Bacheloroppgave

SCM600 Logistikk

Quantifying Footballing: Using Multiple Regression Analysis and ELO Ratings to Identify the Most Important KPI's for Goalkeepers

Mats Mørch

Totalt antall sider inkludert forsiden: 53

Molde, 2.6.20

Høgskolen i Molde
Vitenskapelig høgskole i logistikk

Obligatorisk egenerklæring/gruppeerklæring

Den enkelte student er selv ansvarlig for å sette seg inn i hva som er lovlige hjelpemidler, retningslinjer for bruk av disse og regler om kildebruk. Erklæringen skal bevisstgjøre studentene på deres ansvar og hvilke konsekvenser fusk kan medføre. Manglende erklæring fritar ikke studentene fra sitt ansvar.

Du/dere fyller ut erklæringen ved å klikke i ruten til høyre for den enkelte del 1-6:		
1.	**Jeg/vi erklærer herved at min/vår besvarelse er mitt/vårt eget arbeid, og at jeg/vi ikke har brukt andre kilder eller har mottatt annen hjelp enn det som er nevnt i besvarelsen.**	☒
2.	**Jeg/vi erklærer videre at denne besvarelsen:** • ikke har vært brukt til annen eksamen ved annen avdeling/universitet/høgskole innenlands eller utenlands. • ikke refererer til andres arbeid uten at det er oppgitt. • ikke refererer til eget tidligere arbeid uten at det er oppgitt. • har alle referansene oppgitt i litteraturlisten. • ikke er en kopi, duplikat eller avskrift av andres arbeid eller besvarelse.	☒
3.	**Jeg/vi er kjent med at brudd på ovennevnte er å betrakte som fusk og kan medføre annullering av eksamen og utestengelse fra universiteter og høgskoler i Norge, jf. Universitets- og høgskoleloven §§4-7 og 4-8 og Forskrift om eksamen §§14 og 15.**	☒
4.	**Jeg/vi er kjent med at alle innleverte oppgaver kan bli plagiatkontrollert i URKUND, se Retningslinjer for elektronisk innlevering og publisering av studiepoenggivende studentoppgaver**	☒
5.	**Jeg/vi er kjent med at høgskolen vil behandle alle saker hvor det forligger mistanke om fusk etter høgskolens retningslinjer for behandling av saker om fusk**	☒
6.	**Jeg/vi har satt oss inn i regler og retningslinjer i bruk av kilder og referanser på biblioteket sine nettsider**	☒

Personvern

Personopplysningsloven

Forskningsprosjekt som innebærer behandling av personopplysninger iht. Personopplysningsloven skal meldes til Norsk senter for forskningsdata, NSD, for vurdering.

Har oppgaven vært vurdert av NSD? ☐**ja** ☒**nei**

- Hvis ja:

Referansenummer:

- Hvis nei:

Jeg/vi erklærer at oppgaven ikke omfattes av Personopplysningsloven: ☒

Helseforskningsloven

Dersom prosjektet faller inn under Helseforskningsloven, skal det også søkes om forhåndsgodkjenning fra Regionale komiteer for medisinsk og helsefaglig forskningsetikk, REK, i din region.

Har oppgaven vært til behandling hos REK? ☐**ja** ☒**nei**

- Hvis ja:

Referansenummer:

Publiseringsavtale

Studiepoeng: 15

Veileder: Lars Magnus Hvattum

Fullmakt til elektronisk publisering av oppgaven

Forfatter(ne) har opphavsrett til oppgaven. Det betyr blant annet enerett til å gjøre verket tilgjengelig for allmennheten (Åndsverkloven. §2).

Alle oppgaver som fyller kriteriene vil bli registrert og publisert i Brage HiM med forfatter(ne)s godkjennelse.

Oppgaver som er unntatt offentlighet eller båndlagt vil ikke bli publisert.

Jeg/vi gir herved Høgskolen i Molde en vederlagsfri rett til å gjøre oppgaven tilgjengelig for elektronisk publisering:	☒**ja**	☐**nei**
Er oppgaven båndlagt (konfidensiell)?	☐**ja**	☒**nei**
(Båndleggingsavtale må fylles ut)		
- Hvis ja:		
Kan oppgaven publiseres når båndleggingsperioden er over?	☐**ja**	☐**nei**

Dato: 2.6.20

Table of Content

List of figures

List of tables

List of equations

1.0 Introduction

With the rising following of football-teams and the increasing amount of money poured into the sport, players are raised to higher pedestals of fame than ever before.
When the highlights from football matches are shown, it is always the attacking players who score or assist the goal who receives the recognition while the goalkeeper is left looking frustrated and punching the goal-frame.

The European football market is estimated to be worth €28.4 billion in 2018; therefore there is no lack of money for the big clubs, but what they are continually chasing is points in the league and winning games in cups (Deloitte 2019). This is what football-clubs and players live and die for: to get the ball into the opponent's net and stop them from scoring in our goal. With this understanding, it is easier to grasp why a team would pay over €200 million for a striker, because his main objective is to score goals. And those goals could translate into securing that important win in extra time that gives the team the needed points or further qualification. But it begs for the question of why the same thing is not right; that the goalkeeper's main objective is to make saves not to concede goals, and that those saves will prevent the club from dropping those crucial points in the domestic league or save that vital penalty in the penalty shoot-out in a big tournament?

The business of collecting event-data in football has recently been very lucrative, and several new companies have invested millions of euros in being the provider to gamblers, supporters, and football clubs (Biermann 2019). This means that there is a lot of innovation that can be leveraged by clubs to identify goalkeepers that will suit the clubs playing style and be of the quality searched after. One of the main constraints to this is that in today's industry, traditional scouts are doing the majority of the groundwork of identifying talent, which leads to biased views and the chance of the goalkeeper not being up to the preferred standard. These failed signings are the costliest mistakes clubs make (Biermann 2019). They will spend a lot to sign them, and if they fail to perform at the expected level, their market value significantly drops, leaving them to sell them for a fraction of what they acquired them for. To counter this, key performance indicators (KPIs) are identified, which quantifies the goalkeeper's performance in an objective and unbiased way. This will minimize the

possibility of signing players who are not at up to the expected standard, which will reduce costs for the club. Further, it can reduce the expenditure on scouts, and creating a more competitive team can also be the side effects of using such KPIs to identify potential new signings.

The remainder of this report is structured as follows. Chapter two contains relevant research on the topic of interest and shows the lack of research on the area. In section three, the methodology behind the analysis is presented, and in chapter four, the results are displayed. Chapter five discusses the results and how they can be interoperated and discusses the future impact this might have on the sport. It is all rounded off in chapter six, where the conclusion is put forward.

2.0 Literature review

Ever since Operational Research was used as an approach to aid the military, organisations have used it to improve their business processes and optimise their supply chain (Wright 2009). It occurred to some of them that the same methods could be used for other branches of organisations such as sport: "The possibility of applying the scientific method to the athletic games does not appear to have received much attention by the operation research community" (Mottley 1954).

By adhering to the framework proposed by Rose (2013), a definition of Sports Analytics is recommended as the following:

Sports Analytics refers to the use of data science and statistics to improve any area of a sports team or -organisation, by improving decision making.

The amount of raw data collected by sports teams around the world has recently skyrocketed, and the need to visualise, understand and turn that data into knowledge to act upon has become the focus of almost all sports teams (Hughes and Franks 2005; Mackenzie and Chusion 2013; Memmert and Raabe 2018).

2.1 Baseball

It was Michael Lewis' (2003) bestselling book "Moneyball" and the subsequent $120 million Hollywood production that made Sports Analytics a colloquial term (Hakes and Sauer 2006). The book follows the Oakland A's and Billy Beane's journey from a futile team to reaching the playoffs of the World Series in America with one of the lowest budgets in the league by exploiting market inefficacies (Hakes and Sauer 2006; Baumer and Zimbalist 2013). This was done by focusing on on-base percentage instead of batting average and stolen bases and punts. The rise of sports analytics can best be visualised by the fact that the 30 MLB clubs today employ more than 250 analysts, where most hold a Ph-D in mathematics, statistics, or IT (Biermann 2019). By focusing their player-recruitment and sporting-philosophy on these principles, they revolutionised Major League Baseball (MLB), and changed the game forever, even if they did not tell the story in the exact manner (Harvard Business Review 2012; Lewis 2003; Baumer and Zimbalist 2013).

Today, however, there is no significant advantage for MLB teams that focus a lot of resources on sports analytics specialists, simply because they all do. Every MLB team now employs numerous people to crunch the numbers and extract information to find actionable

data points. This information will then be passed onto specialist coaches who understand the quantitative side of baseball and have a proven record of how to increase the metrics for the players and create on-field success for the baseball team (Murphy 2019). Lastly, MLB has created its own database with all metrics collected for fans to engage with and has created a new value stream with fantasy baseball heavily relying on these statistics (Medium 2019).

2.2 Basketball

Since the introduction of the three-point field goal in basketball in 1979, it would take almost 40 years until the revolution started, and the changes have been dubbed "as the most influential gerrymander in sports history" (Frazier and Sachare 2004; Goldsberry 2019). By adding a new way of scoring points worth 50% more than the former, a new category of superstars would be created. Whereas in 2012, the league average was 18.7 attempts in a game for three-pointers, in 2017, it was 27 attempts (Merrimack College 2019; Goldsberry 2019; Bailey et al. 2018). Figure 1 and 2 shows this shift in a succinct fashion from a game where the most common shoot-position was scattered mainly inside of the three-point line hoop, while in the 2016/2017 season there was a clear tendency for shooters to either shoot from as close to the hoop as possible, or from the three-point line.

