variable for the same model. In general, headed passes, both in current and previous events, have a negative influence on the success rate of passes. Loose balls in the previous event contribute to a higher probability for the player to make a successful pass according to all models, which is reasonable as opponents are assumed to not necessarily be nearby during these actions.

For Model 1 and Model 2, the degree of difficulty decreases with the pass number in the sequence, while for Model 3, all passes with a pass number higher than one are almost equally difficult to make compared to the first pass in the sequence. When including contributions from the interaction terms between pass number and free kick, the likelihood of success for the second pass in a sequence is increased for Model 1 and Model 2, while it is decreased for Model 3. Whether a pass is attempted by the same team performing a free kick in the prior event is insignificant for all models. This implies that a team does not have an advantage over the other team when attempting the first pass after a free kick. For throw-ins, the probabilities of making accurate and tactically good passes increase when adding the interaction term, as opposed to an unchanged negative effect on pass effectiveness. The positive effects in the first two models are reduced when the same team takes the throw-in and attempts the subsequent pass. Apparently, the team seems to be under high pressure from opponents during such set pieces.

Corners within the five previous events have differing impact on the models. While the accuracy of a pass is not affected by the corner variable itself, it is positively affected when a corner is made by the attacking team, which intuitively is opposite of what is expected. The attacking team is subject to a more confined space with opponents surrounding them, which should have made the attempted pass more difficult to make. Tactically good passes are more difficult for the defending team, while the influence on the success rate for the attacking team is minimal. This is counter-intuitive in the same way as for the pass accuracy. The corner effect on pass effectiveness should intuitively be more pronounced for the attacking team as they are closer to the opponent's goal. However, the results suggest otherwise. The defending team has a higher likelihood of success, possibly due to counter-attack possibilities in the open areas ahead after the corner is taken.

All models indicate that a team tend to benefit from playing on their home ground when passing the ball as seen from the positive coefficients. Also, players seem to have an advantage from playing on artificial grass.

Random Effects

The random-effect coefficients of the teams playing in Eliteserien during the 2014-2016 seasons are shown in Figure 5.5 for all models. In the figure, the teams' skills of facilitation and hampering, with respect to the aspects of success, are plotted on the vertical and horizontal axis respectively. The higher the coefficient value, the better a team is on the specific skill.

For all models, Sogndal, Mjøndalen and Tromsø are placed in the third quadrant, indicating both low facilitation and hampering skills. Interestingly, these three teams have been in the lower half of the table and some of them have also been relegated from Eliteserien during the seasons considered. Surprisingly, teams such as Rosenborg, Molde and Strømsgodset, all of which have been seen as favourites at the start of the seasons, and have performed well during the seasons considered, have negative hampering scores in all models. Lillestrøm and Aalesund stand out to be the teams with the best hampering skills for each pass aspect.

Considering the facilitation scores, Odd and Rosenborg are the teams achieving highest scores in terms of pass accuracy and game overview, while Rosenborg, followed by Sarpsborg 08, have the highest pass effectiveness scores. Statistics from

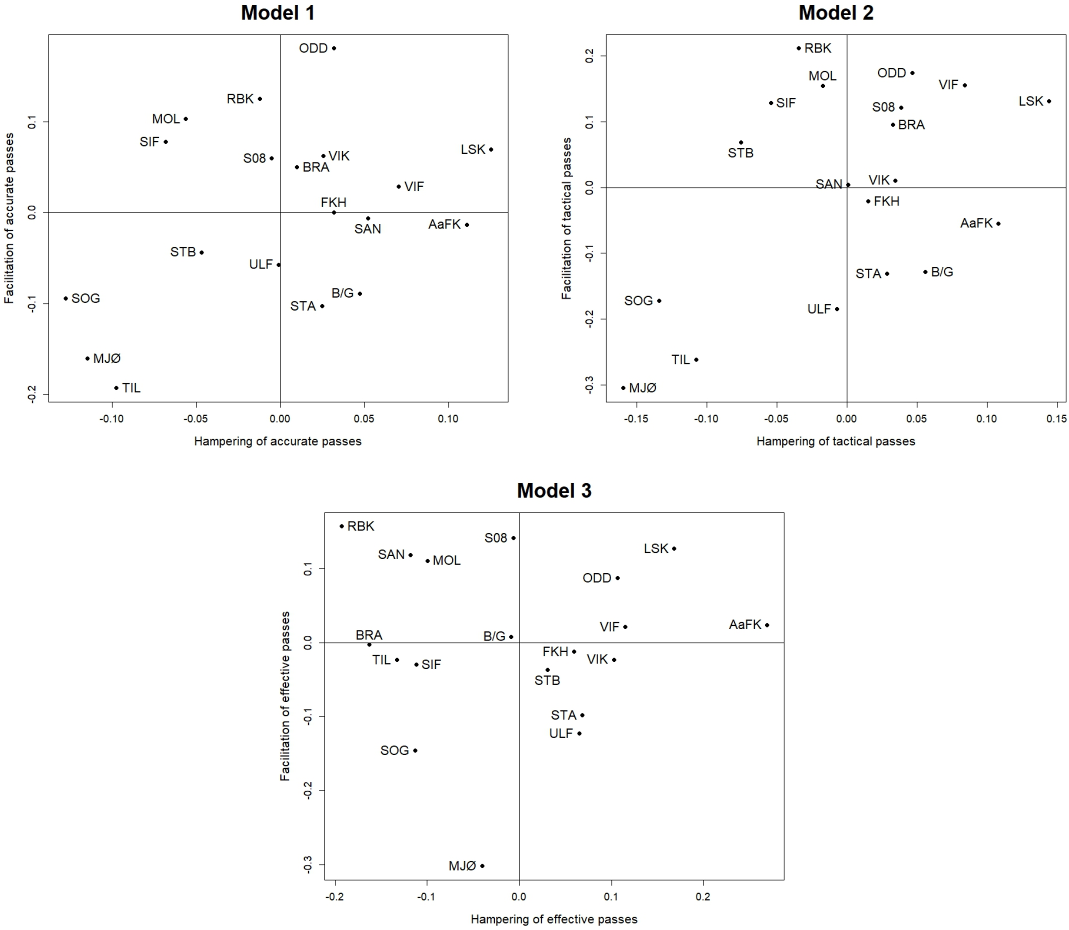


Figure 5.5: Random-effect coefficients of teams in Eliteserien for all models. The coefficients are based on data from the 2014-2016 seasons. The overall team skill of hampering passes is given on the x-axis while the overall team skill of facilitation is given on the y-axis, both skills with respect to the pass aspects considered. Team abbreviations are explained in Appendix D.

WhoScored.com (2018) state that the top five teams in terms of pass accuracy in Eliteserien 2016 were, listed in descending order, Rosenborg, Odd, Strømsgodset, Vålerenga and Molde. Moreover, the teams with the highest number of shots per game in the season were Rosenborg, Molde, Sarpsborg 08, Lillestrøm and Strømsgodset. Although being valid only for the 2016 season, the statistics support the observations in the figures in terms of the facilitation scores.

In general, most teams stay within, or close to, the same quadrant in all models, and the scores are realistic when considering the skill represented on the y-axis. The hampering scores however, are not as intuitive. Especially the occurrences of

Table 5.9: The random-effect coefficients of the top ten passers during the 2014-2016 seasons of Eliteserien for all models. Player position and team abbreviations are explained in Appendix D. If a player has played for more than one team in the seasons considered, the team for which the player has played more matches during these seasons is given.

Panel A: Model 1

#	*Player*	*Team*	*Pos*	*Coef*
1	Johan Lædre Bjørdal	RBK	CD	0.473
2	Martin Ødegaard	SIF	AM	0.422
3	Christian Grindheim	VIF	CM	0.419
4	Daniel Berg Hestad	MOL	CM	0.412
5	Markus Berger	STA	CD	0.401
6	Joona Toivio	MOL	CD	0.397
7	Tomasz Sokolowski	STB	CM	0.394
8	Johan Andersson	LSK	CM	0.391
9	Even Hovland	MOL	CD	0.341
10	Enoch Kofi Adu	STB	CM	0.337

Panel B: Model 2

#	*Player*	*Team*	*Pos*	*Coef*
1	Magnar Ødegaard	TIL	CD	0.278
2	Anthony Annan	STB	CM	0.275
3	Karol Mets	VIK	CM	0.264
4	Giorgi Gorozia	STB	CM	0.263
5	Johan Lædre Bjørdal	RBK	CD	0.260
6	Indridi Sigurdsson	VIK	CD	0.252
7	Anthony Soares	VIK	CD	0.251
8	Jukka Raitala	SOG	FB	0.247
9	Fredrik Michalsen	TIL	CM	0.238
10	Hólmar Örn Eyjólfsson	RBK	CD	0.237

Panel C: Model 3

#	*Player*	*Team*	*Pos*	*Coef*
1	Joona Toivio	MOL	CD	0.148
2	Magnar Ødegaard	TIL	CD	0.140
3	Kristian Brix	B/G	WB	0.137
4	Mohamed Ofkir	LSK	WI	0.124
5	Giorgi Gorozia	STB	CM	0.110
6	Jone Samuelsen	ODD	CM	0.104
7	Maic Sema	FKH	AM	0.099
8	Ernest Asante	STB	WI	0.099
9	Eirik Mæland	FKH	CM	0.098
10	Erlend Hanstveit	BRA	FB	0.098

Lillestrøm on the far right in all models are surprising due to the team being known for a direct playing style.

The random-effect coefficients of the top ten passers during the 2014-2016 seasons in Eliteserien are shown in Table 5.9 for each model. Central defenders top the lists for all models, and central defenders and central midfielders dominate the top lists for both Model 1 and Model 2. As opposed to the two other models, more variation in player positions is present in the top list for Model 3. The fact that more offensive players are included in the top ten list for Model 3 compared to the other models is reasonable as this model considers whether a pass is part of a passing sequence resulting in a shot. Shots are usually attempted by offensive players on the opponent's half, and since many passing sequences are terminated fairly quickly, the patterns seen for Model 3 are supported. The inclusion of defensive players in the top list for the model could be due to longer passing sequences or it could be explained as an effect from counter-attacks where the team is moving the ball forward from defence to attack rather fast.

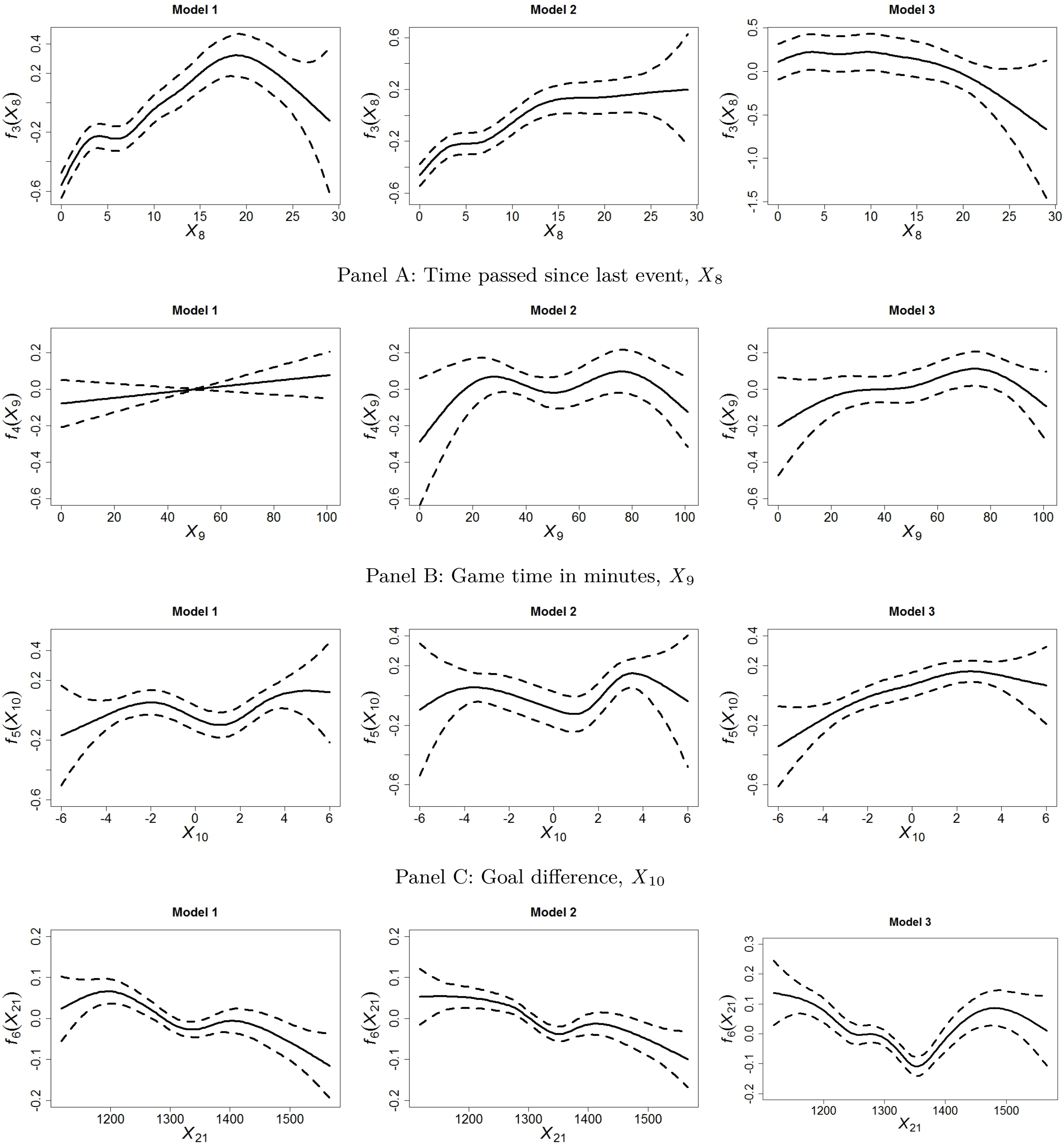
In terms of the magnitude of the coefficients, the range of the Model 1 coefficient values is greater than of the other models, especially compared to Model 3. This could possibly be due to the higher number of observed successful events for Model 1 compared to the other models. In fact, the success rates are 74.9%, 53.7% and 8.7% for Model 1, Model 2 and Model 3 respectively.

To test whether the magnitude of a player's coefficient is influenced by the magnitude of the coefficient of the player's team, the models were run with the team coefficients excluded. The resulting models had higher AICs and were thus not investigated further. However, one should bear in mind a potential correlation between team and player random effects. On the other hand, players playing for teams with differing scores are present in the top lists for all models.

Smooth Terms

The resulting 1-D smooth functions are illustrated in Figure 5.6, and most of them are similar across the models. Panel A illustrates the effect of time passed since the last occurred event measured in seconds. For Model 1 and Model 2, the probability of making an accurate or a tactically good pass is increasing with increased time passed, which is intuitive as opponents are not likely to be nearby the passer in such circumstances. In Model 3, the trend is that effective passes are more difficult to make when the time between events increases. This is not intuitive, but could be explained as an effect from counter-attacks. With the ball moving fast between players in the forward direction, the likelihood for a counter-attack to be successful, and thus also the likelihood for success according to Model 3, could be increased. The negative slopes present for Model 1 and Model 3 after about 18 seconds are probably due to few observations having long time gaps between events.

The effect of game time in minutes is shown in Panel B. It was expected that players would become stressed when the time runs out, which in turn could affect their pass success rates. For Model 2 and Model 3, the pattern is quite similar. Players have an increasing probability of making tactically good or effective passes in the first period of each half of the match, while it is more difficult to make these



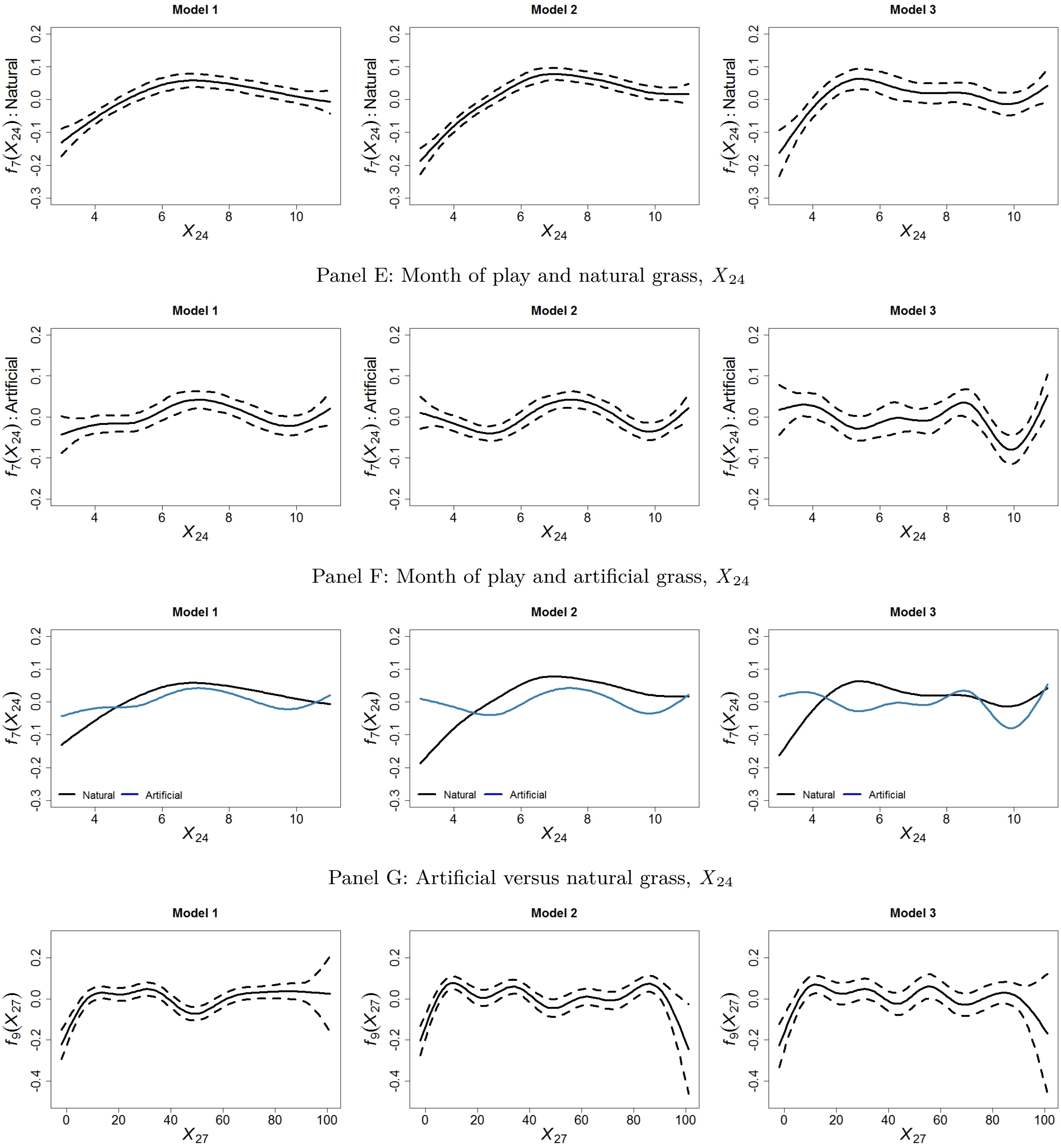
Panel A: Time passed since last event, X_8

Panel B: Game time in minutes, X_9

Panel C: Goal difference, X_{10}

Panel D: Elo rating of the opponent team, X_{21}

Figure 5.6: The resulting 1-D smooth functions with 95% confidence intervals.



