Faculty of Sciences

Soccer Player Performance Rating Systems for the German Bundesliga

Jonathan D. Klaiber

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Christophe Ley

Department of Mathematical Statistics

Academic year 2015–2016

UNIVERSITEIT GENT

Faculty of Sciences

Soccer Player Performance Rating Systems for the German Bundesliga

Jonathan D. Klaiber

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Christophe Ley

Department of Mathematical Statistics

Academic year 2015–2016

Foreword

This master thesis completes the one-year advanced master program in Statistical Data Analysis. It has been an intensive, interesting and, above all, knowledgeable year in which I was able to sharpen my statistical thinking, to practice and acquire new problem solving skills and to make beneficial contact to other statisticians and researchers. Therefore I want to thank all professors, docents and teaching assistants very much for their effort and excellent work during this year.

Special thanks go to professor Christophe Ley for his guidance during the thesis project as well as for his lessons and teaching in other courses which will certainly prove to be valuable for our future career.

Contents

1 Abstract

The objective of this study was to find an optimal performance rating system for the Bundesliga, the German national soccer competition. We implemented a statistic based performance rating system which is called the Player Performance Index (PPI) as described in the article *"On the Development of a Soccer Player Performance Rating System for the English Premier League"* by its developers McHale, Scarf and Folker (2012). The PPI is composed of six subindices that capture different elements of the soccer game. We used the PPI to rate the player performances of the Bundesliga season 2015/2016. It turned out that the first match outcome subindex is suited to predominately capture the performance of defenders and midfielders, the fourth scored-goals subindex to rate the performance of forwards and the sixth clean-sheet subindex to evaluate goalkeeper performance. We concluded that the PPI is an adequate, objective and fair performance rating system with some minor weaknesses. The shortcomings that emerged were the disadvantaging of players that do not play often and the overvaluing of goal scorers at the expense of defenders and goalkeepers. In the second part of this thesis we made a rating system comparison. The comparison was meant to show what kinds of features a good rating system should have and whether there is a difference between statistic based and human expert based systems. The statistic based rating systems PPI and whoscored.com and the human expert based systems of Kicker, Bild and Comunio.de were compared with regard to the end-of-season ratings of the Bundesliga season 2015/2016. Based on this comparison we concluded that the rating distributions of the human expert systems make use of the complete rating scale and appear to be more normally distributed than the distributions of the statistic based systems. Another finding was that the expert based systems rate goalkeepers favourably whereas statistic based systems favour offensive players. Last but not least it became apparent that the PPI is better than the other systems in detecting players that are important for the game but do not stand out with many goals or assists. All in all, we concluded that the human expert based rating systems are appropriate to rate the performance of Bundesliga players as it is done now. However, we argue that a weekly updated statistic based rating system would be a valuable extension for the Bundesliga. The PPI rating system could be this statistic based rating system under the condition that the discussed improvements are applied.

2 Introduction

Since the early days of soccer in the second half of the 19th century, when the Football Association (FA) and the first soccer clubs were formed, enthusiasm for playing and watching others playing the game constantly grew (Dunning et al., 1999). The popularity led to a sense of community of the general public. Especially working-class people, the media and soccer clubs were united by their passion for soccer. Soon the media provided a platform for soccer with articles and discussions that were continued in stadiums, pubs, on the street and in amateur soccer clubs. In the 60s/70s, the most influential soccer media in Germany, the Bild newspaper and the Kicker sports magazine, began to publish school-type grades for players of the just recently founded professional soccer competition, the Bundesliga. The motivation behind these grades was to have an expert based performance evaluation of Bundesliga players which could foster the debates. Other stakeholders in soccer, namely the clubs themselves, were also interested in having an informed external opinion about the performance of their players.

Before continuing, we will give a brief explanation for readers that are not familiar with soccer, or football as it is commonly referred to in Europe. Soccer is played on a (grass) pitch that is 100 - 110 meter long and about 75 meter wide. At each end of the pitch there is a 7.3 by 2.4 meter big rectangular goal. Soccer is played with a leather ball and each of the two teams consists of eleven players. The teams compete with each other to put the ball into the opponents' goal, which is called *"scoring a goal"*. The last pass of the ball before a goal is scored is referred to as *"an assist"*. Players are not allowed to use their arms to touch the ball except for the goalkeeper who is allowed to use all body parts in a rectangular area around his goal. Next to the goalkeeper, there are other specific roles or positions that players can take. The defenders play just before the goalkeeper and their task is to keep the ball away from the own goal and to build up the attacking game from behind. The role of the midfielders is to help out the defenders as well as to distribute the ball to the other players in order to create situations that allow the team to score a goal. Finally, the players that are positioned close to the opposing goal are referred to as attackers, strikers or forwards. Their main task is to score goals. A team wins a soccer game by scoring more goals than the opposing team. If both teams score an equal number of goals, the game outcome is called a *"tie"* or *"draw"*. In multiple-game national competitions (or leagues), the teams receive points according to the outcome of their games. Typically, the winning team gets awarded three points while the losing team gets no points. Both teams receive one point for a tie. Some additional mechanisms apply when soccer tournaments are played, as for example the *FIFA World Cup*. Whereas points are won in the group stage of a tournament similar to league competitions, the post-group stages are knockout rounds. Knockout rounds mean the winning team goes to the next round and the losing team drops out of the tournament. When there is a tie after 90 minutes, an additional 30 minutes playing time is added. If there is still a tie after the additional time has passed, the winner is determined by a penalty shoot-out.

Thesis Objective The objective of this article is to study how the performance of Bundesliga players can be optimally rated. To do this we compute a statistic based rating system for individual soccer player performance that is officially used to rate Premier League players. This rating system is the *EA SPORTS Player Performance Index (PPI)* and we will construct it for the first time for the Bundesliga as described in the article *"On the Development of a Soccer Player Performance Rating System for the English Premier League"* by McHale et al. (2012). After having obtained the PPI ratings, we will compare this statistic based rating system to two human expert based rating systems, *Bild* and *Kicker*, one user community rating system, *Comunio.de*, and the statistic based rating system of the website *whoscored.com*. The goal of this exploratory comparison is to find out what the similarities, strengths and weaknesses of different rating systems are in order to find out how an optimal player performance rating system (for the Bundesliga) should look like.

Explanation of Thesis Setup This thesis is divided into two main parts, one part about the construction of the Player Performance Index and a second part about the comparison between the Kicker, Bild, Comunio.de user, whoscored.com and PPI end-of-season ratings. In the first part we explain in detail what kind of rating system types exist, how a rating system relates to a ranking system and the difficulty of separating the individual from the team performance. The first part continues with explaining how the data for constructing the PPI was collected and how it was preprocessed into a structured format. Furthermore, we shed light on every of the six subindices that form the PPI as a weighted composite score. In the results section of the first part we analyse the results of the most important match outcome subindex 1 that relates player actions to the match outcome. We also examine the Top 15 players according to the PPI final score and discuss the aspects that resulted in a high PPI score for these players. The result section is rounded up by the inspection of the PPI performance rating distribution and relevant distribution characteristics. The first part is concluded by a short summary and discussion of notable facts that became relevant during the PPI construction.