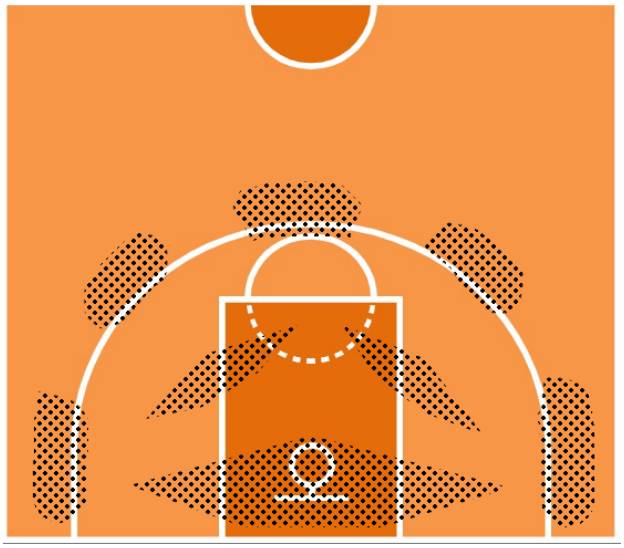

Figure 7 Most Common shot location in the NBA in 2001-2002 Adapted from Goldsberry (2019)

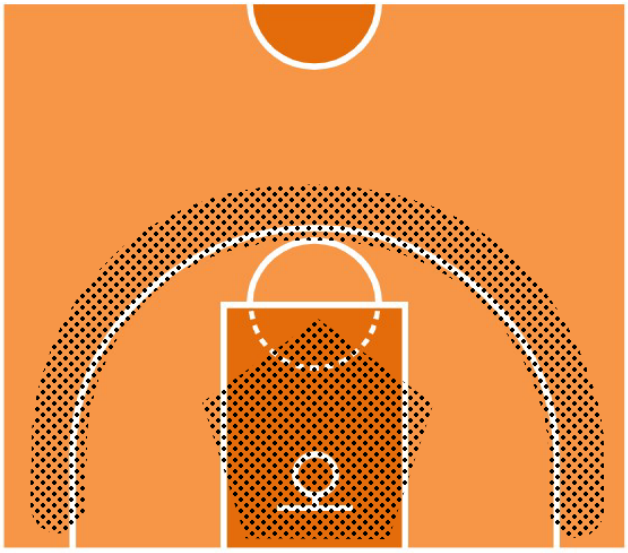

Figure 8 The Most common shot location in the NBA in 2016-2017. Adapted from Goldsberry (2019)

Goldsberry (2019) explains that the shift started in the early 2000s when the NBA began to collect tracking data and event data, meaning it collected "what happened" data and "where did it happen" data for all actions in the NBA. In the beginning, it did not affect the game, but when people within the Huston Rockets started mapping and visualising where shots were taken from, and calculating the average point return on shooting from different distances it became clear to them that some regions of the arena were worth than other areas. This became the backbone for the Houston Rockets under general manager Daryl Morey which is identified as the pioneer for identifying this ineffective in the game and utilising it to its fullest by winning the division titles several times in the last decade (Goldsberry 2019). This style entails setting up possession plays to free up players in certain areas on the three-point line and execute the shot with precision or find available space near the hoop for a player to receive a pass there. Houston Rockets took this tactic to the extreme when they in 2017 shot more than 50% of their shots from outside the three-point line (The Athletic 2017). This eventually left the fans frustrated and started critiquing the games as boring and predictable, even though they were the most winning team in the regular season (Harvard Business Review 2018).

Goldsberry (2019) further explains that one of the most recognisable differences is that the players like Dennis Rodman and Shaquille O'Neal would not have reached the levels of fame they now hold if they were to play the game today, with different skill sets required to be at the pinnacle of the sport today, compared to twenty years ago.

2.3 Football

Football's transformation into a numbers game started when the small British company named Prozone was founded in 1995 (Biermann 2019). By placing eight heat-sensitive cameras to Pride Park Stadium, home of Derby County FC, they were able to capture each player's location every 0.1 seconds and record over 3 000 touches of the ball for each game. In 2005 there were eight, and in 2017, 19 out of 20 Premier League clubs using their services in an attempt to quantify the game to gain an advantage over their opponents (Biermann 2019; Medeiros 2017).

Kuper and Szymanski (2009) found that only 16% of the variation in league position was explained by the expenditure on transfers, whereas spending on wages accounted for 92% of the variation in league position when they looked at the top two leagues in England. This would mean that the more you pay in wages, the higher you will finish in the league, the amount of money you pay in transfer fees does not affect. This is only one example of what Kuper and Szymanski call "systematic failures", others being scouts recommending blond players because they stand out more than the average hair colour of brown and that nationalities matter because of previous players from the same nationality performed at a high level.

Since the 2013/14 season, Sportec Solutions has overseen the recording, storing, and distributing all data for every single game in the Bundesliga and Bundesliga 2. This encompasses 36 teams and 396 games each season were they collect tracking- and event data for all players (Sportec Solutions 2020; Memmert and Raabe 2018). This data set is shared with all clubs, partners and academics to turn the vast amounts of raw data into useful information and knowledge. Further, this collaboration has been a breeding ground for academics, mathematicians, statisticians, and several other professions to try to develop better variables that might lead to a more in-depth insight into the game.

The data collected by Sportec Solution lead to the development of "dangerousity" as a metric. This metric attempts to measure the probability of a goal being scored at every time a player has the ball (Link, Lang, and Seidenschwarz 2016; Biermann 2019). This was

achieved by dividing the football pitch into 2x2m^2 squares with a distance of 34 meters from the goal, as shown in Figure 3.

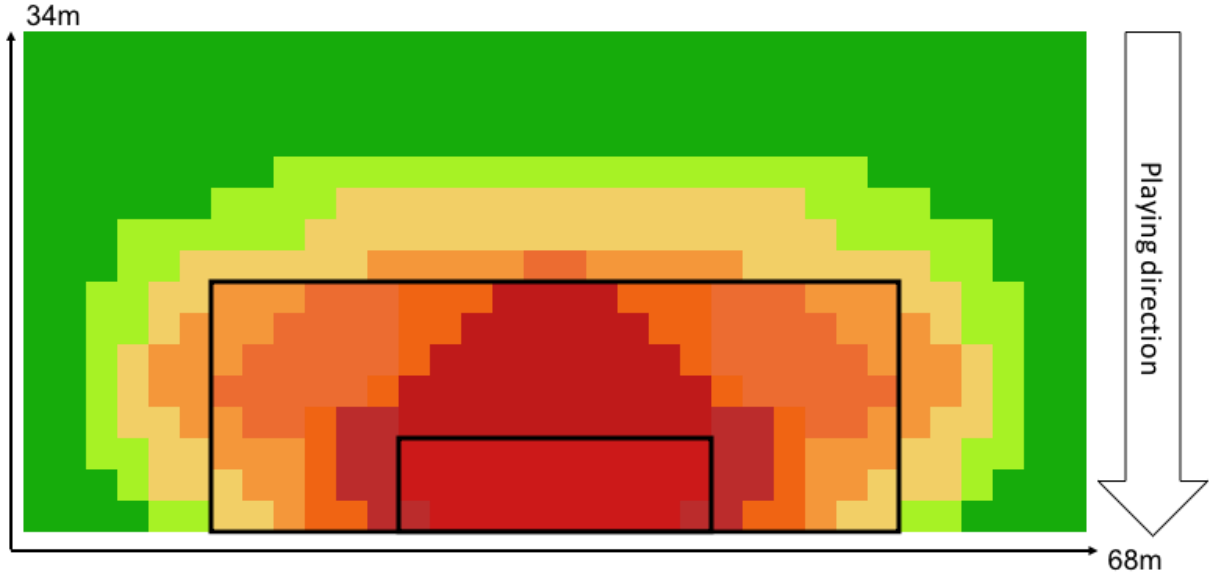

Figure 9 Quantification of Zone by dividing the pitch into 2 x 2 m^2 squares. Adapted from Link, Lang, and Seidenschwarz (Link, Lang, and Seidenschwarz 2016)

The green colour represents a low dangerousity, while a red colour represents higher dangerousity for the attacking team.

This is accomplished by looking at four components:

1) Zone – describes the danger of a player in possession of the ball in a given square
2) Control – the extent a player can implement his tactical intention given the control of the ball he has
3) Pressure – represents the ability that the defending team can prevent the player in possession of the ball by stopping them from scoring
4) Density – the chance the player in possession of the ball has to defend their own goal if they were to lose possession of the ball in the given square they are located.

In the same framework, other metrics to quantify the game was also proposed, such as:

- Action value – how much more dangerous a person can make the situation with their possession of the ball
- Dominance – the difference in performance between different teams

Expected goals (xG) were invented by Sam Green (2012) in 2012, working for OptaPro, trying to find better metrics to evaluate attacking players by looking at the way goals had

been scored previously (Biermann 2019). Based on historical shot data by OptaPro, they were able to quantify the probability for a shot being scored depending only on the distance from goal and if it was a shot from the ground, a volley, or a header (Gregory 2020). This has later proved very useful both for the football teams to more objectively evaluate the players and for the spectators and multiples of media outlets including these in their post-match analysis. This gave the audience a more in-depth look into the game and its decisive moments. Tippett (2019) points out that the real value in xG is in the fact that it can identify why specific types of passes add the most xG to an attack, and by analysing this data, player evaluation can improve. Further, Tippett (2019) and Schoenfeld (2019) points out that this might have consequences for the game on a longer scale, by teams creating tactical approaches and attacking moves that are built around getting the ball in the right spot to maximise xG. The same would also be true for defending, as Williams (2020) argues that defending will be conquered by the teams that minimise the other team's xG.

In later years the xG has been further improved also to include goalkeepers positioning, the number of defenders in between the shot and the goal, and to include whether the shot was taken by the weaker or stronger foot of the attacker. This makes the xG more accurate data, which is essential for any analysis to be conceived as valid, and turn this knowledge into actionable insights (Statsbomb 2018b).

The invention of xG has led to many similar metrics, such as:

- Expected assists (xA) which looks at the probability of the next pass leading to a shot at goal
- Expected goal chain (xGC) is a metric that involves all the players in the sequence of passes leading up to the shot on goal. This can create a pattern of which players are included in the building up to a shot on goal, and reveal which players are mostly involved in chances created for a team.

(Biermann 2019)

2.4 Goalkeeper

With goalkeepers being the only player on the field allowed to use their hands, this creates a different set of skills required for them to master the game and its demands. These demands have been studied in detail with the money pouring in from big European clubs partnering up with sports scientist professionals to get an edge over their other competitors. A lot of research has also been conducted by academics on goalkeepers.

A test to determine the physical skills required for a goalkeeper was developed to assess diving, jumping, sprint running, and directional changes (Knoop, Fernandez-Fernandez, and Ferrauti 2013).
Strand, Krosshaug and Andersen (2011) examined the situations goalkeepers are most prone to injury and identified two situations most injuries occurred. Furthermore, Schmitt, Schlitter, and Boesiger (2010) examined the risk for goalkeepers when they were diving, and found that goalkeepers who rotated when they landed had a significantly lower risk for injury to their hip by doing so.
According to Otte, Millar, and Hüttermann (2019), it was identified that the football team needed to have a goalkeeper coach employed by the team to develop an individual training program suited for each goalkeeper and his skillset.

A large pile of research on goalkeepers and penalties has also been created over the years, with the standardized nature of a penalty kick being the main reason for this (van der Kamp et al. 2018; McMorris and Colenso 1996; Knoop, Fernandez-Fernandez, and Ferrauti 2013; Bar-Eli et al. 2007; Bar-Eli and Azar 2009; Savelsbergh et al. 2005; Peiyong and Inomata 2012). Every single penalty follows the same rules, allowing for research into many directions to be generated and in a game filled with chaos and complexity.

When the back-pass rule was amended in 1992 by FIFA, a different way of playing football was introduced to goalkeepers. They were no longer able to pick the ball up with their hands from a back-pass from a fellow teammate but had to rely on their observational and ball-handball skills to maintain possession. Ito et al. (2004) noticed that the goalkeeper now has a higher amount of passes in games, and the importance of completing these passes successfully has significantly increased in recent years. This development has continued until this day, by successful coaches such as Maurizio Sarri stating that "the goalkeeper is the first attacker" (Mendonca 2018, 146).

This research has created a foundation to create physiologically and mentally strong goalkeepers, as well as understanding the new demands the game requires in terms of distribution. There is, however, a lack of research on key performance indicators (KPI) for goalkeepers. By identifying suitable KPIs for goalkeeper, better quantification of their performances can be made, and thus this is the focus of the rest of this literature review.