Panel E: Month of play and natural grass, X_{24}

Panel F: Month of play and artificial grass, X_{24}

Panel G: Artificial versus natural grass, X_{24}

Panel H: Time played by player, X_{27}

Figure 5.6 (Continued): The resulting 1-D smooth functions with 95% confidence intervals.

passes at the end of the halves, including overtime. In the second period of the first half in Model 3 however, the increase is levelled out. Seemingly, the stressing effect of time is captured by these models. For Model 1 however, a straight line for the function is estimated, giving a linear predictor with a counter-intuitive positive slope. This linear shape is probably due to the newly added variable of player-specific game time, which is highly correlated to game time ($r = 0.83$) as most players play the entire match.

In Panel C, the influence of the goal difference variable is illustrated. Effects from goal differences higher than four in absolute value should be interpreted with care due to few occurrences in the data set. For Model 1, the chance of success is high when a team is falling behind with two goals or leading by at least four goals, while in Model 2, the likelihood of succeeding is highest when a team is falling behind with or leading by four goals. Making accurate or tactically good passes are most challenging when a team is leading by one goal, which makes sense as the opponent team tends to play a pressing game during such circumstances. Thus, it seems like players have a higher likelihood of success according to these models when the goal difference is big, that is, when the match outcome seems to be definite. In terms of effectiveness, a draw score corresponds to a near zero contribution, with teams falling behind having reduced probability of success and teams being in the lead having increased probability of success. Seemingly, teams chasing a goal to even the score are less likely to succeed with their passing sequences, possibly due to a stressing factor of being in need of a goal.

For Model 1 and Model 2, a higher Elo rating of the opponent team makes it harder for a player to achieve success, as seen in Panel D. The same pattern is partly true for Model 3, but when playing against teams with a very high Elo rating, the chance of success is increased. Although this is counter-intuitive, it could be due to counter-attacks. Playing against very good teams with high ball possessions could more often create opportunities for counter-attacks. This might increase the number of observations of successful sequences when playing against such teams, and thus potentially increase the probability of success for Model 3.

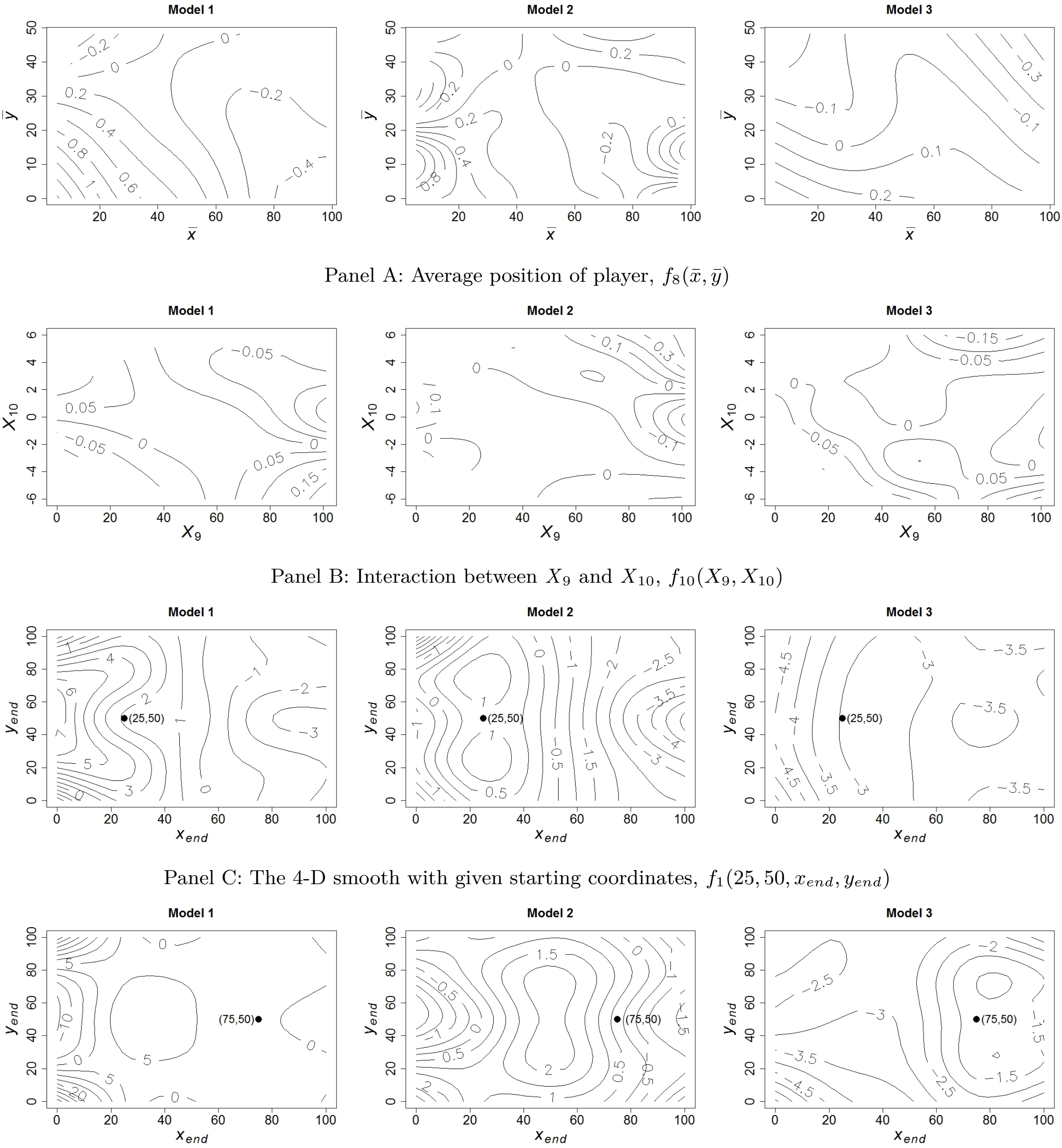
It was perceived that players' performance would increase in the beginning of a season and then decrease in the end of the season due to a fatigue effect. Further, changes in weather conditions throughout the year were anticipated to affect players' passing abilities. Panel E and Panel F display the interactions between month of play and natural grass and month of play and artificial grass respectively. As the shapes of the functions for the two surface types are dissimilar, it is reasonable to assume that it is the effects of the ground conditions, and not the fatigue effects, that are captured. With an upward slope in the beginning of the season for the natural grass graphs, and a downward slope in the end, it seems like the difficulties regarding maintenance of natural grass during the near-winter months are captured by the models. For the artificial grass functions however, varying curves, with small oscillations, indicate little change in the conditions of a field. Comparing the two smooth functions for month of play in Panel G, it is not apparent that playing on artificial grass is more beneficial throughout the year. However, when also considering the positive fixed-effect coefficients for artificial grass, the blue line would be shifted upwards, revealing a clear advantage of playing on artificial grass in all models, especially in the months of March and April.

Panel H illustrates the effect of playing time, measured in minutes, for the player passing the ball. For all models, the likelihood of success is highest when a player has been on the pitch between 15 and 85 minutes when attempting the pass, while players having played the entire match seem to have their performance drop in the last minutes of the game. Thus, the anticipated exhaustion effect seems to be captured by the models, with the effect being more prominent for Model 2 and Model 3.

For the multidimensional smooth functions in Figure 5.7, the resulting contributions are varying between the models. The differing contribution given by the average position of a player is shown in Panel A. When using absolute values for the y-coordinates, the y-axis only ranges from zero to 50, meaning that low y-coordinates correspond to players both on the left-hand and right-hand side of the pitch. All models indicate that the average position of being in the bottom left corner in the figures, where the full backs usually are situated, is the easiest. The corners on the opponent's half, the flanks and the area in front of the opponent's goal are most difficult to succeed in for Model 1, Model 2 and Model 3 respectively. The layout of the variable is very intuitive when seen in light of the results found by Wiig and Håland (2017). With them observing that offensive players seemingly have an overestimated passing ability, and the opposite for defenders, the problem is slightly dealt with as defensive players are given high, positive contributions and offensive players are given negative contributions. Thus, passes attempted by defenders and attackers would be expected to be easier and harder to succeed with by the models respectively.

It was perceived that the combined game time and goal difference would have an impact on the performance of players, especially when a match is approaching the end. For all models, the probability of success in the end minutes is higher for the teams falling behind than the teams being in the lead as seen in Panel B. Thus, the team leading actually has a harder task even if they have a comfortable lead. However, with fewer observations of leads, or defeats, of four or more goals, some of the most extreme contributions must be interpreted more carefully.

Panel C and Panel D both show results from the 4-D smooth function that covers the starting and ending coordinates of a pass. Given a pair of starting coordinates, the contour lines give the contribution for all end-coordinate possibilities. When passing from the defensive half (Panel C) the likelihood of success is higher when passing backwards compared to passing forward according to Model 1 and Model 2. This is intuitive as the opposition is situated in front of the passer at most times. For Model 3, there is little variation in the magnitude of the contribution, but passing the ball forward gives a higher probability of success. This is also intuitive as the ball is moving closer to the opponent's goal where shots are more likely to happen. The same intuitive results are present in the case of a pass made from the offensive half in Panel D as well.



Panel A: Average position of player, $f_8(\bar{x}, \bar{y})$

Panel B: Interaction between X_9 and X_{10}, $f_{10}(X_9, X_{10})$

Panel C: The 4-D smooth with given starting coordinates, $f_1(25, 50, x_{end}, y_{end})$

Panel D: The 4-D smooth with given starting coordinates, $f_1(75, 50, x_{end}, y_{end})$

Figure 5.7: The resulting multidimensional smooth functions are depicted in panels A-D.

5.2.4 Predictions and Player Ratings

In the following, the predictions and player ratings obtained for the developed models are presented. The full predictions used to test the models' predictive powers and the average predictions used to rate players are calculated as explained in Section 5.1.4.

Predictions

In Figure 5.8, the predicted success rate is plotted against the observed success rate for each model and each player in the 2017 season of Eliteserien. Only players attempting a minimum of 100 passes are plotted. The colours represent groups of player positions, and different symbols are used to tell whether a player is new to the 2017 season compared to the 2014-2016 seasons.

For Model 1 and Model 2, more strikers seem to be on the left side of the diagonal compared to the right side, indicating that the models predict a higher likelihood of success for them than what is actually observed. The same pattern, although not as clear, can be seen for wingers and attacking midfielders. Considering the magnitude of the success rates, defenders, defensive midfielders and central midfielders tend to achieve higher success rates than goalkeepers and more offensive players. It is also interesting that the discrepancy between the predicted and observed success rates tend to be smallest for the case of high rates. Overall, the largest discrepancy between the rates seems to be present for strikers.

For Model 3, it is clear that the different groups of player positions cluster together. Moreover, the magnitude of the success rates are increasing in the playing direction, that is, goalkeepers have the lowest success rates while more offensive players obtain the highest rates. This is intuitive as the model measures pass effectiveness and because most of the passing sequences in the data set are short. As opposed to Model 1 and Model 2, the distribution of offensive players in Model 3 seems to be more even.

Player Ratings

The ten best passers in the 2017 season of Eliteserien for each model are presented in Table 5.10. Note that only outfield players attempting more than 209 passes are considered for the analysis. This is the equivalent of having attempted an average number of passes in 20% of the matches, i.e. six matches, that season.

Defensive players dominate the top list for Model 1, while the proportion of offensive players in the lists for the two other models is higher. However, there are still five defensive players in the list for Model 3. As there are many passes made by defenders on their team's own half that are not effective, the model might not be able to give realistic expected values for the defenders who tend to be more active in the offensive play. Hence, compared to other defenders, the more offensive defenders are recognised by the model, but their ratings might not be comparable to those for other player positions.

For comparisons of player ratings in the 2017 season, the developed models were rerun with data from the 2017 season and the resulting random-effect coefficients

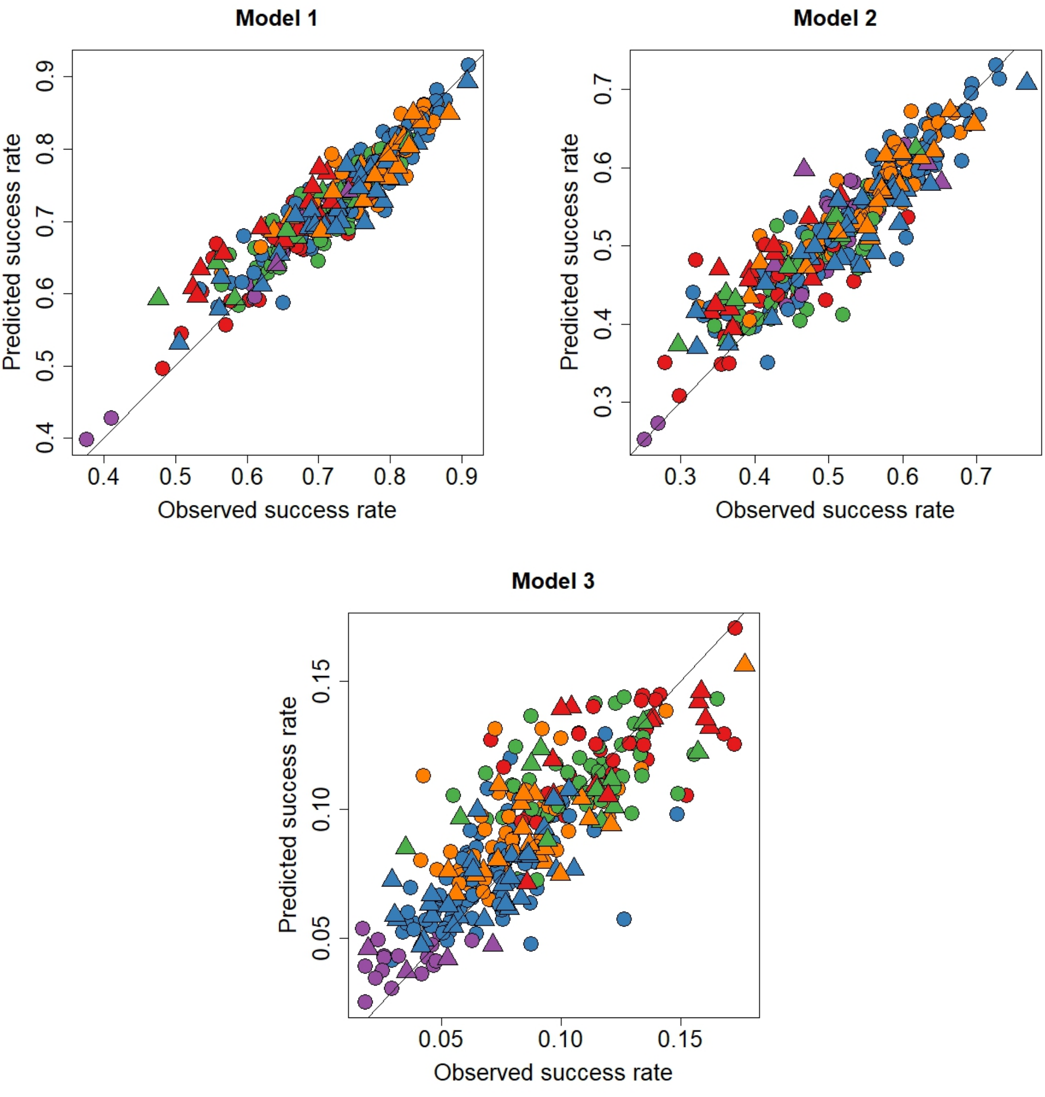


Figure 5.8: The predicted versus observed success rates for each model and each player in the 2017 season of Eliteserien. The colours represent groups of player positions, and different symbols are used to tell whether the player is new to the 2017 season compared to the 2014-2016 seasons (▲) or not (●). A new player is treated as an average player with a random-effect player coefficient of zero. Colours: ● - goalkeepers, ● - defenders, ● - central and defensive midfielders, ● - wingers and attacking midfielders and ● - strikers. Only players with more than 100 passes are considered.

Table 5.10: The top ten passers in the 2017 season of Eliteserien. Explanation of the abbreviations used for the teams and the player positions can be found in Appendix D. Only outfield players with more than 209 passes, the equivalent of playing six matches, are considered, and if a player has played for more than one team in the season considered, the team for which the player has played more matches is given.