The second part about the rating system comparison starts off with an in-depth explanation of the Kicker, Bild and Comunio.de user rating system. This introductory section is followed by a description of how the end-of-season ratings of all Bundesliga players in the different rating systems were collected. The results section begins with a comparison of the rating distributions of the different systems. Moreover, we transform all ratings to a common scale in order to have a scale-independent point of view on the distributions. The results section continues with a comparison of the Top 20 players of each of the five rating systems. We chose to compare the rating systems with regard to ranks, because ranks are scale independent and let us easily compare differently scaled rating systems. Finally, we will compare the end-of-season ranks of the players that were appointed as being the team captains of the 18 Bundesliga teams in the 15/16 season. The second part closes with a short summary of the results and a discussion about what we learnt from the comparison.

This paper will end with a concluding section that puts this study into the broader context of soccer and sports performance rating. Furthermore we debate shortcomings of the current study and lay out what the next steps could be for research that aims in the same direction as ours.

2.1 Rating System Types

The performance of an individual player or team in sports is typically measured in the form of rating and ranking systems. A rating system is broadly defined as the process of assigning a numerical value to the individual and/or team performance on a predefined scale (Stefani, Pollard et al., 2007). Sport performance rating scales usually correspond to well-known scales that are already used in a country, as for example the school grading system in case of the Bild newspaper. Ratings can be used as absolute performance measures if there is a defined value on the scale that indicates an average performance or a predefined standard. If a performance is associated with a higher value than the standard value on the rating scale, it means that the player performed above standard. However, more often rating systems are used for relative performance comparisons, in soccer for example between players within and between teams, between different player positions and/or between different points in time. Of course this is only possible if the same scale is used or a scale transformation is applied and if each player theoretically has the same chance to end up on a certain rating. Stefani et al. (2007) formally define three basic types of sport rating systems that are useful for the construction, evaluation and comparison of rating systems. These types are the subjective, adjustive and accumulative system and they can be applied to rate individual as well as team performances. Next to the different rating system types, it is important to understand the relation between rating and ranking systems and to clarify one of the main difficulties in soccer performance rating, the differentiation between individual and team performance.

Subjective Rating Systems In the subjective rating system, human experts rate the performance. In such systems, usually more than one expert does the evaluation in order to control for possible bias. Subjective rating systems do not have to be solely based on expert opinions but the experts can consult player statistics and other supportive material while forming an opinion about the performance. The person who rates the performance does not have to be a professional expert but could be an amateur who does the performance rating as a hobby. Moreover, amateur communities that discuss and rate the performance together are popular among supporters.

Accumulative Rating Systems According to Stefani et al. (2007), the accumulative rating system is an objective rating system that *"evolve[s] monotonically over a given period"* in which the time period could be the length of a soccer game, an entire soccer league season or any other period. In such a system, the individual player or team gains points for good performances and receives no points when the performance was bad. An example is the Bundesliga competition in which teams receive three points for a win and one point for a tie. Further, Stefani et al. (2007) claim that the ratings used in accumulative systems are arbitrary and ad hoc which leads to the situation that *"the same teams might be ranked differently by different point scales"*. Accumulative rating systems can be extensive when they also explicitly consider the strength of the opponent and/or incorporate a time component that weighs recent performances more than past ones. Stefani et al. (2007) argue that accumulative systems give an incentive to participate in as many games as possible to earn many points. Last but not least, Stefani et al. (2007) conclude that accumulative systems are exclusively useful for performance measurement and not for the prediction of future performances *"because the rating difference of future opponents does not correspond in any obvious manner to predicted score difference or to probability of success."*

Adjustive Rating Systems In the adjustive system, a pre-defined starting point or prior exists. A prior can for example be an average performance or a past performance of the individual/team. The adjustive rating system updates the prior depending on the current performance. This type is usually more sensitive to the current performance in a given game than the accumulative type and it is therefore also more dynamically responsive (Stefani et al., 2007). The most significant difference between the adjustive and accumulative system is that the former one is much less time dependent and can rate a player that only played in half of all league games and a player that participated in all league games equally well. The accumulative system, on the other hand, heavily depends on the time period and thus leads to less points for the players who do not participate in all games.

Rating vs. Ranking Systems The relation between a rating system and a ranking system is that the former is measured on an interval (or ratio) scale while the latter is measured on a less informative ordinal scale. An interval scale as used in rating systems encodes the degree of difference between the ratings which allows comparing how much better one rating is than another. On the other hand, an ordinal scale as used in ranking systems encodes information about which rank is better or worse than another rank, but does not convey information on how much better or worse it is. A rating system contains more information and can always be transformed into a ranking system by ordering the ratings and assigning ranks to the ordering. The vice versa transformation from ranking to rating system is not possible because ranking systems do not contain information about the degree of difference between ranks. Transforming rating into ranking systems makes sense in a situation in which rating systems apply different scales and cannot be conveniently standardized.

Individual vs. Team Performance A good rating system in soccer should be able to find a balance between the influence of the individual and the team performance on the rating. It should identify good performances of players in bad teams and bad performances of players in well performing teams. This feature of a good rating system is hard to satisfy. Even though players might play many accurate crosses, they will receive a poor rating because their teammates could not turn any crosses into goals and consequently none of the players will receive any (rating) points for assists or goals. Soccer is a team sport and therefore team and individual performance are tightly connected and impossible to disentangle. Nevertheless, the aim of a good player performance rating should be to disentangle as much as possible in order to be fair with regard to all players.

2.2 Statistic Based Rating Systems

Statistic based rating systems have traditionally been more popular in English than in continental European soccer. Supporters of the statistic based rating system argue that it is a fair, (mostly) transparent and, above all, objective measurement of player performance. A statistic based rating system can be either of the accumulative or of the adjustive type. Another merit of statistic based rating systems is the extent of data input they use. The available data can significantly influence how a rating system is constructed and in turn how well it evaluates the performance.

Basic rating systems use data that is comprised of goals, assists, shot attempts, pass accuracy and successful tackle percentage of one or more games. If these variables are the foundation of an individual performance rating system, it is mostly an intuitive, basic, offensive-player-focused system. While a system based on this data might be suited to compare offensive capabilities of attack and midfield players and give an impression of their overall performance, it is neither suited for rating defenders and goalkeepers, nor to distinguish between team and individual performance. Nevertheless, it is frequently used as a *"quasi"* performance rating system on soccer television, probably because these numbers are instantly available after a game and straightforward to interpret.

More elaborate rating systems in terms of data use the aggregated player actions and team measures. A player action is for example the number of passes, crosses, blocks or saves of a player. Team measures are for example the number of corners or the league points that a team won. The *Player Performance Index* rating system (McHale et al., 2012) that will be under closer examination in this study is an example of a system that takes the aggregated player actions as data input. It depends on the specific rating system, but typically this data input should allow one to construct a rating system that is able to rate the performance adequately and fairly. Besides, a system based on this data could find a balance between individual and team performance and be able to identify good players in weak teams and weak players in good teams.