The literature agrees that the commonly used metrics such as save percentage (S%), and passing percentage (P%) are not able to accurately predict how well a goalkeeper is performing in a game (Gelade 2014; Castellano, Casamichana, and Lago 2012; Link and Hoernig 2017; Carling et al. 2014). The main argument is that S% does not consider the level of difficulty each shot has but simply aggregates together all shots on target and calculates how many percentages of the shots were saved. Furthermore, there is a bias against keepers in lower-ranked teams as they will have lower S% than goalkeepers in higher-ranked teams by facing a higher number and more difficult shots than the better ones (Gelade 2014).

To find a better metric to look at how goalkeepers are dealing with shots, Statsbomb designed Goals Saved Above Average (GSAA). The idea is that the total xG will say how many goals "should" be conceded during a game. Over time this will show how many goals the goalkeeper "costs" the team and how many goals the goalkeeper has "saved" the team, thus creating an actionable metric clubs have been looking for a long time (Tippett 2019). This metric uses the updated xG from Statsbomb that includes the goalkeepers positioning and the number of players in between the goal and the ball to get the best quality of data and thus increases the accuracy of the model (Statsbomb 2018b; 2018a).

Due to the lack of research on the topic of evaluating goalkeepers performance, Schuckers (2011) tackled this problem for ice hockey goalkeepers, which is very similar to a goalkeepers of football. Schuckers calculated a defensive independent goalie rating (DIGR,) which gives every shot an independent difficulty such that each shot can be viewed objectively and by summarising for the total shots, the total difficulty, and dividing by the number of shots. This gives an independent S% of the difficulty of the shots which can be used to rank the goalkeepers by the ability to save shots in ice hockey. This represents one KPI for an ice hockey goalkeeper that can be used to measure their value objectively. However, the weakness in this model is that it only looks at one metric, the shot-stopping, and not any other areas for an ice hockey team.

Oberstone (2010) presented a quantitative look at what separates the best goalkeepers from the rest in his article by doing a multiple regression that retroactively identifies the specific pitch activities that contributed to increasing the Opta index score. The Opta Index used to be a rating system from the six previous games and based on pitch actions for the players, a score from 0 to 600 was given (OPTA 2011). The analysis contained 34 metrics, ranging

from clean sheets (matches with zero goals conceded), total shots received, and the total number of saves. The remaining metrics that significantly affected the regression analysis were: Shots outside the box, shots from inside the box, punches, short distribution, clean sheets, and goals conceded.

Yam (2019) presents a framework for identifying undervalued goalkeepers by dividing the goalkeeper's actions into three categories, and looking at five variables from Statsbomb's database:

1. Shot stopping - GSAA representing the goalkeepers shot-stopping ability and Positional Deviation from optimum for the goalkeeper
2. Crossing - Crosses Claimed Above Average (CCAA) represents how many crosses the goalkeeper caught, measured against the average goalkeeper in the same league.
3. Distribution - Positive outcome (PO), distribution actions into the opponent's half that results in a free-kick, throw-in, corner, shot, or throw in for the goalkeepers' team. It was also included a variable to look at the difference between a goalkeeper's passing length under normal circumstances and when the goalkeeper is under pressure.

Even though the article is that it only contains data for one season (2017/2018), it managed to identify two goalkeepers who were undervalued; Nick Pope and Dean Henderson. This was achieved by identifying that Dean Henderson had similar metrics to a Premier League goalkeeper despite even though he was playing in League 1 (third tier in English football) at the time. Nick Pope was identified by having the second highest GSAA in the Premier League and also in the top 5 for several other variables in Yam's analysis. Both of them are today playing regularly in the Premier League and are competing for the Nr. 1 jersey for England (Sky Sports 2020; Football365 2020).

The same categorization was created when Statsbomb (2018a) introduced their analysis of goalkeeper performance, and high importance was placed on new metrics developed by Statsbomb to replace, what Statsbomb views, are outdated metrics that do not measure actionable metrics for a goalkeeper.

3.0 Methodology

This is a descriptive study analysing which of the independent variables is most important for a goalkeeper to be playing for a club highly ranked in the ELO rating.

3.1 Quantitative analysis

Regression analysis is chosen because the focus is looking for the relationships between one dependent variable and several independent variables. Firstly, several simple regression analyses are performed to visualise the data and look for trends in the data. Further, a multiple linear regression analysis is run, as it lets us compare the dependent variable to many independent variables and try to account for the variation of the independent variables in the dependent variable synchronically. It also enables us to make predictions about future values for the dependent variable based on values for the independent variable (Uyanık and Güler 2013).

3.2 Data collection

The data collected are composed of Statsbomb's database for all goalkeepers and their metrics, Statsbomb were chosen because they are the market leader in data collection in football, with higher granularity and twice the event data collected for every game (Statsbomb 2020).

An observation is defined as one goalkeeper playing games for one team in one season. If one player is transferred from team X to team Y in the middle of the season, there would be an observation for his metrics for the duration he played at team X, and another observation with his metrics playing for the other team Y, providing the player meets the other requirements for being included in the analysis as well. A total of 1034 observations, spread over 311 teams, were identified in the initial data with 54 different metrics included for each player. To measure the relative skill of a team, the Elo rating system is used as an objective measurement. Elo rating is a system created by Arpad Elo in the 1970s and is meant to calculate the relative skill for a player/team in a zero-sum game (Elo 1978). Originally it was intended to improve the contemporary chess rating system but has later been implemented by FIFA to rank the national teams across the globe (FIFA 2018). To collect the ELO ratings on the website, clubelo.com was used for its function of checking the ELO

rating on any previous date and also for its coverage of European leagues. The ELO is not a direct measure of the individual performance of the goalkeeper, but it measures the performance of the whole team.

Domestic League	***Seasonal coverage***	***ELO coverage***
English 1st and 2nd division	2018/2019 & 2019/2020	Yes
German 1st and 2nd division	2018/2019 & 2019/2020	Yes
Spanish 1st and 2nd division	2018/2019 & 2019/2020	Yes
Italian 1st and 2nd division	1st:2018/2019 & 2019/2020 2nd 2019/2020	Yes
French 1st and 2nd division	2018/2019 & 2019/2020	Yes
Dutch 1st and 2nd division	2018/2019 & 2019/2020	Not for 2nd
Norwegian 1st division	2019	Yes
Belgian 1st division	2018/2019 & 2019/2020	Yes
Austrian 1st division	2018/2019 & 2019/2020	Yes
Danish 1st division	2018/2019 & 2019/2020	Yes

Table 10 Data collection and ELO coverage

It was only data for domestic league games that were collected, as other data was not available through the database provided, and it was not possible to obtain ELO ratings for the Dutch 2nd division; thus, those players were eliminated (n=82) from the analysis.

3.3 Data Processing

After collecting the metrics for the observations, a thorough process of selecting the most accurate metrics to be included in the analysis for all observations is initiated, and the following variables are included:

Name	*Abbreviation*	*Short Description*	*Description*
$\mathbf{Y}$	ELO	ELO Rating	
$\mathbf{X_1}$	GSAA	Goals Saved Above Average	Goals Saved Above Average: Reflects on how many goals should be conceded in regards to the Post-shot xG faced (Statsbomb 2018b). Post-shot xG represents the chance of a goal being scored when the trajectory of the ball is known, which body part it was struck with, and the number of defenders between the attacker and the goalkeeper is known. The aggregated post-shot xG for the match is calculated, and the amount of goals conceded is subtracted to give the GSAA.
$\mathbf{X_2}$	GKAP	Goalkeeper Aggressive Position	Goalkeeper Aggressive Position: How far from the optimal position for facing a shot, the goalkeeper is (on average).
$\mathbf{X_3}$	PE	Positioning Error	On average, how far from the optimal position for facing a shot the goalkeeper is. This is calculated by an algorithm from Statsbomb for every chance (Yam 2019)
$\mathbf{X_4}$	xC	xChain	Total xG from shots coming from possessions a specific player participated in (Biermann 2019)
$\mathbf{X_5}$	Pass%	Pass %	The number of passes that were received by a teammate calculated a percentage of the total number of passes attempted.
$\mathbf{X_6}$	PID%	Pass Into danger %	Percentage of passes made where the recipient was under pressure, or in another dangerous scenario. The danger is defined as having an opponent within 3 meters when the pass is received. Calculated as a percentage of the total number of passes made (Biermann 2019).

$\mathbf{X_7}$	PIP%	Pass Into Pressure %	Percentage of Passes made where the recipient was under pressure, where pressure is defined as having an opponent within 5 meters when the pass is received. Calculated as a percentage of the total number of passes made (Biermann 2019).
$\mathbf{X_8}$	GTS	Game time per season	Number of games played in one season
$\mathbf{X_9}$	PL	Pass length	Average pass length of completed passes
$\mathbf{X_{10}}$	Save %	Save %	Percentage of on-target shots that were saved by the goalkeeper.
$\mathbf{X_{11}}$	Height	Height	Height in cm
$\mathbf{X_{12}}$	Height^2	Height^2	Height in cm squared
$\mathbf{X_{13}}$	Age	Age	Age
$\mathbf{X_{14}}$	Age^2	Age^2	Age squared
$\mathbf{X_{15}}$	xSv%	Expected Save %	Expected save percentage, given the PSxG of shots faced, how many saves is the goalkeeper expected to make. An example could be that if there are 12 shots on target during a game that gives a total Post-shot xG of four, there should be a 75% save rate to concede four goals and save eight shots that did not end up as a conceded goal.
$\mathbf{X_{16}}$	CCAA%	Claims Claimed Above Average	This metric looks at every data point labelled as a cross or High ball in the Statsbomb dataset if this cross/high ball intersects with the 5-yard box at any point it marked as a "claimable cross". It then looks at how often the goalkeeper claims the "claimable crosses", and by looking at a large number of goalkeepers, it will find an average number of crosses that a goalkeeper should come for. Therefore it is possible for this metric to asess if a

			goalkeeper is doing better than the average goalkeeper or if he is not doing that. (Statsbomb 2019)

Table 11 Variable overview

To run the regression analysis, Microsoft Excel is used. However, Microsoft Excel only allows for a maximum of 16 independent variables so a constraint of $n \leq 16$ independent variables are introduced.

By using the same categorisation as Yam (2019) and Statsbomb (2018a) used for their articles, except adding one more category to account for personal details, a comprehensive framework is introduced:

1. Shot stopping to include the most critical aspect of a goalkeeper's life, to keep the ball out of the net. Yam included only X_1 to account for this, but looking at other factors as well, such as X_2 to look at a goalkeeper's ability to sweep behind the defence which will nullify an attack before it gets the chance to develop into a big goal-scoring chance will account for a more substantial portion of the hot stopping aspect. Also, the position of the goalkeeper should be included to account for instances where the goalkeeper is not in the right place before a shot is fired will improve the analysis of a goalkeeper, that's why X_3 is included. Even though it has been established by Gelade (2014) and Castellano, Casamichana and Lago (2012) that X_{10} is an outdated aspect of the game and that does not reflect the quality of the saves, only the number of saves, it will still be included in the analysis to further look at this variable to see if their conclusion is correct, or that the new and revised variable X_{15} is a better measurement of this phase of the game.
2. Crossing. X_{16} is included to account for the goalkeeper's ability to deal with crosses. With more than 17 crosses per game, it is an essential area for goalkeepers to deal with to make it to the top level
3. To look at these aspects of distributing the ball, the percentage of successful passes, passes into danger, passes into pressure, and pass length is included. All of these variables are traditional metrics that have been used for a long time but is judged to be the best representatives of a goalkeeper's abilities to distribute the ball.