Panel A: Model 1

#	*Player*	*Team*	*Pos*	*Expected*	*Actual*	*Obs*	*Ratio*
1	Daniel Braaten	BRA	ST	0.673	0.741	282	1.102
2	Vegar Hedenstad	RBK	FB	0.744	0.812	1348	1.090
3	Bonke Innocent	LSK	DM	0.757	0.823	294	1.087
4	Espen Ruud	ODD	FB	0.706	0.763	1504	1.081
5	Reiss Greenidge	SOG	CD	0.707	0.764	229	1.081
6	Thomas Grøgaard	ODD	FB	0.762	0.812	1219	1.067
7	Taijo Teniste	SOG	FB	0.681	0.727	673	1.067
8	Michael Haukås	VIK	WI	0.659	0.702	242	1.065
9	Martin Ellingsen	MOL	CM	0.761	0.810	406	1.065
10	Birger Meling	RBK	FB	0.778	0.829	981	1.065

Panel B: Model 2

#	*Player*	*Team*	*Pos*	*Expected*	*Actual*	*Obs*	*Ratio*
1	Daniel Braaten	BRA	ST	0.425	0.496	282	1.167
2	Anders Trondsen	S08	CM	0.532	0.608	1152	1.142
3	Herman Stengel	VIF	CM	0.593	0.677	1007	1.142
4	Jacob Rasmussen	RBK	CD	0.673	0.767	675	1.140
5	Vegar Hedenstad	RBK	FB	0.532	0.605	1348	1.137
6	Kasper Skaanes	BRA	WI	0.482	0.545	297	1.133
7	Enar Jääger	VIF	FB	0.646	0.731	1281	1.132
8	Jonatan Nation	VIF	CD	0.605	0.685	1771	1.132
9	Birger Meling	RBK	FB	0.572	0.646	981	1.131
10	Mathias Normann	MOL	CM	0.531	0.600	255	1.130

Panel C: Model 3

#	*Player*	*Team*	*Pos*	*Expected*	*Actual*	*Obs*	*Ratio*
1	Jostein Gundersen	TIL	CD	0.052	0.087	492	1.699
2	Lasse Nilsen	TIL	FB	0.097	0.149	444	1.535
3	Daniel Braaten	BRA	ST	0.120	0.152	282	1.388
4	Kim André Madsen	SIF	CD	0.055	0.076	476	1.373
5	Deyver Vega	BRA	WI	0.111	0.149	302	1.340
6	Gjermund Åsen	TIL	AM	0.117	0.156	867	1.333
7	Martin Broberg	ODD	WI	0.092	0.122	483	1.316
8	Vegard Bergan	ODD	CD	0.058	0.075	771	1.307
9	Vegard Forren	MOL	CD	0.064	0.084	392	1.307
10	Anthony Ikedi	FKH	CM	0.076	0.100	261	1.303

Table 5.11: The random-effect coefficients of the top ten passers during the 2017 season of Eliteserien according to the reran models. Abbreviations for teams and positions are explained in Appendix D. The team for which a player has played most matches in the season considered is given.

Panel A: Model 1

#	*Player*	*Team*	*Pos*	*Coef*
1	Espen Ruud	ODD	FB	0.383
2	Vegar Hedenstad	RBK	FB	0.367
3	Thomas Grøgaard	ODD	FB	0.362
4	Henning Hauger	SIF	DM	0.347
5	André Danielsen	VIK	CM	0.328
6	Taijo Teniste	SOG	FB	0.325
7	Vito Wormgoor	BRA	CD	0.288
8	Bonke Innocent	LSK	DM	0.288
9	Jonatan Nation	VIF	CD	0.288
10	Nikita Baranov	KBK	CD	0.283

Panel B: Model 2

#	*Player*	*Team*	*Pos*	*Coef*
1	Henning Hauger	SIF	DM	0.230
2	Jacob Rasmussen	RBK	CD	0.218
3	André Danielsen	VIK	CM	0.203
4	Mikkel Kirkeskov	AaFK	FB	0.195
5	Ulrik Yttergård Jensen	TIL	CD	0.186
6	Thomas Grøgaard	ODD	FB	0.185
7	Christoffer Aasbak	KBK	WB	0.184
8	Kristoffer Haraldseid	FKH	WB	0.184
9	Jonatan Tollås Nation	VIF	CD	0.184
10	Enar Jääger	VIF	FB	0.180

Panel C: Model 3

#	*Player*	*Team*	*Pos*	*Coef*
1	Gjermund Åsen	TIL	AM	0.090
2	Jostein Gundersen	TIL	CD	0.085
3	Kaj Ramsteijn	AaFK	CD	0.083
4	Liridon Kalludra	KBK	WI	0.083
5	Lasse Nilsen	TIL	FB	0.077
6	Christian Grindheim	VIF	CM	0.076
7	Fredrik Midtsjø	RBK	CM	0.067
8	Eirik Hestad	MOL	CM	0.067
9	Vegar Hedenstad	RBK	FB	0.058
10	Vegard Bergan	ODD	CD	0.058

of the top ten passers are shown in Table 5.11. The idea was to see whether the model predictions could be used to rate players out of sample instead of having to run the models over again for each new season to get new coefficient estimates.

When comparing the player ratings across the two approaches used, five, three and four of the same players on the top ten lists are present on both lists according to Model 1, Model 2 and Model 3 respectively. Two of these players are new to the 2017 data set, but still manages to be spotted by the models through predictions. Five more new players are present in Table 5.11, three of which also plays for Kristiansund, the newly promoted team. In general, the player position patterns are the same across the models for the two approaches.

5.2.5 Comparison with Existing Literature

The models developed in this thesis are modifications of the models considered in Wiig and Håland (2017), with the main difference being the inclusion of smooth functions and the addition of some new variables and interaction terms. Not surprisingly, the results are similar, but the models' fits are improved according to the AUCs for both the ROC and PR curves. When looking at the fixed-effect variables that are exactly the same in the two studies, almost all of them have the same signs of the coefficients and a value in the same range. The tendency is that where deviations are apparent, new interaction terms or new variables that most likely have changed the effect of the variable have been introduced, so that the total contribution might actually be more similar than what is evident.

When comparing to other existing studies, Model 1 is inspired by the model considered in Szczepański and McHale (2016), which was further investigated by Tovar et al. (2017) and McHale and Relton (2018). Model 3 has a nature similar to what is examined in Power et al. (2017), Brooks et al. (2016) and Mackay (2017), and in general papers on other sports, whereas Model 2 is seemingly unique in the way the dependent variable is defined.

The coefficients of Model 1 correspond well with the findings in Szczepański and McHale (2016) in terms of both sign and magnitude. For the two other studies, the fixed-effect variables used are quite different, and the home team advantage variable, which is the only comparable variable, has the same positive effect. Both McHale and Relton (2018) and Tovar et al. (2017) have chosen to force linearity upon continuous variables instead of using smooth functions. Additionally, McHale and Relton (2018) have the advantage of having a much more detailed data set with player tracking data, making it possible to create more accurate variables to handle opponent pressure than the proxies used in this thesis. They suggest that the more accurate variables are the reason why many of their initially proposed variables turn out to be insignificant.

Although Szczepański and McHale (2016) have the most alike smooth functions to the ones considered here, these have somewhat different appearances. Time between events, defined as time between passes in Szczepański and McHale (2016), has the same shape in the start of the graph, but then turn out to be very different. Most likely, this is due to the fact that the x-axes have differing ranges, with the range in this thesis going far beyond the eight seconds in Szczepański and McHale

(2016). Further, the effect of game time has a similar pattern across the two studies. The average player position is defined rather differently, but still shows similar results in the offensive part of the field, while the curvature is more extreme in this thesis for the defensive part. Finally, for the 4-D smooth function, the result is only similar in the defensive region. All other smooth functions are newly proposed, and the random effects are not comparable as different leagues are considered in the two studies.

The predictions made for pass accuracy in the consecutive season seem to be more accurate in Szczepański and McHale (2016). However, as new players are considered in this thesis as opposed to what is seemingly done by Szczepański and McHale (2016), inaccuracies will be present as these players are seen as average.

The measures introduced by Power et al. (2017) are related to the models developed in this thesis. The risk of a pass, defined as the probability of a player executing the pass taking its difficulty into account, relates to Model 1, while the reward of a pass, defined as the probability that the pass made ends with a shot within the next ten seconds, relates to Model 3. To a higher extent than what is seen in this thesis, offensive players tend to achieve higher ratings in terms of making effective passes in Power et al. (2017), which is probably due to the time limit used. In both studies however, the probability of making an effective pass increases closer to the opponent's goal even though these passes are more difficult to successfully execute.

Brooks et al. (2016) also studied a case similar to Model 3, providing insights into which combinations of origin and destination zones that tend to lead to shot opportunities. Although the results are not directly comparable as zones are not utilised in this thesis, some comparisons can be made. Brooks et al. (2016) found that having the ball in the *critical* zone, that is, the zone just in front of the opponent's penalty box, more often leads to shots. This is however not true for the case of making long passes to the critical zone from specific zones on a team's own half. Both observations are supported by the results in this thesis as seen from the corresponding coordinates on both 4-D smooth plots for Model 3 in Figure 5.7.

Although looking at goal effectiveness instead of shot effectiveness, the results obtained by Mackay (2017) are related to Model 3. Mackay (2017) investigated the probability that different types of actions and their previous events result in a goal. Hence, not only passes are considered. Only the variable for headed passes in the current action, while several of the variables concerning actions occurring in the previous event are comparable. In fact, all but one of these comparable variables have the same signs across the two models. The only variable differing in sign is the event of tackles in the previous event, which is found to negatively contribute to goal-scoring opportunities in Mackay (2017).

Just as previously mentioned about the results in Power et al. (2017), offensive players are also favoured in Brooks et al. (2016) and Mackay (2017). This is different from the findings in this thesis, where more variation in player positions is present for Model 3. This might indicate that Model 3 more properly deals with differences in player positions, although it is uncertain how fair the given scores across players positions are.

For the models developed in this thesis, some variables that have not, to the authors' knowledge, been used elsewhere in research related to passing are included to investigate their effects on passes' success. Most interestingly are the results from the surface type variable combined with the course of a season. They show that artificial grass has a considerably big positive effect on the outcome of a pass compared to natural grass for all aspects of success that are considered. Also, the effect of the contribution from natural grass seems to be consistent with the changing conditions of such a surface type throughout the year. Although making a decision about which surface type to use on a field based on these observations is way too drastic, the findings could be seen as proof that teams' should consider different playing styles or passing patters depending upon the quality of the surface.

The inclusion of set pieces as binary variables has also been tested in the analyses, and they reveal some surprising results about passes in Eliteserien. After a corner for instance, the defending team seems to be more likely to have a shot attempt, while the attacking team has a greater chance of succeeding with the passes they make. Having this information in mind, an attacking team may consider to choose a more safer option in the event of a corner, perhaps by reconsidering the number of defenders to be moved forward, as this might minimise the chance of counter-attacks.

5.2.6 Evaluation of Research Question 1

In this chapter, the primary objective was to answer the first main research question outlined in Chapter 1:

> **RQ 1:** Which factors influence the success of a pass in Eliteserien?

To answer this, three GAMMs were developed to assess the probability of success in terms of making accurate, tactically good and effective passes. From the regression results, it is apparent that many factors influence a pass's success, and their contributions are differing across the three aspects considered. The proxies for the implications of a player's position on the pitch reveal quite intuitively that passing the ball backwards, or keeping it on the team's own half, is considerably easier, while the opposite direction however, is more beneficial when looking for shot opportunities. Considering the average player positions, full backs tend to be more likely to succeed.

The proxies for opponent pressure are affecting the outcome of a pass in different ways. For the first pass in a sequence, the variables for types of ball recoveries in the prior event considered have changing, but reasonable signs. Tackles were, at first glance, surprisingly affecting the outcome positively. However, it was established that this variable is not specific enough to handle the opponent pressure related to all types of tackles. In the case of set plays in the previous event, free kicks do not affect the outcome of a pass that much, while for throw-ins, it turned out that the team not taking them has a higher probability of success. For corners within the five previous events, there were rather strange results, which are identified as might being due to counter-attacks.

Interestingly, the effect of varying ground conditions is captured by the models and the perception that playing on artificial grass has an advantage over natural grass is supported. The functions for game time in minutes have intuitive shapes for Model 2 and Model 3, while for Model 1, the function is linear, possibly due to a strong correlation between game time and the time played by a player. Considering the scores, the team in need of a goal seems to be less likely to successfully make effective passes, while, in general, accurate and tactical passes are more likely to be the outcome when the goal difference is big. Headed passes in current and previous event, the Elo rating of the opponent team and the time passed since previous event all give intuitive results in terms of how they affect the difficulty of a pass. Also, there seems to be an home team advantage.

A second question to answer, the sub-question of RQ 1, is: Who were the best passers in the 2017 season of Eliteserien? When setting a reasonable limit of passes to be attempted in order to remove players with few observations, the top lists make sense. Both players playing for different teams and in different positions are recognised among the best passers in the season. Using the notion of average players in the ratings seems to give intuitive results when comparing to the random-effect coefficients in the rerun models for the 2017 data. Similar player positions and many of the same players are represented in the lists for both approaches.

Chapter 6

Identifying Key Players

To identify the key players in teams, network analyses are performed using the results from the passing ability models in Chapter 5. Basic network theory is presented in Section 2.9. Here, the network set-up is introduced, followed by a validation of the network metrics considered and the results. The key Rosenborg players are identified and two case studies from Rosenborg matches are presented. At last, the results are compared to similar existing studies and the second research question is evaluated.

6.1 Networks in the Context of Football

By constructing a social network of team players, it is possible to identify the influence and importance of football players in a team-passing dynamic game through different network metrics. In a passing network, players are represented by nodes and edges represent interactions between the players, where the interactions usually have been passes made.

In the context of football, closeness can be interpreted as a measure of the easiness of reaching a particular player within a team. Players achieving higher closeness scores tend to reach more players in fewer passes (Clemente et al., 2016). The betweenness score gives an indication of how the ball-flow between teammates depends on a specific player. Players with high scores play a key role as connecting bridges between teammates, and they contribute to keep ball possession within the team (Gonçalves et al., 2017). With the usual approach of using the number of passes between players as weights in a network, a low betweenness score is associated with less involvement in the game, and the effect of removing that player from the game is minimal. From the standpoint of a team, betweenness scores that are evenly distributed among players may indicate a well-balanced passing strategy and less dependence on particular players (Pena and Touchette, 2012).

For team sport analysis, the PageRank centrality is a recursive notion of *popularity* or *importance*. From a recipient's perspective, a player is important when receiving passes from other important players, while from a passer's perspective, a player is important when passing the ball to other important players. Basically, the PageRank centrality assigns to each player the probability that the player will receive or pass the ball after some passes have been made (Pena and Touchette, 2012).

Clustering coefficients are computed to get a quantification of players' tendency to cluster together. Such coefficients can be used to assess how close a particular player and his teammates are to become a complete subgraph (Clemente et al., 2016). A high individual clustering score may indicate that a player acts as a middleman for his teammates, and by averaging the players' individual coefficients, a team's clustering coefficient may provide insight into how well-balanced the team is (Pena and Touchette, 2012).

6.2 Network Set-Up

To identify the key players on teams in Eliteserien, both directed and undirected weighted passing networks are created to obtain the chosen network metrics. The passing networks consist of all players in a team for a given season who have made or received at least one pass, and the metrics are calculated in three separate cases: one for each of the aspects of a pass's success considered in Model 1, Model 2 and Model 3 (see Section 5.1.1). The passes that are relevant to analyse for each of the aspects need different weighting, which is further explained in the following section. Furthermore, the network metrics calculated might differ between the aspects. The networks were constructed in RStudio using the *igraph* package (Csardi and Nepusz, 2006).

6.2.1 Defining Weights Matrices

For the case of pass accuracy, the difficulties of the completed passes are used as weights, an approach introduced by McHale and Relton (2018). Both the passer and the recipient are thus seen as better if they pass or receive passes that are more difficult to make respectively. The regression results from Model 1 are used to decide the difficulty of a pass by utilising its average predicted likelihood of success, p_1. The average prediction of a pass can be thought of as its easiness due to it being separated from the skill of the player and the teams. Then, the pass difficulty is given as $1 - p_1$.

When considering game overview, the average predicted probabilities from Model 2, p_2, are used to define the weights in the networks. These predictions give the probability that the pass recipients are able to successfully perform the action in the event succeeding a pass. As passers do a better job if they deliver a pass that is easier to follow up, and as recipients do a better job if they are able to follow up a pass that is expected to be hard to do so with, the weights are separated for the two cases. Thus, two graphs are considered for each team: one handling the receiving view of a pass and one handling the passing view. The former has weights $1 - p_2$ and the latter has weights given by p_2.

The dependent variable for effectiveness of a pass in Model 3 is defined in such a way that similar approaches for edge weights as the ones utilised for the accuracy and game overview networks are not deemed appropriate. This is due to the fact that the pass recipient can not be identified as the player attempting the shot in the end of a passing sequence. Hence, the recipient can not be seen as a good contributor of effectiveness just because he or she was able to receive a pass with a low probability of being effective. Therefore, the sum of effective passes between players is used as weights to favour players that more often are involved in the offensive play. These weights are used for both passers and recipients.

For each of the three network types, the weights associated with all passes going from one player to another in a season are summed up in a weights matrix, W. As a recipient is needed for all passes, only successful passes are considered. Moreover, for the game overview and pass effectiveness networks, only those passes that are successful according to their defined dependent variables in Model 2 and Model 3 respectively are analysed. The weights matrices are normalised by dividing each entry w_{ij} by the total number of matches in which player i has made a pass to player j during the season considered. Also, the edge weights are defined as strengths. Consequently, higher weights between players are favourable, and the shortest paths between nodes in a network are decided as defined in equation (15). This way, the highest scores given by the network metrics are assigned to the key players in a team. The three networks with different specifications of weighting are referred to as Network 1, Network 2 and Network 3. A summary of the networks is given in Table 6.1.

Table 6.1: Summary of the networks.

Network	*Description*
1	Pass accuracy
2	Game overview
3	Pass effectiveness

6.2.2 Network Metric Specifications

For this thesis, three centrality measures for a weighted network are considered: closeness, betweenness and PageRank. Additionally, the Barrat clustering coefficient is calculated. The metrics are defined in Section 2.9, and are computed for each player and each season in the data set. For Network 1, all measures are considered, while only the PageRank centrality is considered for Network 2 and Network 3. If a player has played for different teams in one season, the player is given several scores that season.

As the weights in the networks are considered as strengths, the closeness and betweenness measures are calculated by using inverse weights as distances in equations (16) and (18). Dijkstra's algorithm is utilised in the estimation of closeness, while Brandes's algorithm is used for betweenness. In the *igraph* package, it is specified that if there is no applicable directed path between two nodes, the shortest path between the nodes in the calculation of closeness is set to the total number of nodes in the network. If this is also true for weighted networks, situations in which this default distance is actually shorter than some of the shortest paths between other nodes in the network might occur. However, whether this would appear to be a problem in this thesis is questionable as the majority of the players are bidirectionally connected and have been involved in many passes throughout the seasons. When the inverse weights are used as distances such that the weights can be treated as strengths, most of the distances will have a length between zero and one. Thus, the majority of the shortest paths should be less than the number of nodes in the networks.