Other rating systems take very extensive data into consideration. Dense or extensive data is mostly not aggregated but consists of single player and team events as well as the events' associated time and location. The *"Football Live"* system of the Press Association (McHale et al., 2012) provides for example ball-by-ball data with player action and location data attached to it. This data is detailed to the extent that essentially the entire game could be reproduced, exactly as it happened, in a computer simulation. A player performance rating system that uses extensive data is *whoscored.com*, a website that is operated by the football data supplier *OPTA*. Another one is the *Audi Performance Index* which focuses in particular on the performance of Bayern Munich players and was constructed by the Bayern Munich sponsoring body and car manufacturer *Audi*. Szczepański (2008) developed yet another rating system based on extensive data in which individual player actions influence the rating depending of zones of the pitch wherein they take place. In addition, this rating system incorporates the pressure on the player that is exerted by the opponent players. Another example of a rating system that uses extensive data is the *"B-FASST"* system of Schultze and Wellbrock (2015) that is also time and location dependent. They use a plus-minus metric and compute a performance coefficient that depends on the time a player was on the pitch when goals were scored and on a winning probability that is obtained from the *Bet365* betting company. The plus-minus component depends on time in the sense that goals give more points when they are scored when they are valuable, as for example when the goals so far indicate a draw and fewer points in a situation in which the game is already decided because one team is leading by a high margin. Interestingly, the *"B-FASST"* rating system of Schultze and Wellbrock (2015) could be seen as *"meta"* rating system because it partly consist of the rating system, or at least the system to obtain winning probabilities, from another performance evaluating body, namely the betting company *Bet365*. Similar to the rating system of Szczepański (2008), the *"B-FASST"* rating system makes use of spatial information in order to evaluate the importance of a player action.

2.3 Human Expert Based Rating Systems

Human expert based systems are of the subjective rating system type and give a rating to a player performance based on the experts' knowledge and experience. Thereby, most of the expert systems introduce objectivity by forming an average rating from several experts. Usually, there are two or more observing experts in the stadium to follow the game live and write down and immediately evaluate important events. Due to the high frequency of events happening in soccer games, other experts are watching the game again on tape to capture aspects that had been overlooked by the live observers. Finally, all experts form a common opinion which, in case of agreement, is published and, in case of disagreement, is discussed until a performance rating is found that satisfies all experts. Besides being subjective, many expert based systems, as for example the rating that Kicker magazine and Bild newspaper publish, are also adjustive rating systems. They give the same prior rating to every player and update this rating during the game according to a player's performance. While one subcategory of expert ratings is of the type that was just mentioned, namely employing professional experts to come up with the ratings, another subcategory are community formed ratings. Communities, often people who participate in online soccer management games like *Comunio.de*, rate the performance of players individually and publish the average rating of all community members as the final performance rating. In many cases the performance is discussed just after the game in online bulletin boards, especially the performances for which diverting opinions exist.

2.4 Statistic vs. Expert Based Systems in Comparison

We discuss presumptions about statistic and expert based rating systems and evaluate, for the assumptions that can be evaluated, whether they are likely to be true or not in a comparison between rating systems.

An advantage of statistic based system over experts is that the former are on average less expensive. The data that is used as input for statistic systems is collected and supplied automatically and once the statistical rating model is implemented, it does not cost much to produce the performance ratings. For an expert rating, professional football experts have to be employed who cost a lot of money, even though assessing player performances might not be the only task they are employed for. The statistic based and community based systems are approximately equal in cost because both do not cost much once a computer model or community board have been established. An aspect that is related to the cost is the time it takes until the performance ratings are published. Whereas the statistic based ratings are published as soon as the data has streamed in, the professional experts need some more time to validate their ratings by post-match video analysis and to discuss the final ratings. The user community based ratings do take some time as well because the ratings are not reliable until a sufficient large number of users have rated the performance.

Another argument in favour of statistic based systems is that they appear to be more objective than the expert based systems. The model that rates the performance is transparent to everybody because the function that links the input variables to the output performance rating is publicly known and the input variable data is collected in the same reproducible way in any game. This is at least the case for rating systems developed by academics as the PPI or the *"B-FASST"* system, not for the commercial whoscored.com and *Audi* performance rating systems. A statistic based system guarantees that a player's name and team name are completely independent from the performance rating. This is not guaranteed for the expert ratings. Although the expert opinion is typically the averaged opinion of multiple experts, it could be still affected by some bias. Experts might for example have their favourite players and teams, or they might have expectations about the performance of certain players which will influence their judgement.

An advantage of expert over statistic based rating systems is that the expert systems are more flexible and can adapt more quickly. For example when a team changes their tactic and is suddenly playing fewer passes but it turns out to be effective (hypothetically), then the statistic based model rates the performance wrongly because it was calibrated to reward for a high passing rate while the experts see the new effective playing style and can directly adapt their performance judgement according to this new situation.

3 The PPI model

3.1 PPI Part - Methods

The data of the Bundesliga (BL) seasons 2009/2010 to 2014/2015 was used to build the rating model and was collected in the week of 23.05.2016 to 29.05.2016 from the website whoscored.com. The whoscored.com website is provided by the soccer data supplier OPTA and is frequently used by the community to get soccer statistics. Data from the BL season 2015/2016 was collected to calculate the PPI scores. Performance measures of every player in every match were extracted. The scraped performance measure variables and additional information are listed in Table 6 in the Appendix.

Data Used in This Study OPTA is the world's leading soccer data supplier and operates in all major soccer leagues. During a soccer match, about 2000 on-the-ball events are logged. An event contains several descriptors like the event location on the pitch, the player and team reference and a time-stamp. Events can be described in great detail as for example a successful tackle that prevented a goal, a left-footed volley shot or a diving header. The events are partly logged by computer programs, as for example the location on the pitch, and partly by OPTA experts along the pitch and in the office. In addition, post-match video analysis is used to validate the logged data and add events that could not be logged during the match. This extensive match data is expensive to produce and therefore not freely available to the public. However, OPTA provides less detailed match data to the soccer public by means of the website whoscored.com. The whoscored.com website provides player action measures per match and player. Per match means that all player actions during a match are summed up and presented as a total score. Individual and team player action measures that were obtained from whoscored.com are listed in Appendix Table 6.

3.1.1 Data Collection

Collecting The Raw Data A scraper was built with *Python 2.7* to collect the data. Scrapers are computer software that (automatically) extract information from websites and transform the unstructured data of the web into structured data that can be stored and analysed. The scraping works as follows. First the whoscored.com URL for a given season was read and the web browser was automatically opened and directed to the season's overview page[1]. The software testing framework *"Selenium"* version 2.53 was used to handle all web browser operations and partly to extract the data from within the web page. On the overview page, the season calendar was opened and every match day was

[1] A screenshot of the whoscored.com overview page at the moment of scraping can be accessed via the URL `https://github.ugent.be/jklaiber/ratingsystem/blob/master/data_collection.png`

clicked on, one by one, to obtain the match URLs of all soccer matches on that matchday. The statistics that were necessary to construct the PPI were listed in a tab on the match website. Within the player statistics tab, there were four tables with summarized, offensive, defensive and passing statistics. Player information from these tables was extracted for every player on every matchday as described above. Furthermore, the player information, a match identifier, matchday identifier and the date were saved to check whether the data was valid and structured correctly. After all player data of a matchday had been extracted, the data was saved in a CSV file on the hard disk. The python script that links all scripts together and contains the main functionality to scrape the raw data can be accessed through the URL `https://github.ugent.be/jklaiber/ratingsystem/blob/master/whoscored_playerStats_scraper.py`. Several other scripts with more specific functionality are connected to this main script file. The python helpers script contains python related functions (`https://github.ugent.be/jklaiber/ratingsystem/blob/master/python_helpers.py`). The file helpers script consists of functions that make it possible to save and load text and CSV files (`https://github.ugent.be/jklaiber/ratingsystem/blob/master/file_helpers.py`). The actual scrapers and functions that are related to the scrapers can be found via URL `https://github.ugent.be/jklaiber/ratingsystem/blob/master/scrapers.py` and `https://github.ugent.be/jklaiber/ratingsystem/blob/master/scraper_helpers.py`, respectively. The flow diagram in Figure 7 in the Appendix illustrates the relations between the different software scripts.