4. Personal details such as height and number of games were also included to look at these factors.

Appendix 1 includes the full list of metrics that were collected from Statsbomb.

3.4 Data analysis

Firstly, to ensure that the players who are included in the study are the first-choice goalkeeper, or a worthy replacement on the same level so that the ELO rating will reflect the relative skill of the goalkeeper, a cut off mark is set at ten games per season. This means if a goalkeeper has played less than ten games in a season, their statistical metrics will not be included in the study. Ten games represent between 22% and 38% for the selected domestic leagues and therefore is a suitable cut off point to eliminate players that have not played enough games to generate enough data on their skill level. This step removed 373 observations from the analysis. If players with less than ten games are included, there will be a risk of involving players that played games due to an injury to their first-choice goalkeeper or a red card, thus giving game time to a player which generally would not obtain game time, which would impair the analysis by generating noise.
Thereafter, the height for several players was missing, so this was manually looked up on the transfermarket.com website, it was however, not possible to find the height for all players, so a further 14 observations were eliminated, leaving us with 565 observations to investigate further.

3.4.1 Simple Regression analysis

As the first step in the analysis, several simple regression analyses (SRA) are run to visualise the data at hand. Simple regression analysis is defined by Lane (2018, 464) as a statistical method used to study the relationship between two variables, with one independent variable and one independent variable. Furthermore, by plotting the results in a Cartesian coordinate system, the dependent variable on the Y coordinates and independent variable as X coordinates, the "best-fitted line" according to the observations can be found with the ordinary least squared method. This method minimises the sum of all the residuals, which is the difference from the fitted line and the observation, to create a line that is as close to the observations as possible. All "best-fitted lines" will follow the same function:

$$y = \alpha + \beta x_i + \varepsilon_i$$

$y = \text{dependent variable}$

$\alpha = \text{intersection between y} - \text{axis and } best\ fitted \text{ line}$

$\beta x_i = gradient\ for\ the\ best - fitted\ line\ for\ variable\ i$

$\varepsilon_i = error\ for\ variable\ i$

Equation 4 Simple regression equation

3.4.2 Multiple regression analysis

As the second step of this analysis a multiple regression (MRA) analysis, which is an extension of a simple linear regression, is used to assess the relationship between two or more independent variables and a single continuous dependent variable (Lane 2018). The equation for the multiple linear regression is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

$\hat{Y}$ = predicted or expected value of the dependent variable

β_0 = Value of Y when all the independent (β_1 through β_n) are equal to zero

β_1 to B_n = regression coefficients X_1 to X_n

$\epsilon = error$

Equation 5 Multiple regression equation

A multiple regression analysis of the independent variables is used to explain the dependent variable and is measured by R^2. Each regression coefficient represents the change in $\hat{Y}$ relative to a one-unit change in the respective independent variable, holding all other independent variables constant (Lane 2018).

The last step in the regression analysis is to complete the backward elimination in stepwise regression. The backward elimination starts with all the variables in the analysis and with each step, the variable with the highest P-value is eliminated until all variables have a P-value below a set threshold (Sutter and Kalivas 1993). The limit for this analysis is set at 0.05.

3.4.3 Multiple regression considerations

According to Lane (2018) there are two main concerns when performing an MRA:

- Overfitting. Overfitting occurs when the analysis starts to describe the random error in the data and not the relationship between the variables. This is typically done by including too many variables in the model, which makes it too complicated for the data to find the relationships between dependent and independent variables.
- Multicollinearity. By adding more independent variables to the analysis, relationships between these independent variables are inevitably created. This is a problem because one of the most critical elements of MRA is by looking at the effect on the dependent variable that a change in one independent variable and holding the other independent variables constant. However, if there is a correlation between two independent variables, this will not be possible, as the change in one independent variable will affect another independent variable. We seek to find an analysis where the dependent variable is correlated to all the independent variables, but not that the independent variables are correlated to each other.

4.0 Results

4.1 Simple regression analysis

Firstly, several simple regression analyses are produced to visualise and better understand the data at hand. An example is provided below in Figure 4:

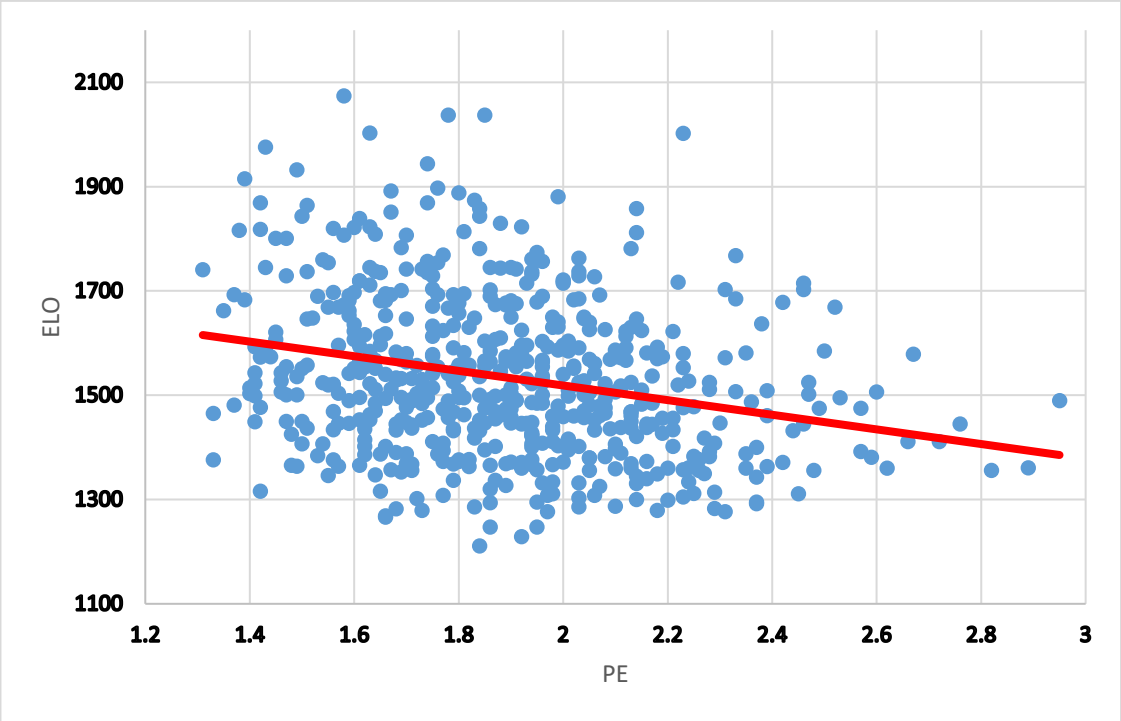

Figure 10 Illustration of an SRA with Positional errors as independent variable and ELO rating as the dependent variable

In this illustration, the y-axis represents the ELO ranking, and the x-axis, the ELO ranking is on the y-axis characterises the number of positional errors per game for goalkeepers. The red line represents the best-fitted line. As demonstrated in Figure 4, there is a clear trend between observations of goalkeepers playing in higher-ranked teams in the ELO rating, positioning themselves more accurately to shots than goalkeepers playing in lower-ranked ELO teams. This explains the downward trend, with goalkeepers in lower-ranked teams in the Elo rating making more mistakes regarding their positioning before shots. The regression coefficient is -140.31 and the interception point being 1799.23, which means that if the goalkeeper were to make one positional error, the ELO ranking would be predicted to be 1799.23 – 140.31 = 1658.69. The position before a shot for a goalkeeper is essential for his probability to save the ball, and thus a considerable increase in their ELO rating should follow if they make more positional mistakes.

Another example can be seen in Figure 5, representing the SRA of GSAA and the ELO rating. Again, a clear trend emerges with goalkeepers playing for higher-rated ELO teams having a higher GSAA score than goalkeepers playing in lower-ranked ELO teams. This time the regression coefficient is 182,49, meaning if the goalkeeper were to improve their GSAA with one unit, the predicted ELO would be increased by 182.49. This seems rational, as the goalkeepers playing in higher-ranked teams in the ELO rating would save the team for more goals compared to players in lower-ranked teams in the ELO rating.

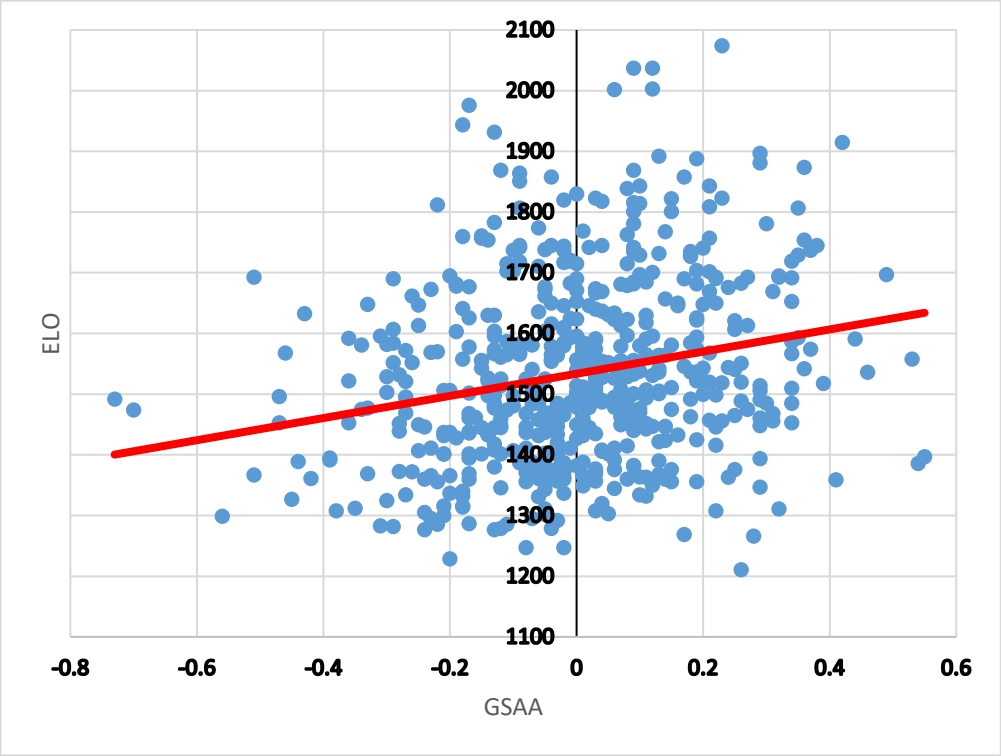

Figure 11 Illustration of an SRA for GSAA and ELO rating

To better understand the following models, it is essential to understand what these statistical metrics mean (Lane 2018). R^2 can be interpreted as how much of the variation for the dependent variable can be explained by the analysis. It is always between 0 and 100%, where 100% represents a perfectly straight line identical to the best-fitted line. The standard error explains what the average distance from the best-fitted line, and the observed value is for the analysis. The P-value is the probability that the same value, or higher, is observed in the sample, given that the null hypothesis is true. A lower P-value would indicate more reliable evidence against the null hypothesis. For our analysis, a P-value of ≤ 0.05 is deemed significant.

Table 3 shows all the SRA in its complete form.