The PageRank centrality is considered for both the pass recipient and the passer, with the two cases being referred to as $PR^w_\mathrm{R}(i)$ and $PR^w_\mathrm{P}(i)$ respectively. The passer's perspective is obtained by switching the direction of the edges in Network 1 and Network 3. When considering Network 2, the two separate graphs are each used to calculate one of the PageRanks: $PR^{w2}_\mathrm{R}(i)$ in the recipient graph and $PR^{w2}_\mathrm{P}(i)$ in the passer graph. Equation (21) is used in the calculations, and the damping factor is set to 0.85 in all cases, a number that is commonly used by researchers.

All centrality measures are normalised by the maximum score of the measure in the corresponding network. Hence, based on a given centrality measure, the key player in a team gets a score of one. To calculate the Barrat clustering coefficient in equation (22), an undirected graph is needed. This is accomplished by collapsing the corresponding edge weights in the directed graph by averaging them.

6.3 Results

In this section, a validation of the network metrics computed for the players in Eliteserien is performed. Then, the key players on Rosenborg in the 2017 season are presented together with case studies from two Rosenborg matches. Further, the network approach used in this thesis is compared with the ones used in other studies and the second research question is evaluated.

6.3.1 Validation

Two approaches are considered to validate the computed network metrics for the players in Eliteserien. First, the correlation for each metric across two consecutive seasons is presented in Table 6.2. Here, only players playing on the same team in both seasons are part of the analysis. Second, the correlation between all metrics, based on the data from all seasons, is shown in Figure 6.1. The idea is that the different metrics all have the same purpose of finding the key players in a team. Consequently, the player rankings should be similar across the measures. Moreover, when playing for the same team in two consecutive seasons, the importance of a player has, most likely, not changed very much.

Table 6.2: Network metric correlation across seasons in Eliteserien for validation. For two consecutive seasons, only players having played on the same team in both seasons are considered.

	2014/2015	*2015/2016*	*2016/2017*
$C_{\mathrm{C}}(i)$	0.778	0.739	0.842
$C_{\mathrm{B}}(i)$	0.494	0.434	0.326
$PR_{\mathrm{R}}^{w1}(i)$	0.798	0.799	0.748
$PR_{\mathrm{P}}^{w1}(i)$	0.550	0.578	0.608
c_i^w	0.152	0.077	0.186
$PR_{\mathrm{R}}^{w2}(i)$	0.612	0.554	0.603
$PR_{\mathrm{P}}^{w2}(i)$	0.670	0.696	0.625
$PR_{\mathrm{R}}^{w3}(i)$	0.626	0.546	0.525
$PR_{\mathrm{P}}^{w3}(i)$	0.423	0.429	0.398

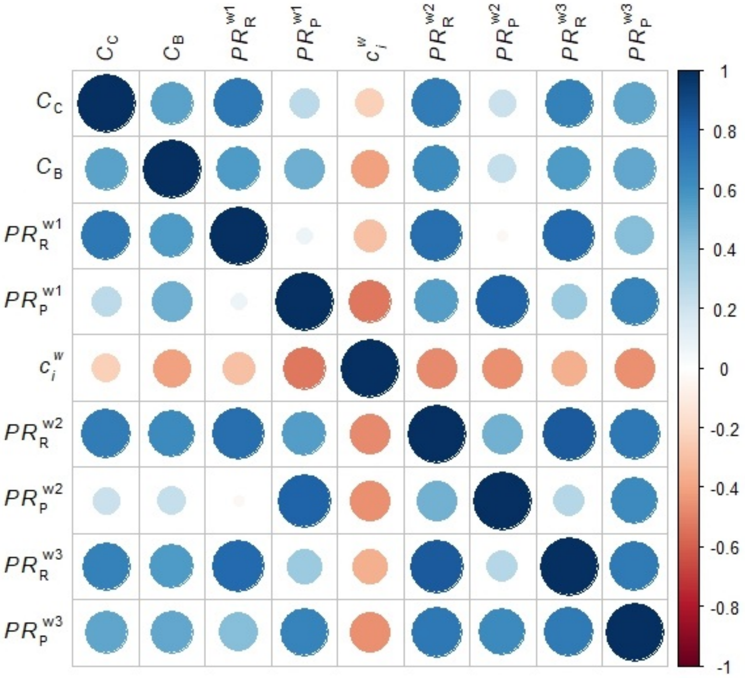


Figure 6.1: Correlation between the different network metrics. Players from the 2014-2017 seasons of Eliteserien are considered.

From Table 6.2, it is evident that there is a tendency for the same players to be ranked similarly in two consecutive seasons. This is especially true for the cases of the closeness measure and the PageRank recipient score of Network 1 as seen from the high correlation across seasons for these measures. As players do develop their abilities across seasons, and their scores depend upon the team's key player for a given season, the correlation for the centrality measures seems to be reasonable as a near-perfect correlation would be impossible. The clustering coefficient has a rather low correlation between seasons, indicating that this metric does not provide consistent scores for players across seasons. This could be explained by looking into how players receive their clustering scores. If a player is connected to exactly two other players in the graph, the strength of the player (node strength) is equal to the sum of the weights in equation (22), giving a coefficient value of one. However, when the number of connections is increasing, but still covering a low percentage of the total number of players on the team, the score is much lower, but increasing with the number of added connections. Hence, players with a slight increase in involvements from one season to another might have very differing scores for these seasons, which would lower the correlation.

The correlation between different network metrics in Figure 6.1 reveals which measures that tend to rate players in the same order. The highest positive correlation can be found for the three PageRank scores, for both recipients and passers. Although the closeness measure has been defined such that both the passes received and the passes made are considered, the measure has a higher correlation with the PageRank passer scores compared to the PageRank recipient scores. The Barrat clustering coefficient has a slightly negative correlation to all other metrics. This could be due to the same reason as explained for the correlation across seasons; some players with little involvement are rated to be the best on this measure, but are not in the top ratings for the other measures due to few connections.

6.3.2 Key Players

Although the network set-ups described are used for all teams and seasons in the data, only the resulting key players on Rosenborg in the 2017 season are presented here due to the thesis focusing on this team and due to spacing issues. However, the general results found across teams are also discussed. Overall, intuitive results were found for all teams, with the teams' perceived most important players achieving higher scores on the calculated network metrics.

Network 1

The key Rosenborg players in the 2017 season in accordance with the network metrics for Network 1 are shown in Table 6.3. Only players having been involved in, i.e. passed or received, a number of passes corresponding to six fully played matches are listed to avoid noise. The cut-off is set to 462 pass involvements, a number based on Rosenborg players' average number of passes made and received per match in 2017. Six matches correspond to having played 20% of the games in a season.

Table 6.3: The key players on Rosenborg in the 2017 season of Eliteserien according to Network 1 are shown. Only players involved in more passes than the equivalent of six matches are considered. The three best players according to each measure are highlighted in bold text. The number of pass involvements, n, is the sum of passes made and received by a player, where passes made also include those that were unsuccessful and thus not part of the network analysis. The abbreviations for player positions are explained in Appendix D.

Name	*Pos*	$C_{\mathrm{C}}(i)$	$C_{\mathrm{B}}(i)$	$PR_{\mathrm{R}}^{w1}(i)$	$PR_{\mathrm{P}}^{w1}(i)$	c_i^w	n
André Hansen	GK	0.319	0.000	0.088	0.693	0.800	1083
Johan Lædre Bjørdal	CD	0.766	0.496	0.205	**0.980**	0.842	1555
Jacob Rasmussen	CD	0.738	0.044	0.255	0.836	0.851	1174
Tore Reginiussen	CD	0.697	0	0.203	0.718	0.915	2259
Jørgen Skjelvik	CD	0.797	0.062	0.275	0.901	0.886	1968
Alex Gersbach	FB	0.907	0.035	0.521	0.670	0.903	854
Vegar Hedenstad	FB	**0.999**	**1.000**	0.672	**1.000**	0.849	2420
Birger Meling	FB	**0.992**	**0.646**	0.909	**0.916**	0.821	1747
Anders Konradsen	DM	0.878	0.115	0.604	0.625	**0.932**	1479
Mike Jensen	CM	0.947	0.487	0.788	0.738	0.836	1927
Marius Lundemo	CM	0.735	0.000	0.400	0.596	**0.944**	1257
Fredrik Midtsjø	CM	0.977	0.425	0.751	0.664	0.910	1119
Anders Trondsen	CM	0.910	0.257	0.566	0.771	0.914	633
Pål André Helland	WI	0.943	0.177	0.665	0.562	0.919	920
Milan Jevtovic	WI	**1.000**	**0.681**	**0.952**	0.628	**0.925**	767
Nicklas Bendtner	ST	0.950	0.133	**1.000**	0.584	0.910	1465
Matthías Vilhjálmsson	ST	0.881	0.310	**0.977**	0.469	0.857	596

As explained earlier, closeness is a measure of the easiness of reaching a player. Hence, when using pass difficulties to find the shortest paths between players, players who tend to receive and pass more difficult passes will get higher scores. In general, midfielders and attacking players tend to receive high scores, which makes sense as these players are situated in positions on the pitch where passes are given higher weights. For Rosenborg however, their full backs have also received high scores, with two of them, Hedenstad and Meling, being ranked among the top three most important players on the team. This suggests that Hedenstad and Meling are important players in the attacking phase of the team's play, perhaps by making good passes to their teammates on the offensive half.

For the betweenness scores, the same three players as for the closeness measure are among the top three ranked players in Rosenborg. However, there are no clear patterns of which player positions or groups of players that are ranked higher. It seems like players are awarded both for having numerous pass involvements and for performing influential passes as seen from the differing scores. This is true for all teams in Eliteserien.

Even though he is one of the most involved players in Rosenborg, the betweenness score for Reginiussen is zero, which at first glance seems to be counter-intuitive. A possible explanation for this could be that he tends to attempt easier passes than other players and for this reason is not achieving weights in the graph that are large enough to be part of someone's shortest paths. By examining the average difficulty of Reginiussen's pass involvements, the explanation is supported. His average pass difficulty, when considering only passes that are included in the network, was 0.053, while for Hedenstad, for instance, this value was 0.135. When including all n pass involvements as given in the tables, the average difficulty of passes increases for both players. This is a natural development as the extra included passes were unsuccessful and thus potentially more difficult to make. However, Reginiussen's pass involvements still had a considerably lower average difficulty.

The top three most important Rosenborg players by the PageRank recipient score are all offensive players, which is also the trend observed across all teams in Eliteserien. This is not unexpected as these players tend to receive more difficult passes due to their location on the field, which will give higher weights on the edges directed towards them. These players are also popular targets as they usually create more goal-scoring opportunities. For the PageRank passer score, the tendency both overall and for Rosenborg players is that defenders have higher scores. Although they on average do not attempt the most difficult passes, the defenders might be seen as popular passers due to them completing more passes, many of which are in between the defenders themselves so that they might boost each others popularity.

The clustering coefficients are in general high for all players in Eliteserien, with players having few pass involvements being awarded with a coefficient of one. As seen from Table 6.3, none of the Rosenborg players who received a score of one made the cut-off, which would be true for all teams. With almost all players being connected on a team, the observation of consistently high scores is reasonable.

Network 2 and Network 3

Moving on to Network 2 and Network 3, the cut-offs used for pass involvements are 308 and 42 respectively, and the computed centrality measures for Rosenborg players in 2017 are presented in Table 6.4. Interestingly, although not being among the most important players and by far not making the cut-off for Network 1 or Network 2, Samuel Adegbenro seems to be an important offensive contributor to the team as he makes the cut-off for Network 3. Adegbenro joined Rosenborg mid-season and has thus fewer connections with his teammates, which is probably the reason why he is not recognised by the other networks.

The PageRank recipient scores for Network 2 identify the full backs Meling and Hedenstad and the strikers Vilhjálmsson and Bendtner as the most important players. Strikers are not surprisingly scoring well here due to their position on the pitch. In general, passes they receive are more difficult to make, and thus also more challenging to follow up. For the more defensive players, high scores on this measure might indicate that they are tactically good contributors to their teams. Across the teams in Eliteserien, the player positions among the highest rated players are differing.

As mentioned before, defenders tend to make easy passes between each other. Thus, it is not surprising that they also obtain the highest scores for PageRank in terms of delivering passes that are easy to follow up due to the way the weights are

Table 6.4: The key Rosenborg players in terms of Network 2 and Network 3 in the 2017 season of Eliteserien. Only players involved in more passes than the equivalent of six matches according to each of the pass perspectives are considered. The three best players according to each measure are highlighted in bold text. The number of pass involvements, n, is the sum of tactical or effective passes made and received by a player. This count also includes passes that were unsuccessful and thus not part of the network analysis. The abbreviations for player positions are explained in Appendix D.

		Overview			*Effectiveness*		
Name	*Pos*	$PR_{\mathrm{R}}^{w2}(i)$	$PR_{\mathrm{P}}^{w2}(i)$	n	$PR_{\mathrm{R}}^{w3}(i)$	$PR_{\mathrm{P}}^{w3}(i)$	n
André Hansen	GK	0.398	0.402	659	0.179	0.465	50
Johan Lædre Bjørdal	CD	0.499	**0.909**	1260	0.520	0.723	106
Jacob Rasmussen	CD	0.624	**1.000**	982	0.479	0.795	70
Tore Reginiussen	CD	0.500	**0.844**	1801	0.710	0.849	154
Jørgen Skjelvik	CD	0.542	0.809	1536	0.551	0.696	125
Alex Gersbach	FB	0.527	0.542	550	0.483	0.504	54
Vegar Hedenstad	FB	**0.770**	0.641	1699	0.745	**0.994**	234
Birger Meling	FB	**1.000**	0.677	1259	**0.938**	**1.000**	149
Anders Konradsen	DM	0.660	0.559	1039	0.883	0.718	151
Mike Jensen	CM	0.715	0.486	1138	0.617	**0.871**	211
Marius Lundemo	CM	0.557	0.615	918	0.578	0.570	94
Fredrik Midtsjø	CM	0.659	0.361	694	**0.944**	0.822	171
Anders Trondsen	CM	0.666	0.497	466	0.751	0.586	55
Samuel Adegbenro	WI				0.585	0.486	46
Pål André Helland	WI	0.471	0.273	449	0.786	0.623	144
Milan Jevtovic	WI	0.520	0.297	432	0.662	0.551	104
Nicklas Bendtner	ST	0.738	0.365	886	**1.000**	0.708	198
Matthías Vilhjálmsson	ST	**0.744**	0.319	380	0.779	0.599	95

defined in Network 2. Passes that are easier to make are often also easier to follow up, such that defenders will have high weights on the outgoing edges due to both higher values of p_2 and many pass attempts. In a way, this PageRank measure is thus not a good indicator of whom are the best players to spot opportunities and make tactically good passes. Consequently, the measure is not considered further beyond this chapter.

Although offensive players might be thought of as being more effective, and thus should be captured by the PageRank score for effectiveness in Network 3, this is not always the case for all teams. This is probably due to the way the weights are defined. By counting involvements and not accounting for difficulty in any way, players with more pass involvements will be considered as more important. Attacking players have fewer pass involvements, but are important in the offensive play through other involvements than those that are considered in the network. Shots are mostly attempted by the attackers, meaning that they might not have been part of the sequence leading up to the attempt itself other than having received the final pass. Hence, they could receive high PageRank recipient scores, but could in principal get low rankings for the passer's score. The PageRank scores for effectiveness are thus a way of identifying the most important players in terms of frequency of offensive pass involvements, and not a way of finding the overall offensive contributor in a team. In Eliteserien, offensive players tend to dominate the top lists in terms of being recipients, while for the PageRank passer score, the player positions vary.

For Rosenborg's case, the top three most important recipients according to Network 3 represent the main outfield player positions: defender, midfielder and attacker. Bendtner, the 2017 league top scorer, is found to be the most important player. When looking at the PageRank passer scores instead, offensive players, wingers included, are ranked lower. The high scores for Meling and Hedenstad support the rationale behind their closeness score; they seem to be important players in Rosenborg's offensive play.

6.3.3 Case Studies

Two case studies involving Rosenborg matches from the 2017 season are presented below. In Case I, Sandefjord played against Rosenborg, with Rosenborg having their season highest ball possession of 70%. Case II is a match between Viking and Rosenborg in which Rosenborg had their season lowest ball possession of 40% (Verdens Gang AS, 2018). Rosenborg's average ball possession in 2017 was 54.4% (WhoScored.com, 2018).