There are two main problems with collecting data via web scrapers. Firstly, the hosts of websites do not like that information is automatically extracted and secondly, the web scraping process itself is unstable. Website hosts do not like that information is extracted via a scraper because it means additional non-human web traffic which is costly and they fear that their data could be sold or illegitimately used after extraction. Many website have ways of detecting web scrapers and blocking their IP addresses in order to prevent them to visit the website. This was also the case for the whoscored.com website. To circumvent being detected and blocked, random elements in the form of variable clicking times were integrated into the scripts to mimic human behaviour. In addition, proxy servers were used to alter the IP address from time to time to create the impression that a new user is visiting. These two measures successfully prevented the scraper of being blocked on the whoscored.com website. Web scraping is inherently unstable because of varying Internet connection speed (with occasional blackouts), other programs that interact with the web browser and/or (whoscored.com) servers that do not respond. In order to deal with this instability, separate *"check files"* were created that monitored which matchdays were scraped and also whether the data was complete. To obtain all data from seven BL seasons, the scraper had to rerun several times to get all the bits of data that had been *"overlooked"* or incorrectly scraped in a previous run. Once all *"check files"* indicated that the data of all seasons was properly scraped, the scraper did not have to rerun anymore and the data could go into preprocessing.

3.1.2 Data Preprocessing

The data was preprocessed to be prepared for the main task of constructing the PPI ratings. In the first preprocessing step, the time that a player was on the pitch was calculated because this information was only available for substitutes on the whoscored.com website, not for non-substituted players. There was also an extra variable added for the received league points per match. In the next step, it was (again) controlled whether the data was correct and whether there was match data in the first leg and their associated rematches in the second leg of the season. Other variables that had to be added were whether a clean-sheet result (no goals have been received) was obtained and the number of saves that the goalkeeper made during a match. The number of goalkeeper saves was calculated as the number of on-target-shots of the opponent team minus the clearances of the own team minus the opponent goals. When a goalkeeper substitution took place, the saves were divided between the goalkeeper and his substitute proportionally to their time on the pitch. The python script that contains data preprocessing functionality can be accessed via `https://github.ugent.be/jklaiber/ratingsystem/blob/master/data_preprocessing.py`. The script that checks whether the correct data was scraped is available via the URL `https://github.ugent.be/jklaiber/ratingsystem/blob/master/matchID_scraper.py`.

3.2 Building the PPI Model

3.2.1 Subindex 1 - Match Outcome

The purpose of match outcome subindex 1 is to capture the influence of player actions on the number of league points that a team wins. The essential idea is that the more players contribute to a victory or tie with their actions, the better these players are and the more points they should receive on match outcome subindex 1. McHale et al. (2012) break down the process that leads to a specific match outcome. They define two elements that are needed to model the expected number of goals for and against the team: The player actions that determine the expected number of shots that a team can fire onto the opponent goal and the shot effectiveness, which is simply the proportion of shots than can be turned into goals. Finally, the goals for and against a team fully determine the match outcome of gaining three, one or none league points. Figure 1 shows the process from player action to match outcome according to McHale et al. (2012).

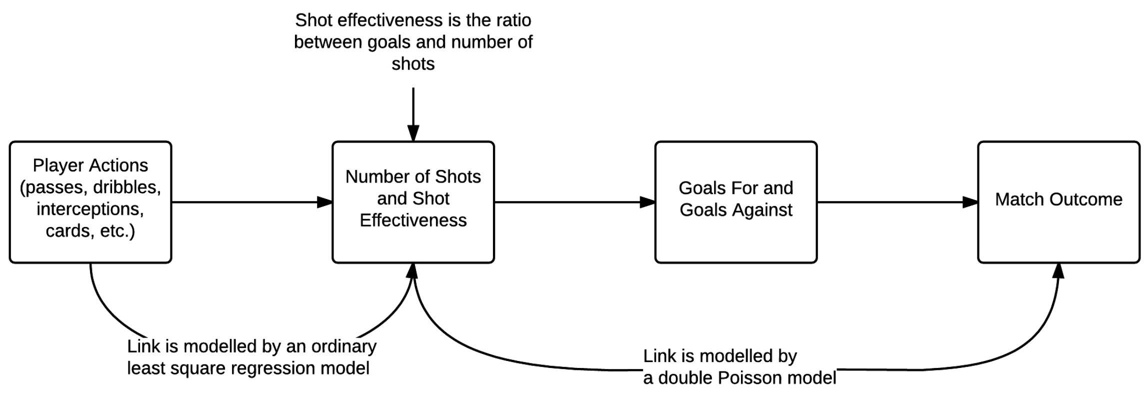

Figure 1: Process that relates player actions to match outcome.

Expected Number of Shots In match outcome subindex 1, the number of shots can be seen as an independent random Poisson distributed variable where the mean is called the *"shot rate"* by McHale et al. (2012). The shot rate is the number of expected shots onto the goal during a full length soccer match. Although McHale et al. (2012) write that *"One could (and probability should) estimate the regression model allowing exactly for the Poisson nature of the dependent variable, and using a log-linear link function so that the logarithm of the expected number of shots is linear in the player contributions"*, the authors decided to estimate the expected number of shots due to simplicity by means of an ordinary least squares (OLS) regression. Modelling the expected number of shots by an OLS regression does not take the dependency between the goals into consideration because, according to McHale et al. (2012), their *"experiments revealed that the final index was no more intuitive with more complicated models for shots"*. The expected number of shots (for home team H) looks as follows:

$$E\{H_S\} = \alpha_0 + \alpha_X H_X + \alpha_Y H_Y \cdots \beta_U A_U + \beta_V A_V + \cdots , \quad (1)$$

where $E\{H_S\}$ is the expected number of shots of the home team, H_X the number of player actions of type X of the home team and A_U the number of player actions of type U of the opposing team. The coefficients α_X and β_U represent the change in the expected number of home team shots when the home player action X increases by one, or the opponent player action U increases by one, respectively.