Name of variable	X variable coefficient	X variable standard error	X variable P-value	Significant	Observations	R^2	Intersection coefficient
GKAP	10.47	3.005	0.001	Yes	565	0.021	1337.384
Age	4.47	1.488	0.003	Yes	565	0.016	1408.840
Age^2	0.08	0.026	0.004	Yes	565	0.015	1474.272
Height	2.8	1.506	0.065	No	565	0.006	1005.692
$Height^2$	0.01	0.004	0.067	No	565	0.006	1273.297
GSAA	182.5	33.350	0	Yes	565	0.051	1533.618
PE	-140.31	22.001	0	Yes	565	0.067	1799.270
xC	670.93	81.788	0	Yes	565	0.107	1432.075
CCAA%	-2.55	2.471	0.302	No	565	0.002	1538.587
PID %	-11.28	1.867	0	Yes	565	0.061	1692.579
PL	-6.07	0.860	0	Yes	565	0.081	1807.617
P%	4.76	0.593	0	Yes	565	0.103	1225.186
PIP%	20.84	3.934	0	Yes	565	0.047	1437.184
GTS	2.08	0.754	0.006	Yes	565	0.013	1485.462
xSv%	2.53	1.946	0.194	No	565	0.003	1355.101
S%	5.41	1.018	0	Yes	565	0.048	1151.861

Table 12 Simple regression analysis overview

Table 2 indicates that there is a clear tendency for most variables to improve the ELO rating if the independent variable is increased with one. However, for variables such as PL, PID%, CCAA%, and PE, this is not true, and most variables have P-values that are less than 0.05, which are deemed significant.

There is a very low R^2 for all of the SRAs, which means that the line best fitted to the observations only accounts for < 0.01 of the variability in the model, which indicates that there is a high level of variability that is not accounted for in the model.

For our SRAs, the P-value is generally significant, meaning they are < 0.05, except for xSv%, CCAA%, height, and $height^2$, with the last two narrowly missing out.

4.2 Multiple regression analysis

The purpose of the multi regression analysis is to identify which of the 16 independent variables explains the most variability for the ELO rating and also has a P-value that is significant. To do this, all the independent and dependent variables are inserted into an Excel workbook, and the analysis is initiated by eliminating the variable that has the highest P-value and then removing it from the analysis. This is done until all variables are significant. The following results were found:

Step	R^2	Adjusted R^2	Standard error	Variable eliminated	The P-value for eliminated variable
1	0.298	0.277	130.642	Age	0.861
2	0.298	0.278	130.527	$Height^2$	0.851
3	0.298	0.280	130.412	GSAA	0.609
4	0.297	0.281	130.325	PL	0.264
5	0.296	0.280	130.354	P%	0.321
6	0.294	0.280	130.352	CCAA%	0.264
7	0.290	0.277	130.614	-	-

Table 13 Multiple regression analysis overviews

Step 7 is named "The final model" from hereafter because all P-values are <0.05 from this point. For our model, the changes in R^2 and adjusted R^2 are marginal, which indicates that the model has neither been strengthened nor weakened by eliminating the six variables from the model.

The variables that remain significant are:

Variable name	Regression coefficient	P-value	Standard error
Interception point	959.013	0.001	290.754
Age^2	0.099	0	0.023
GKAP	8.915	0.003	2.952

PID%	-8.056	0	1.815
Save%	4.675	0	1.130
Height	3.254	0.014	1.314
PIP%	12.138	0.001	3.679
PE	-140.348	0	20.834
90s Played	1.922	0.004	0.661
xSv%	-5.648	0.007	2.081
xC	383.532	0	87.213

Table 14 Multiple regression analysis variable overviews

With the regression coefficients now known, the predicted ELO can be calculated with the following equation:

$$\hat{Y} = 959.013 + 0.099X_{14} + 8.915X_2 - 8.056X_6 + 4.675X_{10} + 3.254X_{11} + 12.138X_7 - 140.348X_3 + 1.922X_8 - 5.648X_{15} + 383.532X_4$$

Equation 6 Predicted dependent variable equation

A diagram visualising the predicted ELO and the actual ELO can be useful to see how accurate the model is. The red line represents an increase in one unit on each axis.

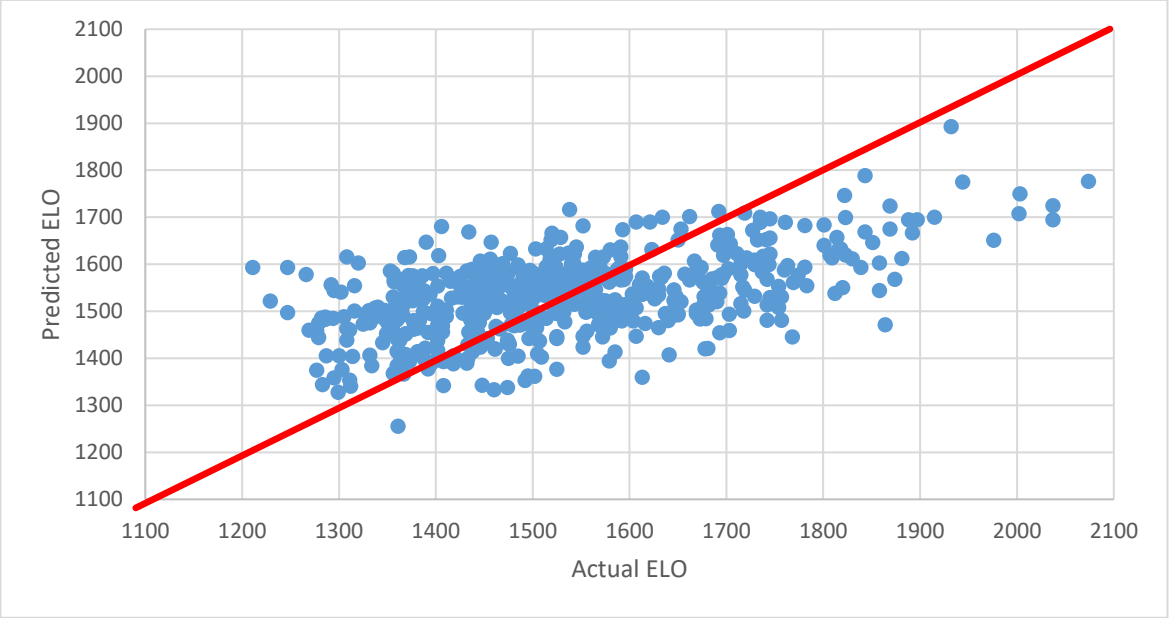

Figure 12 Predicted ELO from the final model and actual ELO rating

Bring (1994) note the importance of standardised regression coefficients to be able to determine the relative importance of the variables. This is achieved by standardising all values from the final model, except the ELO rating in the MRA by the following equation:

$$Z_{ik} = \frac{x_{ik} - \mu_i}{\sigma_i}$$

$x_{ik} = value\ for\ variable\ i\ and\ observation\ k$
$\mu_i = mean\ of\ variable\ i$
$\sigma_i = standard\ diviation\ for\ variable\ i$

Equation 7 Standardising equation

A full list of the mean and standard deviations for the variables is included in Appendix 2. By running the same MRA as is in the final model for the standardized regression coefficients, the following result is found:

R Square	0.290
Adjusted R Square	0.277
Standard Error	130.614
Observations	565

Table 15 Standardised MRA overview

Name	***Coefficients***	***Standard Error***	***P-value***
Intercept	1534.009	5.495	0
Norm Age²	24.584	5.609	0
Norm GKAP	19.013	6.295	0.003
Norm PID	-27.087	6.102	0
Norm Save%	29.026	7.018	0
Norm Height	13.952	5.636	0.014
Norm PIP	19.510	5.914	0.001
Norm PE	-39.909	5.924	0
Norm 90s	16.403	5.637	0.004
Norm xS%	-18.776	6.917	0.007
Norm xC	28.712	6.529	0

Table 16 Standardised variable overview

As expected, when running a standardised version of the final model, the P-value and standard error for the entire model is unchanged, R^2 and adjusted R^2 remains the same because the model has only been standardised, not changed. The reason these standardised regression coefficients are useful is that they are all in one unit, namely standard deviations away from the mean and not in games played, years or cm (Courville and Thompson 2001). When all variables are on the same scale, they are comparable, and it is possible to more reasonably compare the regression coefficients as the relevant impact of each of the variables on the dependent variable. The standardised regression coefficients can be interpreted as "given a one-unit change in Z_i how much $\hat{Y}$ $will\ change$" (Newman and Browner 1991).

5.0 Discussion

5.1 Simple regression analysis

The R^2 values are spread out on the lower end of the scale, with the max being xC with 0.107 and the smallest being xSv% 0.002. This means that the independent variable explains between 0.2% and 10.7% of the variability of the dependent variable. The P-values are significant for all variables except for xSv%, CCAA%, height, and height^2.

The analysis indicates the following form the regression coefficients:

- Highly ranked ELO teams tend to have taller, older, and goalkeepers that coming further out to claim the ball (GKAP), but do not catch more crosses than the league average (CCAA%).
- It also indicates that short passes (PL) to a teammate (P%) that is at least five meters away from an opponent (PIP) will increase the ELO rating of the team.
- Lastly, it implies that goalkeepers who save more shots than is expected of them (GSAA), a higher save percentage and expected save percentage also will increase the ELO rating of the team, while positional errors tend to lead to a lower ELO rating.

All these variables make logical sense, except for the CCAA%. For this variable, the coefficient says that to increase the ELO rating of the team, the goalkeeper tends to claim fewer crosses than the league average, contradicting what is one of the goalkeeper's most important jobs for goalkeepers, namely catching crosses. It is important to note that CCAA% has the highest P-value at 0.302, meaning it is far from significant and that the null hypothesis is most likely true, indicating that there is no relationship between the dependent and independent variables. The R^2 is also the lowest of all variables in the SRA at 0.002, so only a fraction of the variability for the dependent variable can be attributed to changes in the independent variable.

The P-value of xSv% is also very high at 0.193 and has a meagre R^2 score of 0.003; this shows that this variable also might have a true null hypothesis, which means that there is no relationship between the dependent and independent variable.

5.2 Multiple regression analysis

5.2.1 Predicted and actual ELO ratings

The first thing to note about the MRA is that Figure 6 would indicate that our analysis seems to overestimate the ELO rating of teams with a low ELO rating. This can be found by looking at Figure 6 and noticing the high amount of observations that is located above the red line for a low actual ELO rating, and the number of players being below the red line growing as the ELO rating increases. If the model was perfectly calibrated, there would be minimal deviations from the red line for actual and predicted ELO ratings. The smallest deviations from the red line are located to be in the middle of the graph, meaning that the model is best at identifying the observations with an average Elo rating. What seems to happen in the analysis is that there are both highly and lowly ranked ELO teams, which include goalkeepers with high and low variables values and that the regression analysis is not able to differentiate between them. This explains why the lowest predicted value is 1300, and the highest is 1700, whereas the actual ELO is ranging between 1200 and 2000.

5.2.2 Independent variables

After completing the backward elimination process, ten variables remained that had a P-value smaller than 0.05. The R^2 only marginally changed from the first step to the final model, while the adjusted R^2 remained unchanged. The standard error was also consistent throughout the process. When all values from the final model have been standardised, the regression coefficients and standard deviations are similar in absolute terms, and the R^2, P-values, and the standard deviations for the entire model are unchanged from the final model.