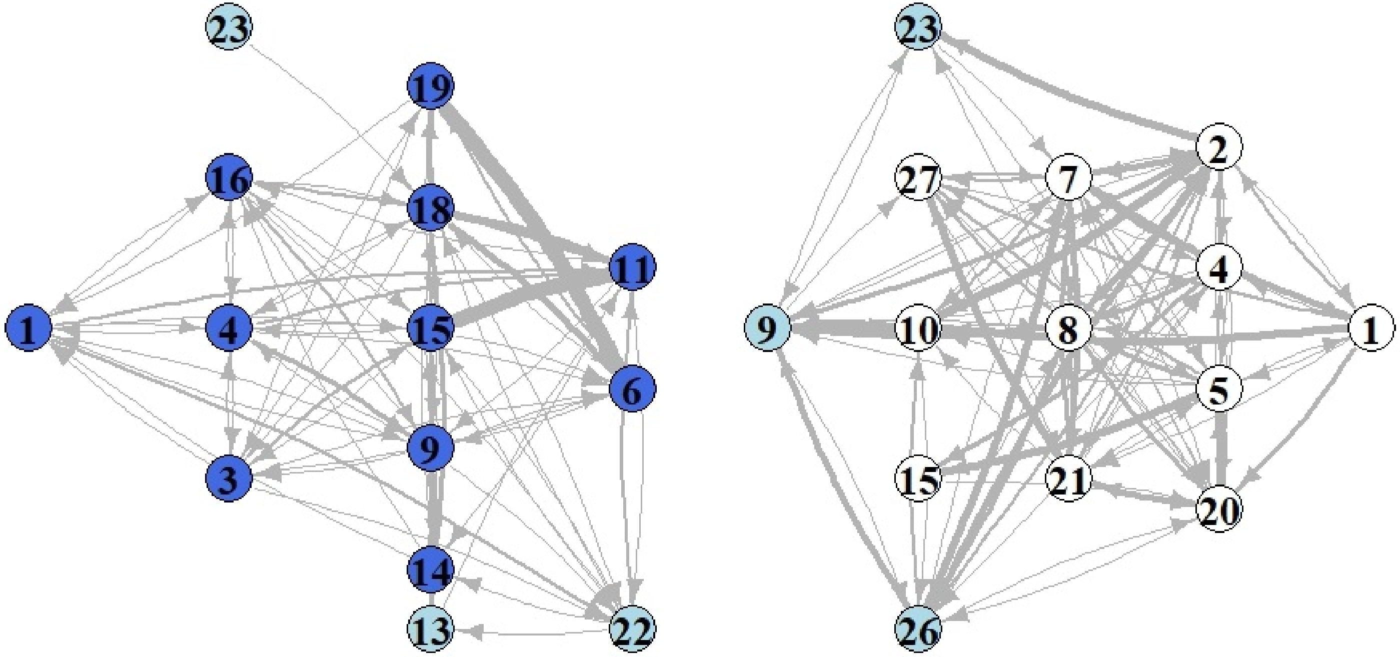
For both case studies, the teams' calculated network metrics are given together with a graphical representation of their passing networks for Network 1. In the networks, nodes represent the players on each team by their shirt number, and they are ordered by the starting formation of the team. The directed edges between nodes are weighted using the weights as described in Section 6.2.1. Hence, thicker lines indicate a stronger relationship between two players, with a stronger relationship being more passes between the players or, in general, higher difficulty on the passes observed between them.

Case I: Sandefjord versus Rosenborg

The match between Sandefjord and Rosenborg was played on the 5th of April 2017, with Rosenborg winning with three goals against zero. Graphical representations of the teams' passing networks are depicted in Figure 6.2, while the computed network metrics for players on both teams are tabulated in Table 6.5 and Table 6.6.

The much higher percentage of ball possession of Rosenborg is evident from the thicker directed edges between their players for Network 1 as shown in Figure 6.2. In general, the connections between Rosenborg players are stronger than that of Sandefjord players. For Rosenborg, there are no clear pattern in the passing network, while for Sandefjord, strong connections exist between offensive players on the left-hand side.

For Network 1, Rosenborg players' closeness scores are on average higher than the closeness scores of Sandefjord players, which is intuitive when taking the differences in ball possession into account. With high closeness scores, players are more easily reached and the ball is more readily kept within the team. Also, the



Panel A: Sandefjord (3-5-2)

Panel B: Rosenborg (4-3-3)

Figure 6.2: Graphical representations of the passing networks for the teams in Case I. Nodes are placed with respect to the starting eleven for each team, coloured based on the team's jersey colour and given names based on the players' jersey numbers. Substitutes are depicted with a separate colour (light blue). The directed edges are weighted based on the difficulty of passes between players, where the difficulty is decided by making use of the regression results from Model 1.

Table 6.5: The overall key figures according to Network 1 for both teams in Case I. The first column for each team indicate the players' jersey numbers.

Sandefjord							Rosenborg						
No.	*Name*	$C_{\text{C}}(i)$	$C_{\text{B}}(i)$	$PR_{\text{R}}^{w}(i)$	$PR_{\text{P}}^{w}(i)$	c_i^w	*No.*	*Name*	$C_{\text{C}}(i)'$	$C_{\text{B}}(i)$	$PR_{\text{R}}^{w1}(i)$	$PR_{\text{P}}^{w1}(i)$	c_i^w
1	I. Jónsson	0.220	0.000	0.091	0.356	0.765	1	A. Hansen	0.219	0.000	0.084	0.557	0.836
3	A. Seck	0.552	0.000	0.151	0.373	0.782	2	V. Hedenstad	0.961	0.769	0.688	0.679	0.667
4	C. Hansen	0.602	0.000	0.102	0.497	0.801	4	T. Reginiussen	0.618	0.308	0.123	0.760	0.665
6	P. Morer	0.799	0.177	0.701	0.349	0.773	5	J. Rasmussen	0.875	0.872	0.260	1.000	0.636
9	H. Storbæk	0.987	0.710	0.833	0.939	0.787	7	M. Jensen	0.955	0.436	0.901	0.393	0.742
11	F. Kastrati	0.991	0.774	1.000	0.380	0.812	8	A. Konradsen	1.000	0.923	0.713	0.648	0.803
13	M. Naglestad	0.712	0.000	0.222	0.089	1.000	9	N. Bendtner	0.774	0.051	0.688	0.191	0.712
14	E. Kebbie	0.857	0.032	0.466	0.426	0.971	10	M. Vilhjálmsson	0.795	0.154	1.000	0.161	0.795
15	E. Vallés	1.000	0.516	0.526	1.000	0.728	15	E. Rashani	0.666	0.026	0.148	0.115	1.000
16	E. Kane	0.868	0.210	0.560	0.525	0.942	20	A. Gersbach	0.934	1.000	0.666	0.642	0.664
18	W. Kurtovic	0.993	1.000	0.725	0.835	0.475	21	F. Midtsjø	0.936	0.359	0.747	0.286	0.874
19	V. Bindia	0.785	0.000	0.521	0.303	0.763	23	P. Helland	0.732	0.051	0.227	0.113	0.936
22	A. Sødlund	0.814	0.016	0.383	0.245	0.743	26	M. Jevtovic	0.944	0.692	0.841	0.323	0.894
23	M. Holt	0.394	0.000	0.068	0.109	NA	27	M. Bakenga	0.855	0.000	0.587	0.180	0.915

Table 6.6: The PageRank scores for Network 2 and Network 3 for players on both teams in Case I. The first column for each team indicates the players' jersey numbers. Players not participating in a passing sequence leading to a shot do not have PageRank scores for Network 3.

Sandefjord						Rosenborg					
No.	*Name*	$PR_{\text{R}}^{w2}(i)$	$PR_{\text{P}}^{w2}(i)$	$PR_{\text{R}}^{w3}(i)$	$PR_{\text{P}}^{w3}(i)$	*No.*	*Name*	$PR_{\text{R}}^{w2}(i)$	$PR_{\text{P}}^{w2}(i)$	$PR_{\text{R}}^{w3}(i)$	$PR_{\text{P}}^{w3}(i)$
1	I. Jónsson	0.303	0.327			1	A. Hansen	0.295	0.257		
3	A. Seck	0.213	0.313			2	V. Hedenstad	1.000	0.541	1.000	1.000
4	C. Hansen	0.320	0.497	0.389	0.826	4	T. Reginiussen	0.459	0.685	0.395	0.653
6	P. Morer	0.562	0.340	1.000	0.684	5	J. Rasmussen	0.756	1.000	0.196	0.254
9	H. Storbæk	0.861	0.768	0.144	0.536	7	M. Jensen	0.643	0.233	0.631	0.580
11	F. Kastrati	0.446	0.137	0.061	0.464	8	A. Konradsen	0.966	0.639	0.668	0.798
13	M. Naglestad	0.294	0.062	0.093	0.236	9	N. Bendtner	0.660	0.150	0.730	0.421
14	E. Kebbie	0.355	0.094			10	M. Vilhjálmsson	0.665	0.155	0.715	0.525
15	E. Vallés	0.852	1.000	0.061	1.000	15	E. Rashani	0.119	0.050	0.157	0.077
16	E. Kane	0.395	0.390	0.348	0.704	20	A. Gersbach	0.443	0.421		
18	W. Kurtovic	1.000	0.521	0.112	0.899	21	F. Midtsjø	0.479	0.167	0.460	0.642
19	V. Bindia	0.549	0.412	0.911	0.527	23	P. Helland	0.278	0.088	0.302	0.166
22	A. Sødlund	0.518	0.356			26	M. Jevtovic	0.513	0.138	0.577	0.386
23	M. Holt	0.090	0.086			27	M. Bakenga	0.345	0.124	0.452	0.627

betweenness scores of Rosenborg players are more evenly distributed across the team compared to the case of Sandefjord, which may indicate that Sandefjord is more dependent on certain players on the team to keep possession of the ball. These observations are somewhat supported by the teams' passing networks.

Considering the PageRank recipient scores for Network 1, the tendency is that offensive players achieve the highest scores, which is reasonable. Different patterns are seen for the PageRank passer scores for the two teams. In the case of Rosenborg, defensive players are achieving the highest scores, while central midfielders have the highest scores for Sandefjord. These differences could be due to the fact that both teams have a set playing style. In a 4-3-3 formation, Rosenborg wants to play out from the back and put pressure on the opponent team on the offensive half when the opponent team is established offensively and defensively respectively (Groven, 2017). For Sandefjord, the central midfielders are crucial players both in the attacking and defensive phases of the play in their 3-5-2 formation (Sandefjord Fotball, 2017).

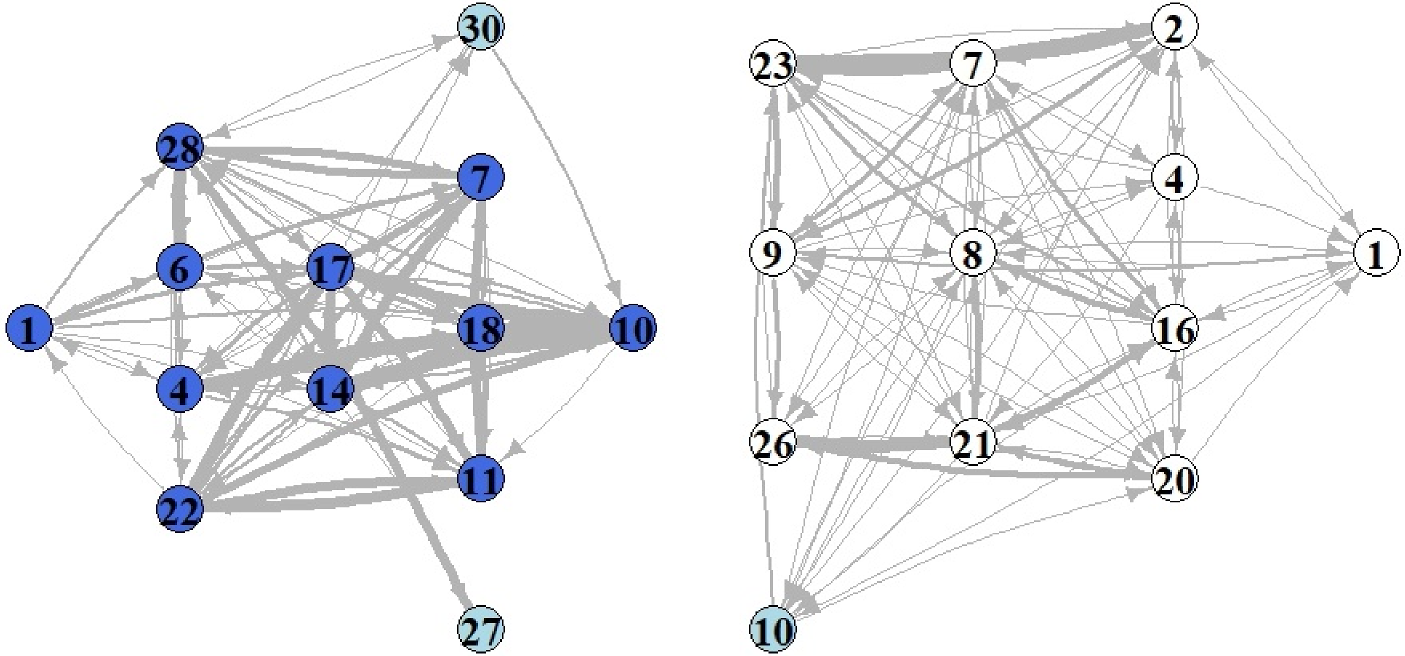
For both teams, the clustering coefficients are relatively high for all players. Rosenborg's average clustering coefficient for the match (0.780) is slightly higher than the team coefficient of Sandefjord (0.779) when considering only players who played at least 20% of the regular game time. These numbers are lower than the corresponding seasonal average team coefficients given in Appendix E. This might be due to the fact that the match was played at the very beginning of the season. At the start of a season, it has been a long period of time since most teams have played proper matches. Hence, the players might need some time to adjust to match situations and are thus not playing their best together just yet.

Moving on to the PageRank recipient and passer scores for Network 2, defensive players are recognised with the highest scores for Rosenborg, while central midfielders are the highest rated players in the case of Sandefjord. The importance of the players in these specific player positions seems to be related to the teams' playing styles as for the case of the PageRank passer measure for Network 1. For Network 3, the distribution of PageRank scores indicates well which players are more involved when shots are attempted as players that have not been involved in sequences leading to shots do not receive a score. The results are intuitive and the passer scores also support the findings that defenders and midfielders seem to be important players in the attacking play for Rosenborg and Sandefjord respectively.

Case II: Viking versus Rosenborg

The match between Viking and Rosenborg was played on April 17th 2017, and Rosenborg won with the final score 0-1. In Figure 6.3, a graphical representation of each team's passing network for Network 1 is shown, and in Table 6.7 and Table 6.8, the calculated network metrics are presented.

As expected, when taking the difference in ball possession into account, the strength of the connections between Viking players is stronger than for the case of Rosenborg players in the passing networks. Interestingly, the connections between Rosenborg's full back, central midfielder and winger on the right-hand side stand out to be strong. Thus, it seems like Rosenborg is more dependent on the players



Panel A: Viking (4-2-3-1)

Panel B: Rosenborg (4-3-3)

Figure 6.3: Graphical representations of the passing networks for the teams in Case II. Nodes are placed with respect to the starting eleven for each team, coloured based on the team's jersey colour and given names based on the players' jersey numbers. Substitutes are depicted with a separate colour (light blue). The directed edges are weighted based on the difficulty of passes between players, where the difficulty is decided by making use of the regression results from Model 1.

Table 6.7: The overall key figures for both teams in Case II. The first column for each team indicate the players' jersey numbers.

		Viking							Rosenborg				
No.	*Name*	$C_{\mathrm{C}}(i)$	$C_{\mathrm{B}}(i)$	$PR_{\mathrm{R}}^{w}(i)$	$PR_{\mathrm{P}}^{w}(i)$	c_i^w	*No.*	*Name*	$C_{\mathrm{C}}(i)'$	$C_{\mathrm{B}}(i)$	$PR_{\mathrm{R}}^{w1}(i)$	$PR_{\mathrm{P}}^{w1}(i)$	c_i^w
1	I. Austbø	0.264	0.000	0.085	0.644	0.884	1	A. Hansen	0.208	0.000	0.081	0.405	0.915
4	M. Ledger	0.604	0.233	0.146	0.829	0.818	2	V. Hedenstad	1.000	1.000	0.417	1.000	0.881
6	K. Mets	0.667	0.000	0.138	0.736	0.982	4	T. Reginiussen	0.766	0.000	0.140	0.709	0.886
7	S. Adegbenro	0.958	0.488	0.645	0.510	0.913	7	M. Jensen	0.989	0.857	0.642	0.811	0.843
10	K. Appiah	0.912	0.442	0.635	0.441	0.897	8	A. Konradsen	0.943	0.393	0.503	0.932	0.842
11	Z. Bytyqi	0.971	0.744	0.606	0.656	0.951	9	N. Bendtner	0.901	0.357	0.896	0.317	0.763
14	A. Danielsen	0.918	0.093	0.440	1.000	0.742	10	M. Vilhjálmsson	0.681	0.000	0.143	0.284	0.823
17	S. Ernemann	1.000	1.000	1.000	0.665	0.827	16	J. Skjelvik	0.835	0.000	0.201	0.986	0.918
18	J. Ryerson	0.751	0.070	0.273	0.752	0.847	20	A. Gersbach	0.874	0.107	0.590	0.453	0.780
22	C. Kronberg	0.900	0.140	0.454	0.740	0.873	21	F. Midtsjø	0.997	0.571	0.688	0.672	0.777
27	M. Bringaker	0.472	0.000	0.291	0.091	1.000	23	P. Helland	0.963	0.357	1.000	0.349	0.858
28	K. Haugen	0.924	0.512	0.649	0.647	0.610	26	M. Jevtovic	0.894	0.214	0.957	0.270	0.936
30	S. Michalsen	0.481	0.000	0.125	0.208	1.000							

Table 6.8: The PageRank scores for Network 2 and Network 3 for players on both teams in Case II. The first column for each team indicate the players' jersey numbers. Players not participating in a passing sequence leading to a shot do not have PageRank scores for Network 3.

		Viking						Rosenborg			
No.	*Name*	$PR_{\mathrm{R}}^{w2}(i)$	$PR_{\mathrm{P}}^{w2}(i)$	$PR_{\mathrm{R}}^{w3}(i)$	$PR_{\mathrm{P}}^{w3}(i)$	*No.*	*Name*	$PR_{\mathrm{R}}^{w2}(i)$	$PR_{\mathrm{P}}^{w2}(i)$	$PR_{\mathrm{R}}^{w3}(i)$	$PR_{\mathrm{P}}^{w3}(i)$
1	I. Austbø	0.241	0.440			1	A. Hansen	0.377	0.310		
4	M. Ledger	0.351	0.529	0.432	0.473	2	V. Hedenstad	0.929	0.626	0.380	0.751
6	K. Mets	0.352	0.834	0.265	0.320	4	T. Reginiussen	0.520	0.620		
7	S. Adegbenro	0.527	0.327	0.426	0.374	7	M. Jensen	1.000	0.469	0.547	0.531
10	K. Appiah	0.673	0.329	0.374	0.301	8	A. Konradsen	0.870	0.777	0.337	0.446
11	Z. Bytyqi	0.673	0.352	0.145	1.000	9	N. Bendtner	0.730	0.477	0.587	1.000
14	A. Danielsen	0.754	1.000	0.081	0.675	10	M. Vilhjálmsson	0.294	0.202	0.484	0.953
17	S. Ernemann	1.000	0.734	0.603	0.320	16	J. Skjelvik	0.771	1.000		
18	J. Ryerson	0.533	0.777	0.081	0.265	20	A. Gersbach	0.663	0.489	0.341	0.596
22	C. Kronberg	0.650	0.523	0.143	0.828	21	F. Midtsjø	0.828	0.503	0.724	0.841
27	M. Bringaker	0.152	0.076	0.388	0.138	23	P. Helland	0.897	0.312	1.000	0.089
28	K. Haugen	0.540	0.554	1.000	0.643	26	M. Jevtovic	0.688	0.263	0.425	0.107
30	S. Michalsen	0.149	0.112	0.365	0.339						

on this side when their ball possession is low compared to the case when they are dominating the play, as they did in Case I.