Variables	Coef.	Std. err.	t-stat	p-value
Home Team Regression Model				
Constant	2.727	0.573	4.762	0.000
Crosses	0.201	0.014	14.785	0.000
Dribbles	0.137	0.020	6.849	0.000
Passes	0.013	0.001	12.771	0.000
Opponent interceptions	-0.028	0.014	-1.933	0.053
Opponent yellow cards	0.390	0.084	4.663	0.000
Opponent red cards	1.010	0.337	3.001	0.003
Opponent tackles won	-0.034	0.018	-1.913	0.056
Opponent cleared	0.016	0.012	1.373	0.170
Away Team Regression Model				
Constant	2.951	0.494	5.967	0.000
Crosses	0.220	0.014	15.446	0.000
Dribbles	0.134	0.019	7.177	0.000
Passes	0.007	0.001	7.522	0.000
Opponent interceptions	-0.040	0.013	-2.990	0.003
Opponent yellow cards	0.267	0.079	3.395	0.001
Opponent red cards	1.061	0.319	3.327	0.001
Opponent tackles won	-0.008	0.015	-0.523	0.601
Opponent cleared	0.041	0.011	3.647	0.000

Table 1: The least square regression model captures how the number of shots is related to team and opponent team player actions. The regression model is once estimated for the home team point of view (above) and away team point of view (below). Abbreviations: Coef.: Coefficient, Std. err.: Standard Error and t-stat: T-Statistic.

Table 1 shows the outcome of the regression model for the home and away team. The result of the regressions is almost the same between home and away team and similar to the results of the original McHale et al. (2012) paper. The main difference between the current results and the results of the original article is the much higher constant of 6.463 (95% Confidence Interval (CI): 4.87 to 8.06) that was found for the Premier League. This discrepancy could be due to an overall higher number of shots in the Premier League compared to the Bundesliga. The coefficients in the regression model can be interpreted as the expected number of extra shots when a player action increases by one, while all other player actions are kept constant. When an opponent player receives a red card for instance and is sent off the pitch, the home team can expect on average one more shot on the opponent goal (1.01, 95% CI: 0.35 to 1.67), controlling for all other variables. The coefficients for crosses, dribbles and passes in Table 1 are all positive which confirms the intuition that more crosses, more dribbles and more passes are associated with more shooting on the goal. In contrast, opponent interceptions and the number of successful tackles by the opponent seem to be negatively associated with the expected number of shots because they hinder a team from shooting (however, both of these coefficients are not significant when a significance level of 0.05 is applied). The intuition with regards to yellow and red cards seems to be confirmed, the coefficients of the cards indicate a relation between an increased number of expected goals and sending players off the pitch or showing them the yellow card. Only the coefficient for opponent clearances is counterintuitive because it indicates a positive association between opponent clearances and expected number of shots, however, this coefficient is not significant for the home team.

Shot Effectiveness The probability of a goal given a shot is called *shot effectiveness* and determines, together with the shot rate, the goals that a team scores and concedes (the conceding goals are estimated by means of the shot rate and effectiveness of the away team). A shot is turned into a goal if the shot is on target, if no opponent player can block the shot and if the goalkeeper cannot save it. The shot effectiveness captures the effects of a player's ability to shoot on instead of off target, the team's blocking capabilities and the goalkeeper's ability to save the ball. A shot effectiveness probability can be estimated for all teams in general (one estimate for all teams), for the home and away team separately (two estimates for all teams) or game-by-game for both of the two teams (one estimate per game and per team). McHale et al. (2012) argue that game-by-game estimates would be accurate but would also lead to an estimated shot effectiveness of zero in games in which a team does not score. A shot effectiveness of zero would lead to zero points for player contributions for that game on match outcome subindex 1 because the expected number of goals are determined by the product of the expected number of shots and the shot effectiveness. McHale et al. (2012) finally chose for the simplest approach of a general estimate for the shot effectiveness, ignoring team and game-by-game specific effects and the home/away effect. The shot effectiveness of BL teams was estimated based on season 09/10 to 14/15 and is $p_{HG|HS} = p_{AG|AS} = p_{G|S} = 0.11$ (Standard Error (SE): 0.02, 95% CI: 0.07 to 0.15) for (home/away) goals given (home/away) shots $p_{HG|HS}$ and $p_{AG|AS}$, respectively. This means that on average approximately one goal is scored from nine shots.

Expected Number of Goals The expected number of goals is the estimated shot effectiveness probability multiplied by the expected number of shots:

$$\lambda_{HG} = p_{HG|HS} \times E\{H_S\} \text{ and,} \tag{2}$$

$$\lambda_{AG} = p_{AG|AS} \times E\{A_S\}, \tag{3}$$

where λ_{HG} is the expected number of home goals, $p_{HG|HS}$ is the probability of a home goal given a home shot (home shot effectiveness) and $E\{H_S\}$ the expected number of home shots (home shot rate) which is estimated by the regression model with the player actions as predictors. The expected away team goals λ_{AG} are constructed analogously.

Relation Between Player Action and Match Outcome McHale et al. (2012) note that player contributions can be linked to the number of rewarded league points (match outcome) for the home team by means of the double Poisson model:

$$E\{\text{PTS}_H\} \tag{4}$$

$$=(3 \text{ points}) \times (\text{probability of a win}) + (1 \text{ point}) \tag{5}$$

$$\times (\text{probability of a tie}) \tag{6}$$

$$= \exp\left(-\lambda_{HG} - \lambda_{AG}\right)\left\{3\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}\frac{\lambda_{HG}^{j}\lambda_{AG}^{i}}{i!j!} + \sum_{i=0}^{\infty}\frac{\lambda_{HG}^{i}\lambda_{AG}^{i}}{i!i!}\right\}, \tag{7}$$

where $E\{\text{PTS}_H\}$ is the expected number of league points for the home team. The same construction of the double Poisson model of the expected number of league points applies to the away team. The number of points that should be awarded to a player on match outcome subindex 1 for a player action X is the change in the expected number of league points when the number of player actions X increases by one. This is the derivative of $E\{\text{PTS}_H\}$ with respect to the (home) player action H_X: $\partial E\{\text{PTS}_H\}/\partial H_X$ (McHale et al., 2012). The points for a player action X are calculated by holding all other player actions at their team-within-match specific values. The sum of points resulting from all player actions in a game is the score on match outcome subindex 1.

The derivative of the double Poisson model for player contribution X cannot be calculated in the form as presented in equation (7) because the model contains sums to infinity. In practice, we determined that the player contribution derivative evaluations are accurate enough with summing up to $i = 10$. To limit the sum at $i = 10$ instead of $i = \infty$ will bias the estimator $E\{\text{PTS}_H\}$ but only for expected shot rate values $E\{H_S\}$ that are very extreme. The sums in equation (7) can be intuitively understood as summing up the expected number of goals (Keller, 1994). From above we know that the expected number of goals is a function of the shot effectiveness $p_{G|S}$ and the expected number of shots $E\{H_S\}$. In order to obtain an expected goals value that is larger than ten (and would in turn bias the estimator of expected number of league points), the expected number of shots $E\{H_S\}$ would have to be larger than 91 because the shot effectiveness $p_{G|S}$, the second factor that determines the expected number of goals, is fixed at $p_{G|S} = .11$. An expected shot rate of 91 is in practice a very extreme value as it would mean to take one shot per minute. McHale et al. (2012) did not make any suggestion about the upper limit of the sums in equation (7) and therefore we decided to apply $i = 10$ as a sensible upper limit based on the argument given.