By looking at Table 5, it is not easy to interpret which variables are the most important. However, Table 7 tackles this problem because all variables use standard deviations as their unit, making it more coherent. The standardized regression coefficients can be interpreted as a change of one standard deviation to an independent variable, and the ELO rating will increase/decrease with the value of the coefficient. Because all values for the standardized regression coefficients (SRCs) are within a range from 13.9 and 39.9 units, in absolute terms, the outcome is very similar for all independent variables. Therefore, we have identified 10 KPIs for goalkeepers playing on good teams, namely all variables form the final model.

It is noteworthy that almost all the new metrics invented by Statsbomb, namely GSAA and CCAA%, were eliminated in the process, contradicting Yam's (2019) research. This means that the independent variables do not contribute to explaining more of the variability of the dependent variable than the other independent variables already do.

Based on predictions from the final model, the following areas are important for a goalkeeper and the teams ELO rating:

- Older, taller, and more experienced goalkeepers that are sweeping behind their defence (GKAP) tend to increase the ELO rating for the team.
- Short passes to teammates that are at least 5 meters away from an opponent is preferred. To take part in a possession series that leads to a shot on target is also estimated by the MRA to increase the ELO rating of the team.
- A high S% would contribute to a higher ELO taring, but the opposite is predicted by the MRA for xSv%.

Age, height^{2,} and GKAP are variables that have positive coefficients, meaning that the older, taller, and more aggressive the goalkeeper is, it is estimated that the ELO rating will increase. However, this only makes partial logical sense. If a goalkeeper is more towering, he will also have a more extensive shot-stopping ability. Still, if a goalkeeper is too tall, he will also struggle to dive quickly down to save shots near his feet and lose some speed and agility due to his long limbs and heavier weight. Both these factors are essential for a goalkeeper to perform at the highest level and contradict the regression coefficients, which indicates that a taller goalkeeper will lead to a higher ELO rating for the team. It is important to note that the Age variable was eliminated from the MRA, but the Age^2 was included. The difference between the Age variable and the Age^2 is that the Age^2 will put a lot more emphasis on older goalkeepers as their variable value will be squared.
What seems likely is that a minimum height is required to be able to catch crosses, have adequate shot-stopping abilities and that at a certain height, if not it can become a hindrance. The same is most likely true for age, as it would be challenging for a goalkeeper to play at a top European club at the age of 48, and at the age of 16. This is because to develop as a goalkeeper, and it is necessary to gain experience and learn from mistakes by playing games. Over time the young goalkeepers will gain knowledge and understanding about the game and also reach his peak performance age. For older goalkeepers, they will lose their physical

performance, and their recovery time would increase, meaning they would struggle to play and train as much as is required. Lastly, if a goalkeeper is coming further out of his goal to claim through balls and sweep behind the defence, it will improve his value and his contribution to the team. There is, however, a fine line between success and failure, as a lot of attributes are necessary to do so. Both speed and agility to get to the ball first, understanding of the developing attack, and the movements of the opponents and bravery to go into though challenges with the opponent are required. But if the goalkeeper is starting too far out of his goal to claim through balls, it leaves him open shots from far out that can go over him and into the goal. Therefore, it is not possible for the goalkeeper to have an extremely high GKAP score, as he cannot be as far forward as his defenders. Just like age and height[2,] it beneficial to have a goalkeeper that is sweeping behind the defence, but not a goalkeeper that is so aggressive that the opposition can be score past him or attempt to claim passes that are not possible to come for.

If the goalkeeper were to try to improve his xC, to increase the ELO score of his team, it would be challenging. This is because the circumstances dependent on the team's performance after the goalkeeper has distributed the ball. If a goalkeeper were to pass 4 meters to a central defender, and then the team creates a fantastic passing sequence, it is not necessary the pass of the goalkeeper that was the defining piece. It can, however, be that the goalkeeper kicks the ball to the striker with a finely timed pass, which leads to the shot on target to lift his xC score. If the goalkeeper is playing on a team consisting of players who are incredibly skilled the xC score of the goalkeeper would most likely be inflated due to the abilities of the other players, casting a shade on the relevance of the xC variables abilities to describe a goalkeepers ability to distribute the ball.

Lastly, the final model found that if passes were to be made to a teammate within three meters, the ELO rating would be predicted to decrease, This is most likely to do with the time a teammate has the ball before he is pressed by an opponent, as it is much harder to maintain possession if an opponent is three meters away from you, compared to further away.

5.2.3 Multi regression considerations

To avoid overfitting the model, a backward elimination process is executed in the multi regression analysis with a cut-off point at 0.05 for the P-value. This means that only significant variables are left in the analysis to counter overfitting.

To check for multicollinearity, a correlation matrix was created for the final model:

	Age²	**GKAP**	**PID%**	**S%**	**Height**	**PIP%**	**PE**	**90s Played**	**xSv%**	**xC**
Age²	1	-0.088	0.027	0.043	-0.163	-0.006	0.034	0.027	0.050	-0.021
GKAP	-0.088	1	-0.146	-0.087	0.060	0.036	0.259	0.011	-0.065	0.397
PID%	0.027	-0.146	1	-0.138	0.069	0.122	0.126	0.091	-0.187	-0.280
S%	0.043	-0.087	-0.138	1	-0.066	0.083	-0.170	0.114	0.591	0.110
Height	-0.163	0.060	0.069	-0.066	1	0.073	-0.037	-0.032	-0.055	0.045
PIP%	-0.006	0.036	0.122	0.083	0.073	1	-0.096	0.123	-0.022	0.246
PE	0.034	0.259	0.126	-0.170	-0.037	-0.096	1	0.078	-0.111	0.055
90s Played	0.027	0.011	0.091	0.114	-0.032	0.123	0.078	1	0.056	0.064
xSv%	0.050	-0.065	-0.187	0.591	-0.055	-0.022	-0.111	0.056	1	0.029
xC	-0.021	0.397	-0.280	0.110	0.045	0.246	0.055	0.064	0.029	1

Table 17 Correlation matrix for the final model

A correlation analysis is often done to determine if a relationship exists between two variables and how strong this association might be (Taylor 1990). The result is then visualised in a correlation matrix containing all independent variables. The correlation coefficient is often represented as the variable *r* and can take a range of values between -1 and 1. An *r* score of 1 would give a perfectly straight line going upwards to the right, and an r score of -1 would create a straight line going downwards to the right if visualised. In absolute terms, a low correlation would give an r score of ≤ 0.35, a moderate correlation would have an r score between 0.36 and 0.67, and a high correlation would be an r score over 0.67 (Taylor 1990).

For Table 8, there is a low correlation, except xC-GKAP and S%-xSv%, which has a moderate correlation. It makes intuitive sense that the S% and xSv% are correlated, as they

very much calculate the same score. A closer look at how the regression coefficients for S% and xSv% developed through step one to six in the MRA is needed to determine if these are correlated.

Step	xSv% Reg. Coefficient	S% Reg. Coefficient	Variable eliminated
Step 1	-8.298	7.388	Age
Step 2	-8.298	7.381	Height^2
Step 3	-8.284	7.364	GSAA
Step 4	-5.820	4.942	PL
Step 5	-6.007	4.861	P%
Step 6	-5.854	4.916	CCAA%
Step 7	-5.648	4.675	

Table 18 Regression coefficient development for S% and xSv% in the MRA

Table 9 shows that the regression coefficients did not change their sign and was only altered in a minor way. The most significant change was provoked when GSAA was eliminated from the analysis. Further analysis is needed to understand the relationship between xSv% and S%, so two separate MRA are run: one without xSv% and one without S% to look at how the analysis would be without them being included. A full overview of these MRAs can be found in Appendix 3. When xSv% is not included in the analysis, S% is eliminated in step 1, and when S% is not included in the analysis, the xSv% is removed in step 4. For both examples, the R^2 of the final model is marginally higher than both of the alternative MRAs. Both these facts lead us to believe that multicollinearity does not affect the analysis in any meaningful way.

It also seems that xSv% is explaining a lot of the same variability for the dependent variable that GSAA does. This is probably because they are both built on post-shot xG and will, therefore, have the same underlying tendencies. When xSv% is not included in the MRA, the GSAA remains a significant variable, as shown in Appendix 3.

Because xSv% is built on post-shot xG, there is nothing the goalkeeper can do about changing this, but it is more up to the defence to lower shots on target, which in turn would reduce the xSv%. This is different from S%, which only depends on the number of saves the goalkeeper makes out of the total amount of shots, which is something the goalkeeper can

do something about. Therefore, to interpret the negative xSv% regression coefficient, it should be viewed as the fewer shots that produce xG against the goalkeeper, the better for the team.

5.3 Future research

As discussed earlier, there have been identified ten KPIs that are the most important variables for the goalkeepers to improve the team's ELO. Roughly half of the KPIs are traditional metrics (Age^2, Height, S%, 90s played) that have been used for a long time, but the rest are more recent innovations. It is positive that the newly innovated metrics are included in the Final Model, as they remain significant. They do however, only account for the 29% of the variability of the ELO rating for the team, which means that further innovation is needed to explain more of the variability and therefore improve the reliability of the analysis. To accomplish this, innovation from the sports analytics companies is required in order to create even better metrics than the one we have today. Based on conclusions from this analysis, clubs are still overvaluing taller and older goalkeepers, a new focus for them should be on metrics that evaluate their shot-stopping based on the difficulty of the shot, such as post-shot xG, metrics that quantify their ability to claim crosses and distribution. To do this, however, new metrics needs to be invented or further developed to encompass the crossing and shot-stopping aspect of goalkeeping in a better way than is available today.

Potential for innovation regarding the CCAA% is a good foundation but needs to be developed further to be more accurate and also to incorporate the starting position of the goalkeeper before the cross is taken. In the same way that the positioning before a shot is taken is essential for the goalkeeper's ability to save the shot stop, the positioning before a cross is just as important. For shot-stopping, GSAA is another way of improving the metric is by including whether or not the goalkeepers have contact with the ground when the shot is taken. This would be represented as a binary metric and would improve it because the goalkeeper needs contact with the ground to dive to the side. A lot of goalkeepers in today's game do not have their feet on the ground when the shot is taken, meaning it takes longer for them to be able to react to the shot. This might increase the goalkeeper's time to react where the shot is coming, which might again lead to a higher probability of saving the shot.

By further developing the dangerousity metric developed by Link (2016) to quantify better the distribution aspect football will lead to a better understanding of which goalkeeper is better equipped to distribute the ball in a fashion that is beneficial for the team.
One aspect that, most likely, it will not be able to quantify is the goalkeeper's ability to communicate. This is a crucial skill that a lot of goalkeepers lack in, which has consequences for the team's performance. By having excellent communication skills, the goalkeeper can prevent shots before they are taken by getting teammates in the right position so that they can block or tackle the attacker before he is able to shoot. This is an intrinsic ability that advances in technology most likely will not be able to measure accurately, but an understanding of that this skill essential for a goalkeeper is nevertheless necessary for practitioners.
Currently, there are two paths being explored by football analytics:

- Closed research – where knowledge is developed by a club and is kept in-house to develop a competitive edge over the rest (New York Times 2019). This is done because clubs are investing heavily in this field and do not want to share their insights with competitors in fear of being copied and surpassed.
- Open research – Clubs, academics, and private organizations group together their forces to develop new knowledge and publish this for others to read and develop further (Evans et al. 2019). This has been the way knowledge has been created for decades by building on what has been done previously and increasing the understanding of new fields of research in a transparent way.