Considering the closeness scores of Rosenborg players in Network 1, many of the same top rated players as for Case I are recognised. Thus, these players seem to be central for Rosenborg independent of the game development and the distribution of ball possession between the playing teams. Moreover, the players achieving the highest closeness scores are also achieving high betweenness scores. Thus, these players are easy to reach and have a high involvement in the game. Compared to Case I, more Rosenborg players have a betweenness score of zero. It seems like higher ball possession is related to having fewer betweenness scores of zero, which intuitively makes sense as a low betweenness score is associated with less involvement in the match.

Similar to Case I, Rosenborg's offensive players dominate the highest ratings according to the PageRank recipient scores for Network 1, while more defensive players achieve the highest PageRank passer scores. For Viking however, the PageRank recipient scores tend to be higher for players playing on the left-hand side of the pitch, while the PageRank passer scores are higher for players playing in positions central on the field. Viking's clustering coefficient of 0.849 is higher than their average coefficient for the entire season which is tabulated in Appendix E, but it is slightly lower than Rosenborg's coefficient from the match (0.852). Although Viking performed poorly in 2017 and was relegated from Eliteserien, they apparently played a good match against Rosenborg, which might explain the difference in Viking's clustering coefficients.

The importance of the players on Rosenborg's right-hand side is seemingly captured by the PageRank recipient scores for Network 2. For Viking, two midfielders are achieving the highest ratings for this measure. Considering the PageRank passer scores for Network 2, the same patterns as seen for Network 1 are present. In general, the highest PageRank scores for Rosenborg players in Network 3 are dominated by offensive players, while more defensive players have high scores for Viking. Compared to Case I, where more defensive players on Rosenborg received higher PageRank passer scores, it seems like the defenders have had a lower offensive contribution in this match. As these players appear to be important for Rosenborg, this might explain the team's lower ball possession in the game.

6.3.4 Comparison with Existing Literature

In this thesis, the key players on teams in Eliteserien are found through network analyses with the majority of the edge weights being based on the results from the predicted probabilities of passes' success in Chapter 5. The idea of using pass difficulties as weights instead of the number of successful passes between players was introduced by McHale and Relton (2018). They only consider the accuracy of passes however, whereas two more aspects explaining the success of passes are investigated in separate networks here. Also, different types of network metrics are considered, making comparisons of the results difficult. In general, fair comparisons of results across leagues are not possible due to different teams, players and playing styles. However, trends seen for player positions may give some insights into which similarities that might be comparable. It is observed that McHale and Relton (2018) have identified mostly attacking players and midfielders in the top lists for teams playing in the English Premier League when calculating the exponential centrality, something which is also found for the closeness scores of Eliteserien players in Network 1.

Using the number of passes between players as weights, Pena and Touchette (2012) evaluated players' individual contributions to teams by calculating the closeness, betweenness and PageRank centrality measures and the Onnela clustering coefficient. In general, high clustering coefficients are observed across the teams, an observation that is supported by the results for Network 1. Also, the betweenness scores seem to vary a lot, with no clear patterns apparent for the different player positions. Other than this, no clear similarities are observed, which could be due to the fact that the scores given in Pena and Touchette (2012) are based on players' performance from one single match, giving a slightly poor basis of comparison between the two studies.

Rojas-Mora et al. (2017) investigated three matches from the group stage of Copa America and calculated the PageRank scores for all players on the field. Although the comparison is made on limited data, the PageRank recipient scores for Network 1 and the PageRank scores for players in Copa America both indicate that players playing in more offensive positions on the pitch are more important to their teams.

By only considering offensive sequences that ended with shots in their network analysis, Peixoto et al. (2017) had a similar approach to what was done for Network 3. The indegree and outdegree centrality measures, which are linked to passes received and passes made respectively, were calculated and revealed that strikers and midfielder scored highest on the respective measures. The PageRank scores for Network 3 are likewise linked to either the passer or the recipient of a pass, and the PageRank recipient scores support the finding that offensive players are more important, whereas the PageRank passer scores did not have a consistent pattern of player positions among the key players.

The network analyses done in this thesis indicate that the formation of a team can reveal information about which of the team's players the play is centred around. By using such information in the pre-game analyses, a team can accommodate their game plan according to the strengths and weaknesses of the opponent team. As the key players found for the teams not only are based on the number of passes between players, but also the difficulty of the passes, the opponent players that make the smartest passing alternatives may be identified and actions to stop them can be taken.

6.3.5 Evaluation of Research Question 2

As introduced in Chapter 1, the second main research question to be answered in this section is:

> **RQ 2:** How can the key players in a football team be identified?

The key players in a football team can be identified through the use of a network analysis. For such purposes, players are represented by nodes and the edges connecting them represent the interactions between the players. The edges can be weighted in accordance with a predetermined criterion. For this thesis, both the number of passes between players and the average predicted probabilities for the passes obtained by running GAMMs are used as weights on the directed edges between players in three different networks. Each network type handles a different aspect of a pass's success.

After the passing network is set up, different network metrics can be computed in order to find the key players of a team in terms of their influence and importance. Here, the closeness, betweenness and PageRank centrality measures and the Barrat clustering coefficient are considered. These measures can, for instance, provide information about which players are more involved in the play, which players who tend to be most popular and how well-balanced a team is. Such information would be beneficial for a team when deciding upon the game plan. Overall, the network analyses performed reveal intuitive results across the teams in Eliteserien, with the perceived most important players receiving higher scores. The PageRank passer scores for Network 2 proved to be unreliable however.

Two sub-questions were added to the second research question, both of them concerning Rosenborg. The first sub-question is: Who were the key Rosenborg players in the 2017 season of Eliteserien? Different network metrics are computed to answer this and they all reveal intuitive information, especially when taking the characteristics of the team's formation into account. For Network 1, three players stand out by achieving consistently high scores across the network metrics. These are the two full backs Vegar Hedenstad and Birger Meling and the winger Milan Jevtovic. Seemingly, Rosenborg's full backs play a key role in the team's offensive play. This observation is also supported by the results from Network 2 and Network 3. Further, the two strikers Nicklas Bendtner and Matthías Vilhjálmsson seem to be key players for Rosenborg in terms of receiving passes that are tactically good and effective.

The second sub-question is: How does Rosenborg depend upon certain key players when having differing ball possessions? To address this question, two matches where Rosenborg had a high and a low ball possession were analysed. When comparing the passing networks from the matches, it is evident that Rosenborg seems to be more dependent on the players playing on the right-hand side when the team struggles with keeping possession of the ball.

Chapter 7

Motif Analysis

In this chapter, the set-up for an analysis of passing motifs is explained, and the results are presented and discussed. Further, the distribution of the motif types used by the teams in Eliteserien is discussed in the context of the regression results. In the end, the results from the motif analysis are compared to similar existing studies and the third research question is addressed.

7.1 Model Set-Up

Passing motifs are sub-sequences of passes and they provide a method for discovering patterns in teams' passing behaviour. In this thesis, a GAM is developed to investigate the influence of different explanatory variables on the effectiveness of motifs in terms of generating shots. The motifs considered have a size of four, i.e. they consist of four players and three passes, which Meza (2017) concluded to be optimal. The passes in a motif 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 five seconds after the previous event (defined as X_8 in Section 5.1.2).

The analysis was performed using the *mgcv* package in RStudio (Wood, 2011), and the model was built on data from the 2014-2017 seasons in Eliteserien, with a total of 203,313 four-sized motifs being included. Fixed-effect variables are selected through the use of a Wald test, and the smooth terms are not tested as fixed effects.

7.1.1 Variables

To analyse the passing motifs in Eliteserien, a GAM is built with all continuous variables treated as smooth terms. The dependent variable is binary and takes the value of one if a motif leads to a shot and zero otherwise, which is similar to the previously developed Model 3. Hence, the effectiveness of motifs is considered. Most of the model's explanatory variables are based on the results from the passing ability models in Chapter 5 and the network analyses of key players in Chapter 6. Table 7.1 contains a summary of the explanatory variables.

The average predicted probabilities from Model 1, Model 2 and Model 3 in Chapter 5 are separately added up for all passes in a motif to serve as smooth functions in the passing motif model. Hence, the number of motifs for analysis is slightly reduced as only passes with applicable outcome according to all three models are considered. The aims of including these smooth terms are to test how the difficulty of the passes, the game overview of the players involved and the effectiveness of the passes influence the effectiveness of motifs.

The network metrics obtained for players in Chapter 6 are used to test whether the involvement of key players in a motif has an effect on its outcome. All but one of the previously considered metrics are included as explanatory variables in the analysis. The PageRank passer centrality for Network 2 is excluded as this measure was identified as a bad measure of key players in Section 6.3.2. The closeness and betweenness centrality measures and the clustering coefficients are summed for all four players involved in a motif, while the PageRank measures are summed for all passers or all recipients in the motif, depending upon which perspective the PageRank measure has.

Table 7.1: 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*
$\hat{Y}_i$	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
$PageRankRecipient_i$	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
$Clustering$	The sum of the Barrat clustering coefficients of all four players	C
$MotifType$	Indication of the motif type	F
$Zones$	The number of unique zones in which the motif takes place	C

A categorical variable is added to identify the motif type. With a motif size of four, five different motif types are possible as seen in Figure 2.4. The reference is chosen to be the ABCD motif, i.e. with four unique players being involved. 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 zones are defined as in Figure 5.1, and a total of six zones may be involved: 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 five-seconds time frame.

7.2 Results

In this section, the resulting passing motif model is discussed. First, the model is validated and the regression results are presented. Then, the distribution of four-sized motifs for teams in Eliteserien is analysed in the context of the results from the regression model. Finally, the model and its characteristics are compared to existing literature, and the third research question is addressed.

7.2.1 Model Validation

The validation of the model is performed using the tools introduced in Section 2.6. From Figure 7.1, the AUC for the ROC curve indicates an acceptable fit according to the guidelines in Table 2.1. However, the PR curve gives a low AUC due to a highly skewed data distribution as also was the case for Model 3 (see Section 5.2.2). Actually, both curves resemble the curves for Model 3. Similarly, the points at which the PR curve has to start and end are not giving much room for a high

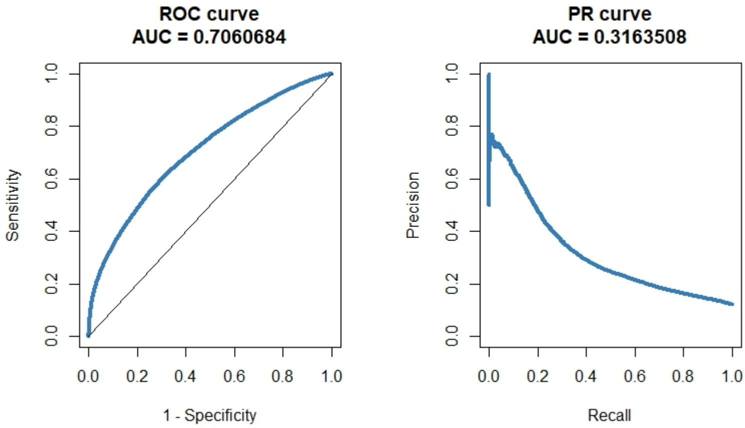


Figure 7.1: ROC and PR curves for the passing motif model.

AUC, which makes the value obtained seem appropriate. A HL test has also been performed. For both sample sizes of 1000 and 5000 observations, the test indicates a good fit of the model as less than ten of the 100 random samples resulted in rejections of the null hypothesis at a significance level of 5%.

7.2.2 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 models explored in Chapter 5. The regression results are shown in Table 7.2. A significance level of 10% is used, and negative values of the fixed-effect coefficients imply a reduced probability of success for the corresponding variables.

Fixed Terms

In the final model, two of the motif types have been removed in a Wald test. Hence, three motif types form the reference for the variable. The initially chosen reference ($ABCD$) is characterised by having four distinct players involved in the motif. With such a pattern, it is more likely that the motif covers a larger area of the pitch. Motif type $ABCA$, which is added to the reference, is similar to $ABCD$ as only the first and last involved player is the same. Consequently, this motif type is also more likely to cover larger areas. 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 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 types seem to be less likely of being effective in terms of resulting in shots.

Smooth Terms

The resulting smooth functions from the motif analysis are displayed in Figure 7.2. 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 in order, 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. Similar reasonable results are also present for the game overview predictions when looking at the region with acceptable confidence interval. 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. Intuitively,

Table 7.2: 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	$-0.503^{***}(0.067)$
Note: *p<0.05; **p<0.01; ***p<0.001	

Smooth terms	
Variable	*Sign.*
Y_1	***
Y_2	***
Y_3	***
$Closeness$	*
$Betweenness$	*
$PageRankRecipient_1$	***
$PageRankPasser_1$	***
$PageRankRecipient_2$	***
$PageRankRecipient_3$	***
$PageRankPasser_3$	***
$Clustering$	***
$Zones$	***

with many passes likely of being effective in a motif, their effectiveness should be positively related to the motif's effectiveness. This belief is captured by the model as seen from 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. As it was revealed that players in more offensive player positions received higher scores on this measure in Section 6.3.2, the negative slope may indicate that defensive or central players more often are 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 parabolic 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. From the betweenness scores calculated for Eliteserien in Section 6.3.2, it was observed that the scores varied among player positions and that players could receive a score of zero even if they had a considerably high number of pass involvements compared to other players on their team. Some of the players who received low scores however, Rosenborg's Reginiussen for instance, did receive high scores for

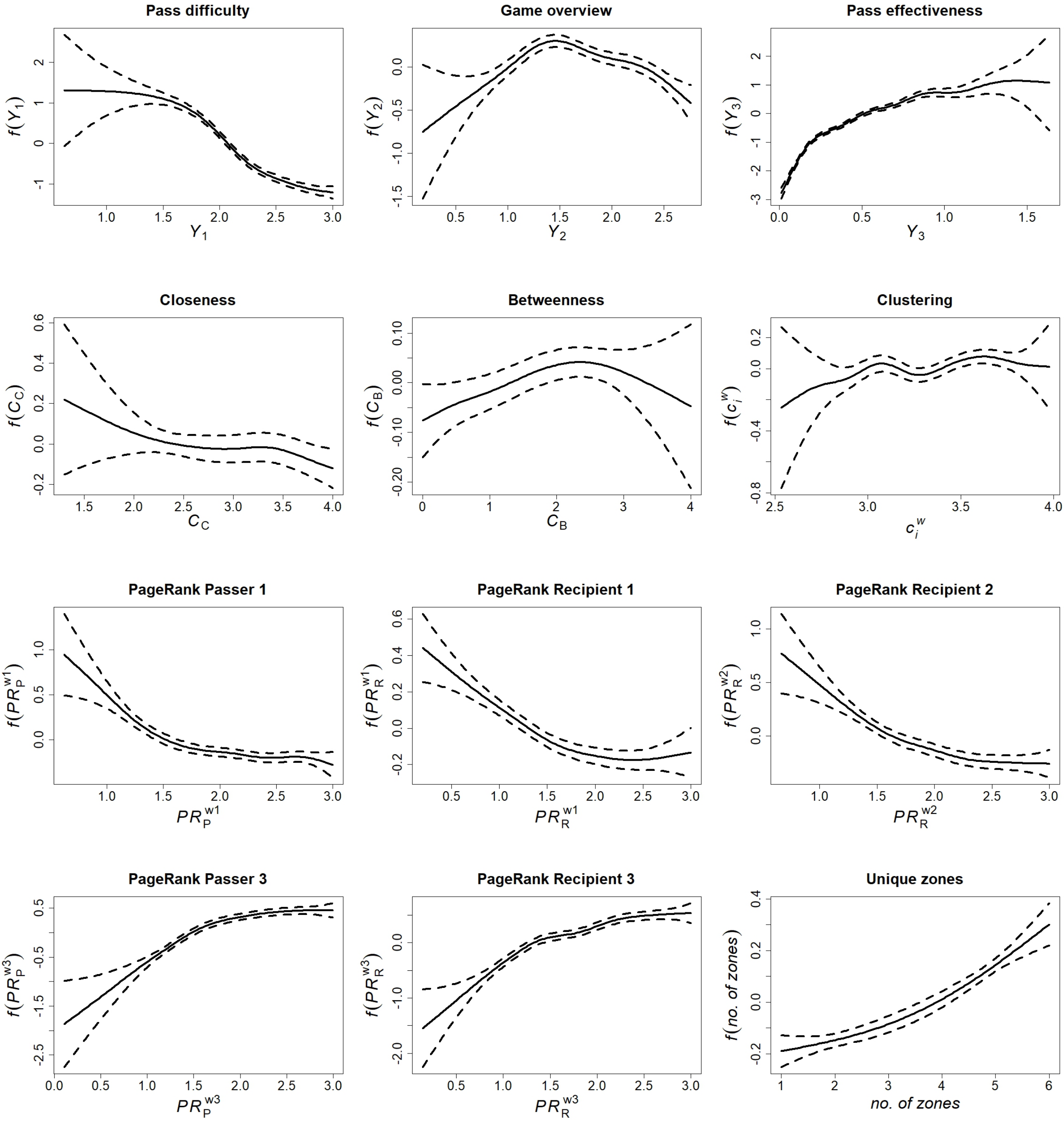


Figure 7.2: The resulting smooth functions from the motif analysis. The dotted lines indicate the 95% confidence intervals of the functions.

the PageRank effectiveness measures. Hence, players that are effective do not necessarily 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 and the PageRank recipient centrality for Network 2, the probability of success decreases with higher sums of the measures. Surprisingly, the involvement of key players in terms of these PageRank measures has less effect on motif effectiveness. Also, 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, the two graphs for the pass accuracy 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 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 utilising 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 five seconds 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.