Finally, the points for match outcome subindex 1 are rescaled so that the total points on subindex 1 match the total number of league points that could be obtained within a season which were 847 league points in the 15/16 BL season. After rescaling, a player's points on match outcome subindex 1 represent the expected increase in league points for the team due to the player actions of this player.

3.2.2 Subindex 2 - Point-Sharing

The purpose of point-sharing subindex 2 is to allocate points to players that played when their team won league points. The ratio between the minutes that a player was on the pitch and the sum of all players' minutes on the pitch is calculated first. For a player who played from the first to the last minute of the game, the playing time ratio is 90 / 900 = 0.091. The final points of point-sharing subindex 2 are calculated by multiplying the playing time ratio with the number of league points that the team won.

3.2.3 Subindex 3 - Appearances

The appearance subindex 3 resembles point-sharing subindex 2 because it rewards players for having played a lot during a game but without relating it to the points that the team won. The appearance index awards players solely for having played. The average number of points won in any game by any team has to be calculated for this index. The average number of league points for a game in BL season 09/10 to season 14/15 was 1.38[2]. The average number of league points is multiplied by the playing time ratio (as in point-sharing subindex 2) to form the points on appearance subindex 3.

3.2.4 Subindex 4 - Scored-Goals

Scored-goals subindex 4 is the goal-scoring index and has been implemented as part of the PPI rating system to reward players for having scored a goal. The average number of league points that a goal is worth has to be calculated. This is simply the total number of league points divided by the total number of goals. The average number of league points per goal was 0.946 in the BL seasons 09/10 to 14/15. Scoring a goal in the Premier League is worth more points than a goal in the Bundesliga because the ratio between points and goals in this competition is 1.039 (McHale et al., 2012). The points on scored-goals subindex 4 are the number of goals of a player multiplied by the average number

[2]In the original PPI article of McHale et al. (2012), the average number of points for a game in the Premier League was 1.34. The lower average number of points means that games ended more often in ties in the Premier League than in the Bundesliga.

of league points per goal. The Bayern Munich forward Robert Lewandowski for example received $30 \times 0.946 = 28.39$ points on scored-goals subindex 4 for the 30 goals he scored in the BL season 15/16.

3.2.5 Subindex 5 - Assists

Analogous to the goal-scoring index, the assist subindex 5 allocates points for providing assists. An assist is the final pass to a player who scores a goal. The average number of points per goal (0.946) is multiplied by the number of assists to get the points on assist subindex 5.

3.2.6 Subindex 6 - Clean-Sheet

Clean-sheet subindex 6 is concerned with rewarding players for not having received any goals, a so-called *"clean-sheet"*. To keep a balance between the subindices, the points that are awarded for a clean-sheet are made to be equal to the points that are awarded for assists. All assists are summed up (total assists BL 15/16: 583) and multiplied by the average number of league points per goal. The total number of clean-sheet games (total clean-sheet games BL 15/16: 163) is then divided by the result of this multiplication to get the number of points that a team receives for a clean-sheet (points per clean-sheet BL 15/16: 3.38). The clean-sheet points of a team have to be fairly distributed among the players according to their contribution to the clean-sheet result. Defensive actions that contribute to a clean-sheet result are the number of blocked shots, the number of clearances, the number of successful tackles, the number of interceptions and the number of goalkeeper saves. In order to have a fair distribution of clean-sheet points, we calculate which player positions (goalkeeper, defence, midfield and forward) are linked to these defensive actions. The counts of defensive actions in the BL season 15/16 were 2477 for goalkeepers, 23827 for defenders, 8789 for midfielders and 1845 for forwards. We assume (as in McHale et al. (2012)) that the average team consists of one goalkeeper, four defenders, four midfielders and two forwards. By this assumption, we can distribute the clean-sheet points fairly among the team positions be taking into consideration how many contributions have been made and how many players played on a certain position. The resulting weights for the distribution of clean-sheet points are 0.21 for goalkeepers, 0.13 for defenders, 0.05 for midfielders and 0.04 for forward players. Exactly the same weight allocation was found in the Premier League by McHale et al. (2012) (except that the forward weight was 0.03, but this might as well be due to rounding). Of the 3.38 points that are rewarded for a clean-sheet result, a goalkeeper receives $0.21 \times 3.38 = 0.71$, a defender 0.44, a midfielder 0.17 and a forward 0.14 points on clean-sheet subindex 6.

3.2.7 Final PPI Score

The final PPI score is a weighted sum of all six subindices. In the original article of McHale et al. (2012) the weighted PPI looks as follows:

$$PPI = 100 \times (0.25I_1 + 0.375I_2 + 0.125I_3 + 0.125I_4 + 0.0625I_5 + 0.0625I_6)$$

where I_1 to I_6 represent the points on subindex 1 to 6 and PPI the final PPI score. The authors multiply the final PPI by 100 because it was found undesirable to have to report ratings with two decimal places. McHale et al. (2012) chose these weights to satisfy the desire of their clients *"to have an index that produced stable ratings, but that was varied enough from week to week to generate discussion among fans, the media, and pundits."* Further, they explain that subindices 2 to 6 can be seen as the stabilizing components of match outcome subindex 1, because subindex 1 can vary a lot from game to game. Lastly, McHale et al. (2012) explain that each subindex was constructed to have the property that the total subindex points are approximately equal to the total league points that can be won in a season (BL 15/16: 847 league points). McHale et al. (2012) reason that with this property, a player's points represent his/her share of league points won by the team. For the BL season 15/16, the total points on match outcome subindex I_1 and point-sharing subindex I_2 both equal exactly 847 points, on appearance subindex I_3 they equal 846 points (due to rounding), on goal-scoring index I_4 = 819 points, on assists index I_5 = 552 points and on clean-sheet index I_6 = 925 points. The total points on assists index I_5 are less than the other subindices because an assist gives an equal number of points as a goal (0.94), but because there were fewer assists (283) than goals (866), the summed up assist I_5 subindex points are less than 847. The total points of 925 on the clean-sheet subindex I_6 came about due to the distribution of the clean-sheet points among the team members proportionally to their contribution. In McHale et al. (2012), the clean-sheet points are distributed under the assumption that there is one goalkeeper, four defenders, four midfielders and two forwards. We adopted this team position distribution to develop exactly the same index for the Bundesliga as describe in the original article. However, the actual or empirical distribution of team positions with regard to the BL season 15/16 shows that there were on average one goalkeeper, 4.9 defenders, 3.5 midfielders and 1.5 forwards on the pitch. This means that the share of clean-sheet points that defenders receive should be lower when the empirical distribution of team positions is considered (and the share of midfielders and forwards higher) because there are more defenders in the Bundesliga, 4.9 instead of 4, and therefore the points should be divided by more defending players (0.13 weight with 4-4-2 distribution and 0.09 weight with empirical 4.9-3.5-1.5 distribution). The clean-sheet subindex I_6 has an exceeding number of 925 total points due to the over-weighing of defenders in the distribution of the clean-sheet points (with empirical weighing, the I_6 total points are 802). In an improved version of the PPI rating system, it could be considered to distribute the clean-sheet points according to the empirical team position distribution.