With the competitive nature of football clubs, this trend will undoubtedly continue, but to tackle a challenge like quantifying football in a meaningful and useful way, cooperation and knowledge-sharing are needed, maybe not directly between clubs, but between academics, conferences and companies.

Another factor in deepening the knowledge of the quantitative side of football is by hiring more sporting directors, who oversee all footballing operations in a football club. The importance of a sporting director is highlighted by Parnell et al. (2018), which notes that by consolidating all responsibilities for the different departments within a football club, a holistic and long-term plan can be created. This plan will surpass any single manager leading the first team, and in this way, continuity and longevity can be given to a sports analytics department. In the current climate, absolute power is given to the manager of a team, and

the identity of the club can often be sacrificed to mould around the manager's footballing principles. By having a sporting director, a reliable and durable plan can be devised for football clubs, compared to today's situation of sacking the manager every ten months and starting from scratch again, but also assign more influence over the transfer policy for a club (Biermann 2019).

As noted by both Bierman (2019) and Knutson (2020), the success of analytical movements does not depend on being able to confirm already established doctrines to gain the approval of the footballing environment. It depends on finding small and meaningful advantages and adding these together to leverage them effectively. It is only then that an analytical revolution can truly begin.

6.0 Conclusion

In this report, 16 variables were included in a multi regression analysis, and in the end, ten key performance indicators remained significant. The KPIs explained 29% of the variability of the ELO rating the goalkeeper is playing for, where 4/10 were traditional metrics, and 6/10 were newly invented, which gives us an even split between them. There is still a lot of variability that needs to be explained by the independent variables to make this analysis usable by commercial entities. However, it has identified a clear trend between the ELO rating and the KPIs. The analysis also highlighted the importance of new innovations in the search for better ways to quantify football. This could be accomplished by either improving the existing metrics or finding new ways to break down the game into pieces that can be quantified in a better way than is done today. This can result in reduced costs for the club because less expenditure will be lost on players that do not live up the expectations of them and on wage and travel costs related to the club's scouting network.

In a business always chasing short term fixes, a longer planning horizon needs to be implemented for stability, consistency, and longevity. By having a sporting director that is in charge of footballing operations, the influence in the transfer philosophy of the sporting analytics department can be increased and over time, the impact can spread to other areas such as hiring managers and coaches, as well as players.
For this to be a reality though, innovations from the sports analytics community are needed to become more widespread in the footballing community.

7.0 Reference list

Bailey, Nate, Karan Bhuwalka, Hin Lee, and Tim Zhong. 2018. "Understanding Features of Successful 3 Point Shots in the NBA."

Bar-Eli, Michael, and Ofer H. Azar. 2009. "Penalty Kicks in Soccer: An Empirical Analysis of Shooting Strategies and Goalkeepers' Preferences." *Soccer and Society* 10 (2): 183–91. https://doi.org/10.1080/14660970802601654.

Bar-Eli, Michael, Ofer H. Azar, Ilana Ritov, Yael Keidar-Levin, and Galit Schein. 2007. "Action Bias among Elite Soccer Goalkeepers: The Case of Penalty Kicks." *Journal of Economic Psychology* 28 (5): 606–21. https://doi.org/10.1016/j.joep.2006.12.001.

Baumer, Benjamin, and Andrew Zimbalist. 2013. *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*. University of Pennsylvania Press.

Biermann, Christph. 2019. "FOOTBALL HACKERS : The Science and Art of a Data Revolution."

Bring, Johan. 1994. "How to Standardize Regression Coefficients." *The American Statistician* 48 (3).

Carling, Christopher, Craig Wright, Lee John Nelson, and Paul S. Bradley. 2014. "Comment on 'Performance Analysis in Football: A Critical Review and Implications for Future Research.'" *Journal of Sports Sciences* 32 (1): 2–7. https://doi.org/10.1080/02640414.2013.807352.

Castellano, Julen, David Casamichana, and Carlos Lago. 2012. "The Use of Match Statistics That Discriminate between Successful and Unsuccessful Soccer Teams." *Journal of Human Kinetics* 31 (1): 139–47. https://doi.org/10.2478/v10078-012-0015-7.

Courville, Troy, and Bruce Thompson. 2001. "Use of Structure Coefficients in Published Multiple Regression Articles: β Is Not Enough." *Educational and Psychological Measurement* 61 (2): 229–48. https://doi.org/10.1177/0013164401612006.

Deloitte. 2019. "World in Motion Annual Review of Football Finance 2019." *Deloitte Annual Review of Football Finance 2019*, no. May: 40.

Elo, Arpad. 1978. *The Rating of Chess Players, Past and Present*.

Evans, Nicolas, Daniel Memmert, Christopher Clemens, Alexia Putellas, Robert Moreno, Raúl Peláez, Carlos Rodriguez, et al. 2019. “FOOTBALL ANALYTICS : NOW AND BEYOND.”

FIFA. 2018. "Revision of the FIFA / Coca-Cola World Ranking." 2018.

https://resources.fifa.com/image/upload/fifa-world-ranking-technical-explanation-revision.pdf?cloudid=edbm045h0udbwkqew35a.

Football365. 2020. "What Will Happen to Dean Henderson This Summer ?" 2020. https://www.football365.com/news/dean-henderson-man-utd-sheffield-united-summer.

Frazier, Walt, and Alex Sachare. 2004. *He Complete Idiot's Guide to Basketball*. Penguin Putnam.

Gelade, Garry. 2014. "Evaluating the Ability of Goalkeepers in English Premier League Football." *Journal of Quantitative Analysis in Sports* 10 (2): 279–86. https://doi.org/10.1515/jqas-2014-0004.

Goldsberry, Kirk. 2019. "Sprawlball: A Visual Tour of the New NBA."

Green, Sam. 2012. "Assessing the Performance of Premier League Goalscorers." OptaPro. 2012. https://www.optasportspro.com/news-analysis/assessing-the-performance-of-premier-league-goalscorers/%0Ahttps://www.optasportspro.com/about/optapro-blog/posts/2012/blog-assessing-the-performance-of-premier-league-goalscorers/.

Gregory, Sam. 2020. "BLOG : Expected Goals in Context Understanding a Team' s Underlying Performance," 1–8.

Hakes, Jahn K., and Raymond D. Sauer. 2006. "An Economic Evaluation of the Moneyball Hypothesis." *Journal of Economic Perspectives* 20 (3): 173–85. https://doi.org/10.1257/jep.20.3.173.

Harvard Business Review. 2012. "Four Steps to Measuring What Matters" 3: 9–10. https://hbr.org/2012/10/how-to-pick-the-right-metrics.

———. 2018. "Moreyball : The Houston Rockets and Analytics." 2018. https://digital.hbs.edu/platform-digit/submission/moreyball-the-houston-rockets-and-analytics/.

Hughes, Mike, and Ian Franks. 2005. "Analysis of Passing Sequences, Shots and Goals in Soccer." *Journal of Sports Sciences* 23 (5): 509–14. https://doi.org/10.1080/02640410410001716779.

Ito, Kosaku, Masamitsu Ito, Ryosuke Wakasugi, Takashi Takemiya, and Toshio Asami. 2004. "Effectiveness of Amendments of the Laws of the Game to the Goal Keeper in Soccer," 1–7.

Kamp, John van der, Matt Dicks, Jose Antonio Navia, and Benjamin Noël. 2018. "Goalkeeping in the Soccer Penalty KickTorhüten Beim Strafstoß Im Fußball." *German Journal of Exercise and Sport Research* 48 (2): 169–75.

https://doi.org/10.1007/s12662-018-0506-3.

Knoop, Marco, Jaime Fernandez-Fernandez, and Alexander Ferrauti. 2013. "Evaluation of a Specific Reaction and Action Speed Test for the Soccer Goalkeeper." *Journal of Strength and Conditioning Research* 27 (8): 2141–48. https://doi.org/10.1519/JSC.0b013e31827942fa.

Knutson, Ted. 2020. "Waiting For The Revolution At Soccer Analytics Bootcamp." 2020. https://deadspin.com/waiting-for-the-revolution-at-soccer-analytics-bootcamp-1836224038.

Lane, David. 2018. "Introduction to Statistics." 2018. http://onlinestatbook.com/Online_Statistics_Education.pdf.

Lewis, Michael. 2003. *Moneyball: The Art of Winning an Unfair Game*. *W.W. Norton & Company*. https://doi.org/10.1111/j.0022-3840.2005.140_11.x.

Link, Daniel, and Martin Hoernig. 2017. "Individual Ball Possession in Soccer." *PLoS ONE* 12 (7): 1–15. https://doi.org/10.1371/journal.pone.0179953.

Link, Daniel, Steffen Lang, and Philipp Seidenschwarz. 2016. "Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data." *PLoS ONE* 11 (12): 1–16. https://doi.org/10.1371/journal.pone.0168768.

Mackenzie, Rob, and Chris Chusion. 2013. "Performance Analysis in Football: A Critical Review and Implications for Future Research." *Journal of Sports Sciences* 9 (1): 79–96. https://doi.org/10.1080/02640414.2012.746720.

McMorris, Terry, and Sion Colenso. 1996. "Anticipation of Professional Soccer Goalkeepers When Facing Right- and Left-Footed Penalty Kicks." *Perceptual and Motor Skills* 84 (3): 931–34. https://doi.org/10.2466/pms.1996.82.3.931.

Medeiros, Joao. 2017. "How Analytics Killed the Premier League's Long Ball Game." Wired Uk. 2017. http://www.wired.co.uk/article/premier-league-stats-football-analytics-prozone-gegenpressing-tiki-taka.

Medium. 2019. "Data Science in the Major Leagues Data Science : On The Field." 2019. https://medium.com/@connor.anderson_42477/data-science-in-the-major-leagues-3bd251333471.

Memmert, Daniel, and Dominik ^{R}aabe. 2018. *Data Analytics in Football*. Routledge. https://doi.org/10.1007/978-3-319-57870-5_9.

Mendonca, Pedro. 2018. "PREPARING FOR COMPETITION - The Method of Maurizio Sarri." In .

Merrimack College. 2019. "How NBA Analytics Is Changing Basketball." 2019.

https://onlinedsa.merrimack.edu/nba-analytics-changing-basketball/.

Mottley, Charles 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): 335–38. https://doi.org/10.1287/opre.2.3.335.

Murphy, Ryan H. 2019. "Ben Lindbergh and Travis Sawchik, The MVP Machine: How Baseball's New Nonconformists Are Using Data to Build Better Players." https://doi.org/10.1007/s11138-019-00493-6.

New York Times. 2019. "How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory." *New York Times*, 1–7. https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html.

Newman, Thomas, and Warren Browner. 1991. "In Defence of Standardized Regression Coefficients." *Epidemiology Resources*.

Oberstone, Joel. 2010. "Comparing English Premier League Goalkeepers: Identifying the Pitch Actions That Differentiate the Best from the Rest." *Journal of Quantitative Analysis in Sports* 6 (1). https://doi.org/10.2202/1559-0410.1221.