7.2.3 Distribution of Motif Types in Eliteserien

The distribution of four-sized motifs for each team playing in Eliteserien 2017 is shown in Table 7.3. Additionally, the number of shots attempted per game is given for the teams. The aim of utilising 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 stands 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

Table 7.3: 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

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.

7.2.4 Comparison with Existing Literature

In this thesis, the effectiveness of motifs consisting of four players is analysed by including players' predicted probabilities of succeeding with the passes involved and players' contribution to their team in terms of importance and influence. The approach of using regression models to investigate passing motifs is seemingly new to the literature. Consequently, comparisons with existing literature is difficult.

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 utilised, 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 and Network 2 in this thesis turned out to have negative slopes too.

By using motifs of size four, Gyarmati et al. (2014) studied teams' playing styles in the Spanish La Liga. FC Barcelona, the league winners, stands out by using the three compact motif types more often than the other teams in the league. Interestingly, the team uses the motifs *ABCA* and *ABCD*, which were found to be more likely to be effective in this thesis, 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, something which is supported by the results from the motif analysis in this thesis.

Bekkers and Dabadghao (2017) investigated teams' playing styles by studying possession motifs and motifs resulting in a shot immediately after the last pass. Although the results are not directly comparable, some of the same patterns as can be seen for the passing motif model in this thesis are found with the more compact motif types being less likely to end in a shot.

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 this thesis are seemingly new to the literature on passing motifs. For instance, it is found that the more space utilised on the pitch during a motif, the more likely it is of being effective. Hence, counter-attacks seem to be proven 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.

7.2.5 Evaluation of Research Question 3

The third main research question is answered in this chapter:

> **RQ 3:** What determines the success of a passing motif in Eliteserien?

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. 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 likely of being 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. Easier passes in terms of accuracy and game overview decrease the probability of success whereas having more passes that are effective in a motif increase the likelihood for the motif to be effective.

The participation of key players in a motif has differing results for the different network metrics considered. However, the tendency is that for the metrics where offensive contribution already is taken into account, more intuitive effects on the effectiveness of motifs are present. Some of the network metrics are highly correlated, which might be the reason why some counter-intuitive results are present.

A sub-question handles the use of passing motifs in Eliteserien: Are the top performing teams in Eliteserien inclined to use the more effective motif types? As seen from the distribution of motif types used for each team in the 2017 season, all teams use the *ABCD* motif the most. However, there are no observed 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 which are equally affecting the outcome of a motif.

Chapter 8

Conclusion

In this thesis, the passing behaviour of football players in the Norwegian top division Eliteserien has been evaluated by answering three main research questions through regression and network analyses. The results obtained and their applications can provide coaches and players with valuable information that can be transferred to training sessions with the aim of increasing performance. A data set consisting of 749,859 passes and 203,313 motifs from 960 matches in the 2014-2017 seasons of Eliteserien was used in the analyses. However, the analyses done can easily be performed for any other league for which similar event-data is available, including all the top leagues in Europe. Also, they can be done for other sports where passing between players is essential.

Three generalised additive mixed models were developed to determine players' passing abilities and each handles a different aspect of a pass's success defined as accuracy, game overview and effectiveness. Game overview is an indirect term that is used to assess players' abilities to make tactically good passes. The AIC criterion was used to determine whether a variable should be treated as a smooth function or a fixed effect, and Wald tests were performed for variable elimination. In general, all models proved to have reasonably good fits and the signs of the coefficients of their fixed-effect variables made sense. Also, the shapes of the smooth functions were more or less intuitive.

A recurrent theme in many of the findings for the pass effectiveness model was that they might be explained as effects from counter-attacks. This indicates that teams can benefit from putting a low pressure on the opponents and awaiting counter-attack opportunities. The effect of the Elo rating found for opponent teams supports this thought when a match is played against one of the strongest teams in the league.

The seasonal effects on ground conditions seem to be captured by the models as the perceived difficulty of maintaining natural grass during the near-winter months was supported. Additionally, passes appear to be more easily made on artificial grass. From this, it is suggested that teams should choose a game plan depending upon the state of the surface the match is to be played on. Players should also carefully decide which types of passes that are likely to be successful on the different grass types. Passes along the ground might be easier to steer in the right direction when playing on artificial grass as the ground conditions are more predictable, while when playing on natural grass, it might be favourable to consider long passes in the air to avoid that varying ground conditions change the direction of the attempted pass.

The results from the passing ability models were further utilised in network analyses to identify the key players in teams. Three networks were considered and each handles one of the defined aspects of a pass's success. Both the number of passes between players and the predicted probabilities of the passes' success were used as weights in the networks. Close to all chosen network metrics revealed intuitive results, with the perceived most important and influential players receiving the highest scores. With such intuitive results, the networks may function as excellent tools for a team to make discoveries about opponent teams. When finding out which players that are more involved in the opponent's team, and also getting an indication about the quality of the passes they make, teams can alter their game plan to target opponent players that have more important involvements in the game.

For Rosenborg, the two full backs Vegar Hedenstad and Birger Meling were found to be key players when analysing the passing networks from the 2017 season. Seemingly, they play an important role in the team's offensive play. Further, Rosenborg seemed to be more dependent on the players playing on the right-hand side in matches where the team had a low ball possession. However, the left-back Meling had not yet made his debut for Rosenborg for the cases studied in this thesis, and considering his apparent importance for the team, he might be a valuable piece for regaining balance on the left-hand side.

To investigate the effectiveness of four-sized passing motifs, a generalised additive model, with results from the passing ability models and the network analyses included as variables, was considered. 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 scored 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 should consider looking into which motif types are the most effective and use them. This may provide the top-scoring teams with more goal-scoring opportunities that they can take advantage of.

Chapter 9

Recommendations for Further Research

The major limitation of the analyses performed in this thesis is that ball event-data is used, making it difficult to develop proxies that truly reflect opponent pressure. To address this, player tracking data should be utilised in further research. This will enable researchers to develop better proxies for the independent variables and make more tailored variable choices for the models. As Model 3 is considerably different from Model 1 and Model 2 in what the dependent variable handles, it would be reasonable to search for more specific variables to separate the models from each other.

With more detailed data than what is used in this thesis, it would be possible to define variables more accurately. It was, for instance, observed a counter-intuitive effect for tackles in the event prior to a pass, which most probably is due to the loose specification of the variable. Another aspect to consider for future work is to further investigate the degree to which the outcome of a pass is affected by the quality of the passer's team. As of now, the effect of a possible correlation between the player and team random-effect coefficients is unknown, and a better solution could be considered to avoid the potential issue.

When finding the key players in teams through network analyses, one of the measures proved to give unreasonable results. The PageRank passer scores for the game overview network favoured defenders and would not recognise the true tactical pass makers for the teams in Eliteserien. Hence, the weights used for this network should be reconsidered in order to get reliable results for the passing aspect that is explored.

For the structure of the networks, some changes can be made to investigate other features of the teams. Separate nodes can be added to the graphs to represent opponent players. With this approach, it is possible to extract information about which players have the most interactions with the opponent team or which players who most often regain ball possession. This would reveal useful information for coaches when deciding upon a team's playing strategy for a match. Moreover, by using the weights proposed in this thesis, it would be possible to identify the importance of the passes that are intercepted. Ideally, all passes made by a player can thus be utilised and not only those that successfully reach a teammate.

On an overall team level, it could be tested whether there is a connection between teams' clustering coefficients and their choice of passes. Does a team with a higher clustering coefficient tend to make better decisions regarding passing alternatives?

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 investigate whether passing motifs that are parts of counter-attacks are more likely to be effective. 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 on 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 considerable 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. This was tried in this thesis, but the vast number of potential combinations when having 21 zones and a motif size of four made the analysis too extensive. If finding a good way of dealing with the combinations however, 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 would 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 are found to be more effective in this thesis.

Bibliography

Allison, P. D. (2014). 'Measures of fit for logistic regression'. In: *Proceedings of the SAS Global Forum 2014 Conference.*

Arriaza-Ardiles, E., J. Martín-González, M. Zuniga, J. Sánchez-Flores, Y. de Saa and J. García-Manso (2018). 'Applying graphs and complex networks to football metric interpretation'. *Human Movement Science* 57, pp. 236–243.

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.* World Scientific, pp. 67–92.

Bartley, A. C. (2014). 'Evaluating goodness-of-fit for a logistic regression model using the Hosmer-Lemeshow test on samples from a large data set'. PhD thesis. The Ohio State University.

Bate, R. (1988). 'Football chance: tactics and strategy'. In: *Science and Football.* Ed. by T. Reilly, A. Lees, K. Davids and W. Murphy. London, UK: E. & F.N. Spon, pp. 293–301.

Bekkers, J. and S. Dabadghao (2017). 'Flow Motifs in Soccer: What can passing behavior tell us?' In: *MIT Sloan Sports Analytics Conference.*

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang (2006). 'Complex networks: Structure and dynamics'. *Physics Reports* 424.4-5, pp. 175–308.

Boor, S., C. Hanson and C. Ross (2018). *Rising stars.* `https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/deloitte-football-money-league.html`. Manchester, United Kingdom: Deloitte Sports Business Group.

Brandes, U. (2001). 'A faster algorithm for betweenness centrality'. *Journal of Mathematical Sociology* 25.2, pp. 163–177.

Bransen, L. (2017). 'Valuing passes in football using ball event data'. Master of Science. Erasmus University Rotterdam. URL: `http://hdl.handle.net/2105/41346`.

Brier, G. W. (1950). 'Verification of forecasts expressed in terms of probability'. *Monthey Weather Review* 78.1, pp. 1–3.

Brin, S. and L. Page (1998). 'The anatomy of a large-scale hypertextual web search engine'. *Computer Networks and ISDN Systems* 30.1-7, pp. 107–117.

Brooks, J., M. Kerr and J. Guttag (2016). 'Developing a data-driven player ranking in soccer using predictive model weights'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, pp. 49–55.

Clemente, F. M., F. M. L. Martins and R. S. Mendes (2015). 'There are differences between centrality levels of volleyball players in different competitive levels?' *Journal of Physical Education and Sport* 15.2, p. 272.

Clemente, F. M., F. M. L. Martins and R. S. Mendes (2016). *Social network analysis applied to team sports analysis.* Netherlands: Springer International Publishing.

ClubElo (2018). *API.* URL: `http://clubelo.com/API` (visited on 23/01/2018).

Coleman, B. J. (2012). 'Identifying the "players" in sports analytics research'. *Interfaces* 42.2, pp. 109–118.

Conte, D., T. Favero, M. Niederhausen, L. Capranica and A. Tessitore (2017). 'Determinants of the effectiveness of fast break actions in elite and sub-elite Italian men's basketball games'. *Biology of Sport* 34.2, p. 177.

Corbeil, R. R. and S. R. Searle (1976). 'Restricted maximum likelihood (REML) estimation of variance components in the mixed model'. *Technometrics* 18.1, pp. 31–38.

Courel Ibáñez, J. (2017). 'Tactical behaviour analysis in NBA basketball: Predictive study of use and effectiveness of players' actions and interactions during the inside pass'. PhD thesis. Universidad de Granada.

Courel, J., E. Suárez, E. Ortega, M. Piñar and D. Cárdenas (2013). 'Is the inside pass a performance indicator? Observational analysis of elite basketball teams'. *Revista de Psicología del Deporte* 22.1.

Csardi, G. and T. Nepusz (2006). 'The igraph software package for complex network research'. *InterJournal, Complex Systems* 1695.5, pp. 1–9.

Davis, J. and M. Goadrich (2006). 'The relationship between Precision-Recall and ROC curves'. In: *Proceedings of the 23rd international conference on Machine learning.* ACM, pp. 233–240.

Deb, S. and D. Dey (2017). 'Spatial modeling of shot conversion in soccer to single out goalscoring ability'. *arXiv preprint arXiv:1702.05662.*

Dey, P., M. Ganguly and S. Roy (2017). 'Network centrality based team formation: A case study on T-20 cricket'. *Applied Computing and Informatics* 13.2, pp. 161–168.

Dijkstra, E. W. (1959). 'A note on two problems in connexion with graphs'. *Numerische Mathematik* 1.1, pp. 269–271.

Duch, J., J. S. Waitzman and L. A. N. Amaral (2010). 'Quantifying the performance of individual players in a team activity'. *PloS One* 5.6, e10937.

Eliteserien (2018). *Stadioner.* URL: `https://www.eliteserien.no/stadioner` (visited on 22/01/2018).

Fawcett, T. (2006). 'An introduction to ROC analysis'. *Pattern Recognition Letters* 27.8, pp. 861–874.

Fewell, J. H., D. Armbruster, J. Ingraham, A. Petersen and J. S. Waters (2012). 'Basketball teams as strategic networks'. *PloS One* 7.11, e47445.

Freeman, L. C. (1977). 'A set of measures of centrality based on betweenness'. *Sociometry*, pp. 35–41.

Freeman, L. C. (1978). 'Centrality in social networks conceptual clarification'. *Social Networks* 1.3, pp. 215–239.

Frick, B. (2007). 'THE FOOTBALL PLAYERS' LABOR MARKET: EMPIRICAL EVIDENCE FROM THE MAJOR EUROPEAN LEAGUES'. *Scottish Journal of Political Economy* 54.3, pp. 422–446.

Fu, H.-H., D. K. Lin and H.-T. Tsai (2006). 'Damping factor in Google page ranking'. *Applied Stochastic Models in Business and Industry* 22.5-6, pp. 431–444.

Gama, J., P. Passos, K. Davids, H. Relvas, J. Ribeiro, V. Vaz and G. Dias (2014). 'Network analysis and intra-team activity in attacking phases of professional football'. *International Journal of Performance Analysis in Sport* 14.3, pp. 692–708.

Gelman, A. et al. (2005). 'Analysis of variance—why it is more important than ever'. *The Annals of Statistics* 33.1, pp. 1–53.

Gómez, M.-Á., J. Moral and C. Lago-Peñas (2015). 'Multivariate analysis of ball possessions effectiveness in elite futsal'. *Journal of Sports Sciences* 33.20, pp. 2173–2181.

Gómez, M.-A., A. Lorenzo, S.-J. Ibañez and J. Sampaio (2013). 'Ball possession effectiveness in men's and women's elite basketball according to situational variables in different game periods'. *Journal of Sports Sciences* 31.14, pp. 1578–1587.

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.

Goodenough, A. E., A. G. Hart and R. Stafford (2012). 'Regression with empirical variable selection: description of a new method and application to ecological datasets'. *PLoS One* 7.3, e34338.

Greig, F. (2017). *Dexterity, accuracy and FIFA skills – the people behind Opta's football data.* iNews. URL: `https://inews.co.uk/sport/football/opta-data-analyst-stats-football/` (visited on 09/03/2018).

Groven, M. H. (2017). *Analysen, Rosenborg-Lillestrøm: - Kan frustrere med destruktive grep og kynisk fotball.* Eurosport. URL: `https://www.eurosport.no/fotball/eliteserien/2017/analysen-rosenborg-lillestrom-kan-frustrere-med-destruktive-grep-og-kynisk-fotball_rbk_sto6171357/story.shtml` (visited on 10/04/2018).

Gyarmati, L., H. Kwak and P. Rodriguez (2014). 'Searching for a unique style in soccer'. *arXiv preprint arXiv:1409.0308.*

Gyarmati, L. and R. Stanojevic (2016). 'QPass: a Merit-based Evaluation of Soccer Passes'. *arXiv preprint arXiv:1608.03532.*

Haave, H. S. and H. Høiland (2017). 'Evaluating Association Football Player Performances Using Markov Models'. Master of Science. Norwegian University of Science and Technology.

Haave, H. and H. Høiland (2016). 'Analysing Corner Kicks in Norwegian Association Football'. Project Report. Norwegian University of Science and Technology.

Hastie, T. and R. Tibshirani (1986). 'Generalized Additive Models'. *Statistical Science* 1, pp. 297–318.

Hawkins, D. M. (2004). 'The problem of overfitting'. *Journal of Chemical Information and Computer Sciences* 44.1, pp. 1–12.

He, M., R. Cachucho and A. Knobbe (2015). 'Football player's performance and market value'. In: *Proceedings of the 2nd workshop of sports analytics, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD).*

Higham, D. G., W. G. Hopkins, D. B. Pyne and J. M. Anson (2014). 'Performance indicators related to points scoring and winning in international rugby sevens'. *Journal of Sports Science & Medicine* 13.2, p. 358.

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

Hughes, M. and I. Franks (2005). 'Analysis of passing sequences, shots and goals in soccer'. *Journal of Sports Sciences* 23.5, pp. 509–514.

Judd, C. M., G. H. McClelland and C. S. Ryan (2011). *Data analysis: A model comparison approach.* Abingdon, UK: Routledge.

Kang, B., M. Huh and S. Choi (2015). ‘Performance analysis of volleyball games using the social network and text mining techniques’. *Journal of the Korean Data and Information Science Society* 26.3, pp. 619–630.

Kharrat, T., J. L. Peña and I. McHale (2017). ‘Plus-Minus Player Ratings for Soccer’. *arXiv preprint arXiv:1706.04943*.

Kramer, A. A. and J. E. Zimmerman (2007). ‘Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited’. *Critical Care Medicine* 35.9, pp. 2052–2056.

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, pp. 381–400.

Macdonald, B., C. Weld and D. C. Arney (2013). ‘Quantifying playmaking ability in hockey’. *arXiv preprint arXiv:1307.6539*.

Machol, R. E. and S. P. Ladany (1976). *Management science in sports*. Vol. 4. Amsterdam, Netherlands: North-Holland.

Mackay, N. (2017). ‘Predicting goal probabilities for possessions in football’. Master of Science. Vrije Universiteit Amsterdam. URL: https://beta.%20vu.%20nl/nl/Images/werkstuk-mackay%5C_tcm235-849981.%20pdf.

Maher, M. J. (1978). ‘Optimal Strategies in Sports’. *Journal of the Operational Research Society* 29.6, pp. 619–620.

McCulloch, C. E., S. R. Searle and J. M. Neuhaus (2011). *Generalized, Linear, and Mixed Models*. Vol. 651. Hoboken, New Jersey: John Wiley & Sons.

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, pp. 339–347.

McHale, I. G., P. A. Scarf and D. E. Folker (2012). ‘On the development of a soccer player performance rating system for the English Premier League’. *Interfaces* 42.4, pp. 339–351.