Software and Code to Estimate PPI Model The match outcome subindex 1 derivatives of the expected number of league points with respect to the player actions were computed with the software *Wolfram Mathematica version 10.3*. The computed derivatives of all relevant player actions were saved on the hard disk in a text file (`https://github.ugent.be/jklaiber/ratingsystem/tree/master/derivatives`). The code to create the PPI ratings can be reached via URL `https://github.ugent.be/jklaiber/ratingsystem/blob/master/model_building.R` and was constructed with the statistical software *R version 3.3.0* (R Core Team, 2014). We created a script that reads in the derivative text files, changes mathematical symbols from *Mathematica* to *R* representation and solves the derivative equations for a given player action. The code is listed on `https://github.ugent.be/jklaiber/ratingsystem/blob/master/derivativeFunction.R`.

3.3 PPI Part - Results

3.3.1 Match Outcome Subindex 1

Table 2 lists the best-scoring players on match outcome subindex 1. The table consists exclusively of defenders and midfielders. This is not an unexpected finding because there are three factors that cause a high individual score on match outcome subindex 1. These factors are a high number of dribbles, crosses and passes. Defenders and midfielders are the player positions that engage the most in dribbles, crosses and passes because their task is to organize the defence and create situations that allow supplying balls to strikers. Not surprisingly, the task of the majority of match outcome subindex 1 Top 20 players within their teams is to build up the soccer game from behind. The absence of forwards and goalkeepers can be explained by the constant shot effectiveness that was not allowed to be influenced by game, team or home/away effects. The shot effectiveness is meant to capture the ability of players to direct a shot on the target, to block shots and of goalkeepers to save the ball. McHale et al. (2012) hold the shot effectiveness constant which means that directing a shot on the target, blocking and saving effects cannot influence the shot effectiveness and in turn cannot influence match outcome subindex 1. However, the goal-scoring subindex 4 and the clean-sheet subindex 6 are meant to balance the final PPI score in favour of forwards and goalkeepers, respectively. An interesting observation is that predominately Bayern Munich, Dortmund and Gladbach players occupy the Top 20 table of match outcome subindex 1. These teams are known to have high ball possession, on average 66.4% (1^{st}), 61% (2^{nd}) and 54.6% (4^{th}) respectively in the 15/16 season, which is typically associated with a high number of passes and crosses *("one-touch-football")* and consequently with a high score on match outcome subindex 1.

Name	Team	Position	Crosses	Dribbles	Passes	Subindex 1
David Alaba	Bayern Munich	Defence	21	15	2450	7.51
Mats Hummels	Dortmund	Defence	3	23	2367	7.33
Julian Weigl	Dortmund	Defence	2	25	2341	7.19
Xabi Alonso	Bayern Munich	Defence	53	4	2341	7.18
Granit Xhaka	Gladbach	Midfield	2	29	2316	7.13
Ilkay Gündogan	Dortmund	Midfield	17	55	2022	6.27
Arturo Vidal	Bayern Munich	Midfield	35	6	2028	6.23
Philipp Lahm	Bayern Munich	Defence	24	11	1989	6.08
Sokratis	Dortmund	Defence	3	10	1874	5.79
Andreas Christensen	Gladbach	Defence	0	7	1855	5.70
Pascal Groß	Ingolstadt	Midfield	218	19	1808	5.70
Naldo	Wolfsburg	Defence	2	11	1844	5.57
Rafinha	Bayern Munich	Defence	36	12	1805	5.55
Lewis Holtby	Hamburg	Defence	75	44	1714	5.33
Niklas Süle	Hoffenheim	Defence	0	5	1706	5.21
Joel Matip	Schalke	Defence	2	12	1707	5.17
Vladimir Darida	Berlin	Midfield	97	16	1666	5.14
Dante	Wolfsburg	Defence	0	3	1701	5.13
Thiago Alcántara	Bayern Munich	Midfield	30	49	1572	4.88
Salif Sané	Hannover	Defence	11	33	1579	4.83

Table 2: PPI match outcome subindex 1 scores of the Top 20 players of the 15/16 Bundesliga season. Name, team, position, along with the absolute scores of three player actions that determine the subindex 1 points, crosses, dribbles and passes, and the final subindex 1 points in descending order are listed in this table.

3.3.2 PPI Rating Top 15

The Top 15 players with the highest PPI score of the Bundesliga season 2015/2016 are depicted in Table 3. The list seems to be a good mix between different positions and teams. Bayern Munich and Dortmund players are overrepresented in the Top 15 due to the good performance of these two teams in the BL season 15/16. This is expressed by high scores on point-sharing subindex 2 (league points won × time on pitch) which is the subindex with the largest weight on the PPI final score. The absence of any Leverkusen players (highest ranked Leverkusen player: 19. Javier Hernandez), the team that finished third in the 15/16 season, is somewhat unexpected. A possible explanation is that there are no single Leverkusen player that performed exceptionally well but that all of the team's players were about equally good. The Top 15 PPI list is dominated by forward players because of the high scores on the scored-goals subindex 4. For every goal 0.94

points are awarded on scored-goals subindex 4 which results in higher total scores on this than on other subindices. In addition, scored-goals subindex 4 significantly contributes to the final PPI score with a weight of 0.125. The good PPI scores of Lewandowski, Müller, Aubameyang and Kalou are predominately caused by the high number of goals that these players scored. The large gap between Lewandowski and the successive players of nearly 100 PPI points is related to the high number of 30 goals that he scored in the 15/16 season. The good position of Mikhitaryan, Raffael, Kagawa and Stindl are due to their goals and their ability to provide assists. The defensive players Alaba, Neuer, Alonso and Lahm are in the Top 15 due to their high score on match outcome subindex 1 *("they organize the game from behind")* and high scores on the clean-sheet subindex 6.

Name	Team	Position	Rated Games	Match Outcome Subindex 1	Point-Sharing Subindex 2	Appearance Subindex 3	Goal-Scoring Subindex 4	Assist Subindex 5	Clean-Sheet Subindex 6	PPI Rating
	PPI Subindex Weights:			0.25	0.375	0.125	0.125	0.0625	0.0625	×100
Robert Lewandowski	Bayern Munich	Forward	32	2.37	7.00	3.71	28.39	1.89	2.59	751
Thomas Müller	Bayern Munich	Midfield	31	3.29	6.19	3.28	18.93	4.73	3.97	646
Pierre-Emerick Aubameyang	Dortmund	Forward	31	1.62	5.55	3.44	23.66	4.73	1.51	626
Henrikh Mkhitaryan	Dortmund	Forward	31	4.27	5.97	3.59	10.41	14.19	2.05	607
David Alaba	Bayern Munich	Defence	30	7.51	6.86	3.50	0.95	0.00	7.14	545
Manuel Neuer	Bayern Munich	Goalkeeper	34	3.42	7.89	4.22	0.00	0.00	14.51	525
Mats Hummels	Dortmund	Defence	30	7.33	5.68	3.48	2.84	1.89	5.40	521
Raffael	Gladbach	Forward	31	3.71	4.25	3.72	12.30	9.46	1.11	518
Shinji Kagawa	Dortmund	Midfield	29	4.50	5.03	3.01	8.52	6.62	2.32	501
Arturo Vidal	Bayern Munich	Midfield	30	6.23	5.18	2.86	3.79	4.73	5.12	495
Xabi Alonso	Bayern Munich	Defence	26	7.18	5.15	2.71	0.95	1.89	7.16	475
Philipp Lahm	Bayern Munich	Defence	26	6.08	5.85	3.06	0.95	0.95	6.45	468
Andreas Christensen	Gladbach	Defence	31	5.70	5.01	3.90	3.79	0.95	3.05	452
Lars Stindl	Gladbach	Forward	30	4.42	4.28	3.62	6.62	7.57	0.84	452
Salomon Kalou	Berlin	Forward	32	2.12	4.06	3.48	13.25	1.89	1.91	438