OPTA. 2011. "What Is the Opta Index." 2011. https://www.tottenhamhotspur.com/news-archive-1/what-is-the-opta-index/.

Otte, Fabian W., Sarah-Kate Millar, and Stefanie Hüttermann. 2019. "How Does the Modern Football Goalkeeper Train? – An Exploration of Expert Goalkeeper Coaches' Skill Training Approaches." *Journal of Sports Sciences* 00 (00): 1–9. https://doi.org/10.1080/02640414.2019.1643202.

Parnell, Daniel, Paul Widdop, Ryan Groom, and Alex Bond. 2018. "The Emergence of the Sporting Director Role in Football and the Potential of Social Network Theory in Future Research." *Managing Sport and Leisure* 23 (4–6): 242–54. https://doi.org/10.1080/23750472.2018.1577587.

Peiyong, Zhou, and Kimihiro Inomata. 2012. "Cognitive Strategies for Goalkeeper Responding to Soccer Penalty Kick." *Perceptual and Motor Skills* 115 (3): 969–83. https://doi.org/10.2466/30.22.23.PMS.115.6.969-983.

Rose, Robert. 2013. "Defining Analytics." *ORMS Today*, 2013. https://doi.org/10.1016/b978-0-12-401696-5.00001-3.

Savelsbergh, Geert J.P., John Van der Kamp, A. Mark Williams, and Paul Ward. 2005. "Anticipation and Visual Search Behaviour in Expert Soccer Goalkeepers." *Ergonomics* 48 (11–14): 1686–97. https://doi.org/10.1080/00140130500101346.

Schmitt, Kai Uwe, Maja Schlittler, and Peter Boesiger. 2010. "Biomechanical Loading of

the Hip during Side Jumps by Soccer Goalkeepers." *Journal of Sports Sciences* 28 (1): 53–59. https://doi.org/10.1080/02640410903369927.

Schuckers, Michael E. 2011. "DIGR: A Defense Independent Rating of NHL Goaltenders Using Spatially Smoothed Save Percentage Maps." *MIT Sloan Sports Analytics Conference*, 8.

Simon, Kuper, and Szymanski Stefan. 2009. *Soccernomics*. Nation Books.

Sky Sports. 2020. "Nick Pope for England ? Burnley Goalkeeper' s Stats Make Him No 1." 2020. https://www.skysports.com/football/news/11096/11947759/nick-pope-for-england-burnley-goalkeepers-stats-make-him-no-1.

Sportec Solutions. 2020. "SPORTEC PRODUCTS & SERVICES Broadcast & Digital Media Services," 1–4.

Statsbomb. 2018a. "Intro to Goalkeeper Analysis." 2018. https://statsbomb.com/2018/11/intro-to-goalkeeper-analysis/.

———. 2018b. "The Dual Life of Expected Goals (Part 2) | StatsBomb," no. Part 1: 2018–21. https://statsbomb.com/2018/05/the-dual-life-of-expected-goals-part-2/.

———. 2019. "The XClaimables ! Measuring Keeper Aggressiveness." 2019. https://statsbomb.com/2019/01/the-xclaimables-measuring-keeper-aggressiveness/.

———. 2020. "StatsBomb Media." 2020. https://statsbomb.com/media/.

Strande, E, T Krosshaug, and T E Andersen. 2011. "Injury Risk for Goalkeepers in Norwegian Professional Football." *ISMJ International SportMed Journal* 4 (4): 1–18. https://doi.org/10.1136/bjsm.2011.084038.

Sutter, Jon, and John Kalivas. 1993. "Comparison of Forward Selection, Backward Elimination and Generalized Simulated Annealing for Variable Selection." *Microchemical Journal* 47: 60–66.

Taylor, Richard. 1990. "Interpretation of the Correlation Coefficient: A Basic Review." *Journal of Diagnostic Medical Sonography* 6 (1): 35–39. https://doi.org/10.1177/875647939000600106.

The Athletic. 2017. "A High-Scoring Revolution Has the Rockets Soaring." 2017. https://www.theatlantic.com/entertainment/archive/2017/11/a-high-scoring-revolution-has-the-rockets-soaring/547103/.

Tippett, James. 2019. *The Expected Goals Philosophy: A Game-Changing Way of Analysing Football*. Independently Published.

Uyanık, Gülden Kaya, and Neşe Güler. 2013. "A Study on Multiple Linear Regression Analysis." *Procedia - Social and Behavioral Sciences* 106: 234–40.

https://doi.org/10.1016/j.sbspro.2013.12.027.

Williams, Josh. 2020. "Liverpool Are Using Incredible Data Science during Matches , and Effects Are Extraordinary." 2020. https://www.liverpool.com/liverpool-fc-news/features/liverpool-transfer-news-jurgen-klopp-17569689.

Wright, Mike. 2009. "50 Years of or in Sport." *Journal of the Operational Research Society* 60 (SUPPL. 1): 161–68. https://doi.org/10.1057/jors.2008.170.

Yam, Derrick. 2019. "A Data Driven Goalkeeper Evaluation Framework," 1–18. https://statsbomb.com/data/.

Appendix 1

Metric Name	Short Description	Description
M_1	Age	The age of the player in years
M_2	Height	The height of the player in cm
M_3	OP Passes	Average passing completion % allowed by the opponent
M_4	Pass%	Proportion of Passes made that were completed successfully
M_5	Long Balls	The average number of long balls, longer than 30 meters, per game
M_6	L/R Footedness%	Aggregate the total number of passes with the left foot and with the right foot, divide the smallest amount by the biggest. A higher number would indicate more two footedness
M_7	Long Ball%	The amount, in percentage, long balls account for in all passes attempted
M_8	Pr. Long Balls	Number of long passes to a player who is under pressure (pressure is defined as an opponent within 5 meters when receiving the ball)
M_9	UPr. Long Balls	Number of long passes when the goalkeeper is under pressure
M_{10}	Passes Pressured%	The proportion of passes that is passed to a teammate which has a defender within 5 meters
M_{11}	Pr. Pass%	Proportion of Passes made that were completed successfully, while under pressure
M_{12}	Pr. Pass% Dif.	Subtract the Pr Pass% from the Pass%, and find the difference between regular passes, and passes completed under pressure
M_{13}	Deep Progressions	Passes, dribbles and carries into the opposition final third per game
M_{14}	xGBuildup	The total xG of possession a player was involved in an outside shot or assist
M_{15}	xGChain	Total xG from shots coming from your possession
M_{16}	Carries	Amount of meters the ball was carried forward during a game
M_{17}	Carry%	In percentage, how successful (lose/keep possession of the ball) the carries were
M_{18}	Carry Length	The average distance the ball was carried forward each time

M_{19}	OP F3 Passes	Passes into the opponents final third
M_{20}	PintoB	Number of passes into the opponent's box
M_{21}	OP Passes Into Box	Number of passes by the opposition into our box
M_{22}	SP PintoB	No adequate definition was found for this variable
M_{23}	Passes Inside Box	Passes inside the opponent's box to another teammate
M_{24}	Touches In Box	Number of touches inside the opponent's box
M_{25}	PinTin	No adequate definition was found for this variable
M_{26}	Through balls	A pass splitting the opponent's defence for a teammate that has a shot
M_{27}	Box Cross%	The amount, in percentage, of crosses, account for the total amount of passes
M_{28}	Successful Crosses	How many of the crosses a player took reached a teammate
M_{29}	Crossing%	How many, in percentage, of the crosses reached a teammate
M_{30}	Pass Length	The average length of passes, in maters
M_{31}	Pr. Pass Length	The average length of passes under pressure, in meter
M_{32}	Pass Length Ratio	The ratio between regular passes, and passes under pressure
M_{33}	Pr. Pass Length%	In percentages, how long are the passes under pressure compared to regular passes
M_{34}	Pr. Pass Length Dif.	The difference between passes under pressure, and regular passes
M_{35}	Succ. Pass Length	The average length of passes that reached a teammate
M_{36}	Succ.Pr. Pass Length	The average length of passes that reached a teammate, when the passer was under pressure
M_{37}	Goals Conceded	Number of goals conceded
M_{38}	PSxG Faced	Post Shot xG: Total number of xG face from shots on target
M_{39}	GSAA	Goals Saved Above Average: Reflects how many goals should be conceded in regards to the Post-shot xG faced
M_{40}	Save%	The number of shots on target divided by how many saves a goalkeeper made
M_{41}	xSv%	Expected save percentage, given the PSxG of shots faced, how many saves is the goalkeeper expected to make.
M_{42}	Shot Stopping%	Expected save percentage, given the PSxG of shots faced - calculated as the PSxG/Saves
M_{43}	xG Faced	Total xG from all shots, including those off-target

M_{44}	Shots Faced	Number of on-target shots faced
M_{45}	Shots Faced OT %	The percentage of shots faced by the goalkeeper that were on-target
M_{46}	All Shots Faced	Number of shots faced, including those off-target
M_{47}	Positioning Error	How far away from the optimal position for facing shots a goalkeeper were, on average
M_{48}	GK Aggressive Dist.	How far from the goal a keeper is coming forward to perform defensive actions
M_{49}	Claims%	Claims Claimed Above Average reflects how many crosses the GK should have claimed compared to the. Ref: https://statsbomb.com/2019/01/the-xclaimables-measuring-keeper-aggressiveness/ league average
M_{50}	Pass into Danger%	Passes reaching 15 meters within the opponent's goal
M_{51}	Pass into Pressure%	Percentage of Passes made where the recipient was under pressure
M_{52}	Positive Outcome	Number of times a player is involved in a sequence that soon results in a positive outcome, such as throw in opponents final third, shot on/off target, corner or free-kick on final third
M_{53}	Positive Outcome%	How frequently a player is involved in a positive outcome, expressed as a percentage of the total positive outcome score a team achieves while the player is still on the pitch

Appendix 2

Standard deviation and mean for independent variables in MRA for standardized variables.

Name	μ	σ
Age²	801.433	249.230
GKAP	18.789	2.132
PID%	14.053	3.362
Save%	70.681	6.208
Height	189.281	4.287
PIP %	4.646	1.607
PE	1.890	0.284
90s Played	23.345	8.532
xSv%	70.637	3.324
xC	0.151	0.0748

Appendix 3

MRA without xSv% included as a variable:

Step	R^2	The standard error for the sample	Variable eliminated	Variable eliminated P-value
Step 1	0.294	130.8	S%	0.929
Step 2	0.294	130.7	$Height^2$	0.871
Step 3	0.294	130.5	Age^2	0.865
Step 4	0.294	130.4	PL	0.248
Step 5	0.293	130.5	P%	0.34
Step 6	0.292	130.4	CCAA%	0.06
Step 7	0.290	130.3		

MRA without S% included as a variable:

Step	R^2	The standard error for the sample	Variable eliminated	Variable eliminated P-value
Step 1	0.295	130.7	$Height^2$	0.874
Step 2	0.295	130.6	Age	0.87
Step 3	0.295	130.5	xSv%	0.649
Step 4	0.294	130.4	PL	0.238
Step 5	0.293	130.5	P%	0.34
Step 6	0.291	130.5	CCAA%	0.07
Step 7	0.287	130.3		