McHale, I. G. and Ł. Szczepański (2014). ‘A mixed effects model for identifying goal scoring ability of footballers’. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 177.2, pp. 397–417.

Meza, D. A. P. (2017). ‘Flow Network Motifs Applied to Soccer Passing Data’. In: *Proceedings of MathSport International 2017 Conference*. Ed. by C. D. Francesco, L. D. Giovanni, M. Ferrante, G. Fonseca, F. Lisi and S. Pontarollo. Padova, Italy: Padova University Press, pp. 305–319.

Miller, D. L. (2017). *Why is the default smoothing method ”REML” rather than ”GCV.Cp”?* URL: https://github.com/DistanceDevelopment/dsm/wiki/Why-is-the-default-smoothing-method-%5C%22REML%5C%22-rather-than-%5C%22GCV.Cp%5C%22%5C%3F (visited on 06/02/2018).

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, pp. 824–827.

Mottley, C. M. (1954). 'Letter to the Editor—The Application of Operations-Research Methods to Athletic Games'. *Journal of the Operations Research Society of America* 2.3, pp. 335–338.

Müller, O., A. Simons and M. Weinmann (2017). 'Beyond crowd judgments: Data-driven estimation of market value in association football'. *European Journal of Operational Research* 263.2, pp. 611–624.

Opsahl, T., F. Agneessens and J. Skvoretz (2010). 'Node centrality in weighted networks: Generalizing degree and shortest paths'. *Social Networks* 32.3, pp. 245–251.

Page, L., S. Brin, R. Motwani and T. Winograd (1999). *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab.

Patterson, H. D. and R. Thompson (1971). 'Recovery of inter-block information when block sizes are unequal'. *Biometrika* 58.3, pp. 545–554.

Paul, P., M. L. Pennell and S. Lemeshow (2013). 'Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets'. *Statistics in Medicine* 32.1, pp. 67–80.

Peixoto, D., G. M. Praça, S. Bredt and F. M. Clemente (2017). 'Comparison of network processes between successful and unsuccessful offensive sequences in elite soccer'. *Human Movement* 18.5, pp. 48–54.

Peña, J. L. and R. S. Navarro (2015). 'Who can replace Xavi? A passing motif analysis of football players'. *arXiv preprint arXiv:1506.07768*.

Pena, J. L. and H. Touchette (2012). 'A network theory analysis of football strategies'. *arXiv preprint arXiv:1206.6904*.

Piette, J., S. Anand and L. Pham (2011). 'Evaluating basketball player performance via statistical network modeling'. In: *MIT Sloan Sports Analytics Conference.*

Piette, J., A. Braunstein, B. B. McShane and S. T. Jensen (2010). 'A point-mass mixture random effects model for pitching metrics'. *Journal of Quantitative Analysis in Sports* 6.3.

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, p. 1173.

Power, P., H. Ruiz, X. Wei and P. Lucey (2017). 'Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data'. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, pp. 1605–1613.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

rbk.no (2018). *Rosenborg Ballklub.* URL: `http://rbkmedia.no/statistikk/` (visited on 26/04/2018).

Reep, C. and B. Benjamin (1968). 'Skill and chance in association football'. *Journal of the Royal Statistical Society. Series A (General)* 131.4, pp. 581–585.

Refaeilzadeh, P., L. Tang and H. Liu (2009). 'Cross-validation'. In: *Encyclopedia of Database Systems.* Springer, pp. 532–538.

Rein, R., D. Raabe and D. Memmert (2017). '"Which pass is better?" Novel approaches to assess passing effectiveness in elite soccer'. *Human Movement Science* 55, pp. 172–181.

Rojas-Mora, J., F. Chávez-Bustamante, J. del Río-Andrade and N. Medina-Valdebenito (2017). 'A Methodology for the Analysis of Soccer Matches Based on PageRank Centrality'. In: *Sports Management as an Emerging Economic Activity.* Springer, pp. 257–272.

Sæbø, O. D. and L. M. Hvattum (2015). 'Evaluating the efficiency of the association football transfer market using regression based player ratings'. *Norsk Informatikkonferanse (NIK).*

Sáez Castillo, A., J. Rodríguez Avi and J. M. Pérez Sánchez (2013). 'Expected number of goals depending on intrinsic and extrinsic factors of a football player. An application to professional Spanish football league'. *European Journal of Sport Science* 13.2, pp. 127–138.

Sammonds, C. (2017). *Charles Reep: Football Analytics' Founding Father.* The Innovation Enterprise Ltd. URL: `https://channels.theinnovationenterprise.com/articles/charles-reep-football-analytics-founding-father` (visited on 26/04/2017).

Sandefjord Fotball (2017). *Sportsplan.* `https://drive.google.com/file/d/0B9wYsNKQFBUFMkRpejFDaFM3OFk/`. (accessed on 10/04/2018).

Statistics Solutions (2018). *Selection Process for Multiple Regression.* URL: `http://www.statisticssolutions.com/selection-process-for-multiple-regression/` (visited on 12/03/2018).

STATS LLC (2017). *AI and the Growing Use of Technology in Sport.* URL: `https://www.stats.com/industry-analysis-articles/ai-growing-use-technology-sport/` (visited on 11/04/2018).

Steinberg, L. (2015). *CHANGING THE GAME: The Rise of Sports Analytics.* Forbes Media LLC. URL: `https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/#6e02bce04c1f` (visited on 13/02/2018).

Szczepanski, L. (2015). ‘Assessing the skill of football players using statistical methods’. PhD thesis. University of Salford.

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, pp. 513–533.

The Guardian (2017). *The transfer hunters: how Premier League scouting set-ups compare.* URL: `https://www.theguardian.com/football/2017/jun/17/transfers-premier-league-scouting-recruitment` (visited on 03/04/2018).

Total Sportek (2007). *25 World’s Most Popular Sports (Ranked by 13 factors).* URL: `http://www.totalsportek.com/most-popular-sports/` (visited on 12/02/2018).

Tovar, J., A. Clavijo and J. Cárdenas (2017). *A strategy to predict association football players’ passing skills.* Tech. rep. UNIVERSIDAD DE LOS ANDES-CEDE.

UEFA (2017). *2017/2018 UEFA Europa League revenue distribution.* URL: `https://www.uefa.com/uefaeuropaleague/news/newsid=2493323.html#/` (visited on 13/02/2018).

Verdens Gang AS (2018). *VG LIVE.* URL: `https://vglive.no/` (visited on 04/04/2018).

Vicente-Vila, P. and C. Lago-Peñas (2016). ‘The goalkeeper influence on ball possession effectiveness in futsal’. *Journal of Human Kinetics* 51.1, pp. 217–224.

Weisstein, E. W. (2018). *Connected Digraph.* MathWorld—A Wolfram Web Resource. URL: `http://mathworld.wolfram.com/ConnectedDigraph.html` (visited on 15/03/2018).

WhoScored.com (2018). *WhoScored.com.* URL: `https://www.whoscored.com/` (visited on 25/04/2018).

Wiig, A. S. and E. M. Håland (2017). ‘Evaluating Passing Ability in Norwegian Association Football’. Project Report. Norwegian University of Science and Technology.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences.* Vol. 100. Oxford, UK: Academic Press.

Wood, S. (2018). *Random effects in GAMs.* R Documentation. URL: `https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/random.effects.html` (visited on 15/03/2018).

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, pp. 3–36.

Wood, S. N. (2017). *Smooth Models.* `https://people.maths.bris.ac.uk/~sw15190/mgcv/SmoothModels.pdf`. Lecture slides from the Swedish Winter Conference 2017 (accessed on 29/01/2018).

Wood, S. N., Y. Goude and S. Shaw (2015). ‘Generalized additive models for large data sets’. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.1, pp. 139–155.

Appendix A

Ground Surface Types in Eliteserien

Table A.1: An overview of types of grass used on football pitches in Eliteserien during the 2014-2017 seasons. Data was obtained from Eliteserien (2018).

Team	*Natural Grass*	*Artificial Grass*
Aalesund		✓
Bodø/Glimt		✓
Brann	✓	
Haugesund	✓	
Kristiansund		✓
Lillestrøm	✓	
Mjøndalen		✓
Molde		✓
Odd		✓
Rosenborg	✓	
Sandefjord	✓	
Sandnes Ulf	✓	
Sarpsborg 08		✓
Sogndal		✓
Stabæk	✓	
Start		✓
Strømsgodset		✓
Tromsø		✓
Viking	✓	
Vålerenga	✓	✓

Appendix B

Model Selection Results

The results from the selection of zone interactions are presented in Table B.1, and the results from the smooth term selection are presented in Table B.2.

Table B.1: Averaged Brier scores resulting from a 10-fold cross validation of the plain GLMM with differing number of zone interactions included. The total number of possible interactions was 352 after excluding the reference zone $z = 2$ and the interactions for which none observations were made. Interactions are added to the model in descending order of number of occurrences for the particular interaction. The second column, *Obs*, indicate the lowest number of observations made for an interaction included in the model. Only the numbers of interactions to be used, and not which ones, are tabulated.

No. of interactions	*Obs*	*Brier score*
10	9108	0.129250
30	4415	0.128802
50	3671	0.128412
75	2483	0.128321
100	1617	0.128112
150	685	0.127911
200	360	0.127654
250	132	0.127527
300	24	0.127489
352	1	0.127490

Table B.2: A summary of the AIC scores resulting from the stepwise selection of smooth terms. As the full GAMM obtained a lower AIC score than the plain GLMM, the stepwise selection was performed backwards. The smooth switched indicates which smooth term was replaced by a fixed effect in the given model estimation, where no switches imply a full GAMM and all imply a plain GLMM. Interaction terms are altered accordingly when the stand-alone variable is switched.

Step 1

Smooth switched	*AIC score*
None	428453
$f_1(x_{start}, y_{start}, x_{end}, y_{end})$	442698
$f_2(X_7)$	428095
$f_3(X_8)$	428655
$f_4(X_9)$	428467
$f_5(X_{10})$	428470
$f_6(X_{21})$	428461
$f_7(X_{24})$	428467
$f_8(\bar{x}, \bar{y})$	428519
$f_9(X_{27})$	428490
All	442860

Step 2 ($f_2(X_7)$ switched)

Smooth switched	*AIC score*
None	428095
$f_1(x_{start}, y_{start}, x_{end}, y_{end})$	442414
$f_3(X_8)$	428265
$f_4(X_9)$	428110
$f_5(X_{10})$	428113
$f_6(X_{21})$	428103
$f_7(X_{24})$	428157
$f_8(\bar{x}, \bar{y})$	428165
$f_9(X_{27})$	428129
All	442860

Appendix C

Regression Results

Table C.1: Model 1 regression results. The random-effect coefficients are not given here, they are presented in Section 5.2.3. Significance level is indicated by *.

Fixed effects	
Variable	*Coefficient(SE)*
$X_{7.2}$	$0.351^{***}(0.068)$
$X_{7.3}$	$0.498^{***}(0.014)$
$X_{7.4}$	$0.580^{***}(0.013)$
X_{11}	$0.343^{***}(0.015)$
X_{12}	$-0.100^{***}(0.024)$
X_{13}	$-0.327^{***}(0.024)$
$X_{14.1}$	$0.154^{**}(0.052)$
$X_{14.3}$	$0.757^{***}(0.043)$
X_{15}	$-0.311^{***}(0.016)$
X_{16}	$-1.227^{***}(0.014)$
X_{17}	$0.138^{***}(0.008)$
X_{19}	$0.184^{***}(0.045)$
X_{22}	$0.515^{***}(0.021)$
X_{23}	$0.123^{***}(0.011)$
$X_{25.1}$	$0.226^{***}(0.036)$
$X_{25.3}$	$-0.215^{\cdot}(0.122)$
$X_{7.2}{*}X_{18}$	$0.306^{***}(0.036)$
$X_{7.2}{*}X_{19}$	$0.225^{\cdot}(0.122)$
Intercept	$1.299^{***}(0.068)$

Smooth terms	
Variable	*Sign.*
$f_1(x_{start}, y_{start}, x_{end}, y_{end})$	***
$f_3(X_8)$	***
$f_4(X_9)$	
$f_5(X_{10})$	***
$f_6(X_{21})$	***
$f_7(X_{24}) : Artificial$	***
$f_7(X_{24}) : Natural$	***
$f_8(\bar{x}, \bar{y})$	***
$f_9(X_{27})$	***
$f_{10}(X_9, X_{10})$	***

Note: $^{\cdot}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table C.2: Model 2 regression results. The random-effect coefficients are not given here, they are presented in Section 5.2.3. Significance level is indicated by *.

Fixed effects	
Variable	*Coefficient (SE)*
$X_{7.2}$	$0.311^{***}(0.012)$
$X_{7.3}$	$0.471^{***}(0.011)$
$X_{7.4}$	$0.515^{***}(0.012)$
X_{11}	$0.189^{***}(0.029)$
X_{12}	$-0.596^{***}(0.062)$
X_{13}	$-0.148^{***}(0.021)$
$X_{14.1}$	$0.097^{*}(0.042)$
$X_{14.2}$	$0.532^{***}(0.066)$
$X_{14.3}$	$0.387^{***}(0.036)$
X_{15}	$-0.294^{***}(0.014)$
X_{16}	$-0.968^{***}(0.014)$
X_{17}	$0.117^{***}(0.006)$
X_{18}	$-0.181^{**}(0.057)$
X_{20}	$-0.172^{***}(0.050)$
X_{22}	$0.277^{***}(0.017)$
X_{23}	$0.187^{***}(0.009)$
$X_{25.1}$	$0.120^{*}(0.060)$
$X_{25.3}$	$-0.367^{***}(0.109)$
$X_{7.2}{}^{*}X_{18}$	$0.543^{***}(0.063)$
$X_{7.2}{}^{*}X_{19}$	$0.447^{***}(0.109)$
Intercept	$-0.517^{***}(0.085)$

Smooth terms	
Variable	*Sign.*
$f_1(x_{start}, y_{start}, x_{end}, y_{end})$	***
$f_3(X_8)$	***
$f_4(X_9)$	**
$f_5(X_{10})$	***
$f_6(X_{21})$	***
$f_7(X_{24}) : Artificial$	***
$f_7(X_{24}) : Natural$	***
$f_8(\bar{x}, \bar{y})$	***
$f_9(X_{27})$	***
$f_{10}(X_9, X_{10})$	***

Note: *p<0.05; **p<0.01; ***p<0.001

Table C.3: Model 3 regression results. The random-effect coefficients are not given here, they are presented in Section 5.2.3. Significance level is indicated by *.

Fixed effects	
Variable	*Coefficient*
$X_{7.2}$	$0.240^{***}(0.019)$
$X_{7.3}$	$0.223^{***}(0.018)$
$X_{7.4}$	$0.223^{***}(0.020)$
X_{11}	$0.325^{***}(0.036)$
X_{12}	$-0.134^{**}(0.050)$
X_{13}	$0.091^{*}(0.036)$
$X_{14.3}$	$0.338^{***}(0.059)$
X_{15}	$-0.110^{***}(0.025)$
X_{16}	$-0.785^{***}(0.026)$
X_{17}	$0.092^{***}(0.010)$
X_{19}	$-0.106^{***}(0.028)$
X_{20}	$0.363^{***}(0.080)$
X_{22}	$0.289^{***}(0.029)$
X_{23}	$0.127^{***}(0.014)$
$X_{25.1}$	$-0.156^{\cdot}(0.088)$
$X_{7.2}{*}X_{18}$	$-0.093^{*}(0.040)$
Intercept	$-3.248^{***}(0.118)$

Smooth terms	
Variable	*Sign.*
$f_1(x_{start}, y_{start}, x_{end}, y_{end})$	***
$f_3(X_8)$	***
$f_4(X_9)$	
$f_5(X_{10})$	***
$f_6(X_{21})$	***
$f_7(X_{24}) : Artificial$	***
$f_7(X_{24} : Natural)$	***
$f_8(\bar{x}, \bar{y})$	***
$f_9(X_{27})$	***
$f_{10}(X_9, X_{10})$	

Note: $^{\cdot}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Appendix D

Abbreviations

Table D.1: Team name abbreviations and the seasons of the 2014-2017 seasons of Eliteserien in which the teams have played.

Abbreviation	*Explanation*	*Seasons*
AaFK	Aalesunds FK	’14,’15,’16,’17
B/G	FK Bodø/Glimt	’14,’15,’16
BRA	SK Brann	’14,’16,’17
FKH	FK Haugesund	’14,’15,’16,’17
KBK	Kristiansund BK	’17
LSK	Lillestrøm SK	’14,’15,’16,’17
MJØ	Mjøndalen IF	’15
MOL	Molde FK	’14,’15,’16,’17
ODD	ODDs BK	’14,’15,’16,’17
RBK	Rosenborg BK	’14,’15,’16,’17
S08	Sarpsborg 08 FF	’14,’15,’16,’17
SAN	Sandefjord Fotball	’15,’17
SIF	Strømsgodset TF	’14,’15,’16,’17
SOG	Sogndal Fotball	’14,’16,’17
STA	IK Start	’14,’15,’16
STB	Stabæk Fotball	’14,’15,’16,’17
TIL	Tromsø IL	’15,’16,’17
ULF	Sandnes Ulf	’14
VIF	Vålerenga Fotball	’14,’15,’16,’17
VIK	Viking FK	’14,’15,’16,’17

Table D.2: Player position abbreviations.

Abbreviation	*Explanation*
AM	Attacking midfielder
CD	Central defender
CM	Central midfielder
DM	Defensive midfielder
FB	Full back
ST	Striker
WB	Wing back
WI	Winger

Appendix E

Team Clustering

Table E.1: The average team clustering coefficient for each team in the 2017 season of Eliteserien. Only players with pass involvements above the equivalent of six matches are considered for each team. For the entire league, this number corresponds to 364 pass involvements.

Team	$\bar{c}_i^w$
Aalesund	0.906
Brann	0.906
Odd	0.886
Rosenborg	0.883
Sarpsborg 08	0.864
Haugesund	0.860
Lillestrøm	0.854
Strømsgodset	0.847
Vålerenga	0.844
Tromsø	0.842
Molde	0.833
Sandefjord	0.824
Viking	0.807
Stabæk	0.805
Kristiansund	0.797
Sogndal	0.761