Table 3: The Top 15 players of the Bundesliga season 15/16 according to the final PPI score with name, team, position and the number of games that they have participated in. Scores on subindex 1 to 6 are listed in the table as well as the weights that make up the final PPI score *("PPI Subindex Weights")*. The light red shaded cells indicate subindex points of players that are exceptionally high and contributed considerably to their high overall PPI rating scores.

3.3.3 Rating Distribution

Figure 2 shows the distribution of PPI ratings for the Bundesliga season 2015/2016. The plot on the left depicts the PPI rating distribution of all players and on the right only for those players that played in more than ten season games. The large bar on the left side of the left graph represents the ratings of substitutes and of players that have not made it into the squad at all. When comparing the two plots in Figure 2, we see that more than 200 BL players have appeared in less than ten out of 34 games. These players have a low PPI score because of low ratings on the point-sharing and appearance subindices that directly translate (successful) playing time on the pitch into PPI ratings and the goal-scoring, assist and clean-sheet subindices 4, 5, and 6 that indirectly result in a low rating because there is less time to make goals, assists and clean-sheet results. In total 543 players had been registered by the Bundesliga clubs to compete in the 15/16 BL season. These players engaged on average in 15.62 (Standard Deviation (SD): 11.6) league games (23.61 (SD: 7) in the right distribution). The PPI distribution on the right is made up of players for whom we can expect to have a reliable performance rating. The shape of the rating distribution on the right seems to follows a Gaussian curve with a slight skew to the right that is caused by a few highly rated players. These outliers are caused by the many goals that these players scored as explained in the *"PPI Rating Top 15"* paragraph.

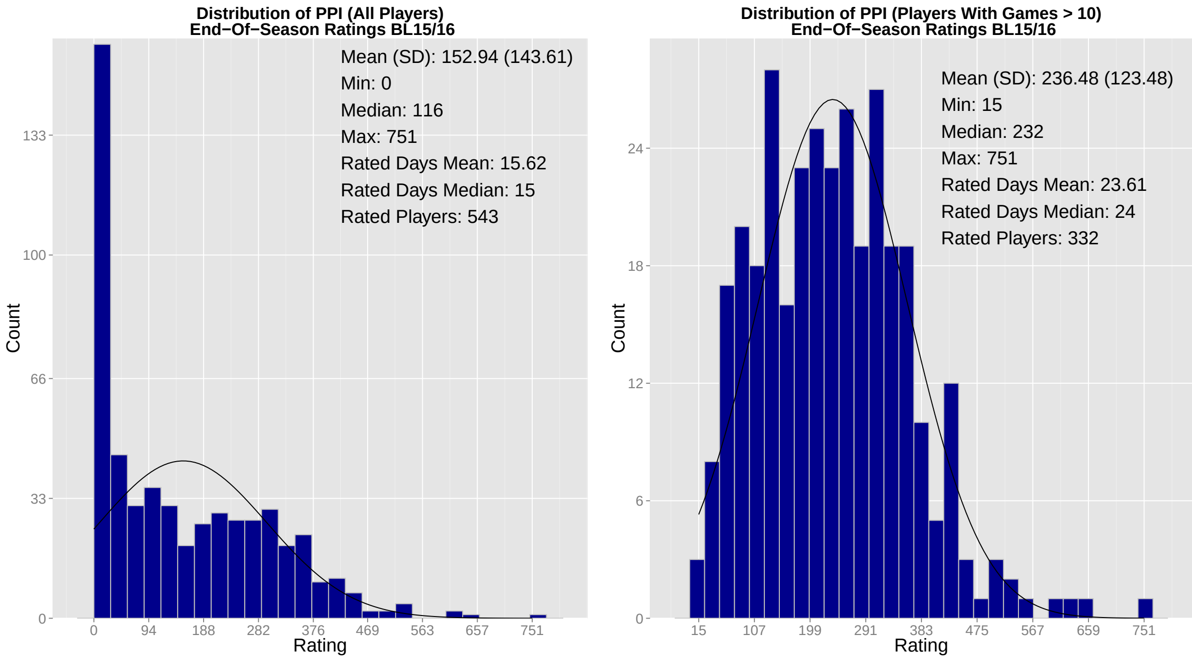

Figure 2: The PPI rating distribution of the Bundesliga season 15/16 for all Bundesliga players is depicted on the left and the PPI rating distribution for Bundesliga players who played in more than ten games is depicted on the right. Distribution characteristics are displayed as text on the plot and the black line indicates the best-fitting normal curve.

3.4 PPI Part - Discussion

We have applied the Player Performance Index rating system successfully to players of the Bundesliga season 2015/2016. Based on the obtained results we have the impression that the PPI rating system can provide a fair, adequate and objective performance rating. The results of the match outcome subindex 1 indicate that the influence of midfielders and defenders is well captured but to a lesser extent the influence of strikers and especially goalkeepers. In an extended version of the PPI, constant shot effectiveness estimate could be replaced by game- and team-specific estimates to better incorporate the effect of forward and goalkeeper actions on the match outcome. A practical difficulty that will arise from using game-by-game shot effectiveness estimates is that the double Poisson model, that links the player actions to the teams' expected league points, would have to be derived for both teams per game. Keeping in mind that we needed a considerable amount of time with a fast and efficient math software program to obtain the extensive derivations for a constant shot effectiveness estimate, a continuous re-derivation of the double Poisson model would computationally be very intensive. Even in a situation in which the derivation could be done with the statistical software R, the running time of such a script to obtain end-of-season PPI scores would be immense.

Critics might suggest that a weakness of the implemented PPI rating system is that strikers get too good PPI ratings because of a high score on the goal-scoring subindex 4. There are two possible solutions to solve this issue. Either the weighing of subindices could be adapted to reduce the weight of goal-scoring subindex 4 in favour of a higher weight of clean-sheet subindex 6. The two subindices 4 and 6 could both be seen as equally important in terms of winning league points because a high number of goals as well as receiving no goals do in many cases lead to winning league points. A suggestion is to weigh scored-goals subindex 4 and clean-sheet subindex 6 equally heavy with a weight of 0.09375 each to make the goal-scoring subindex less and the clean-sheet subindex more important. Although the weights of the two subindices would be equal, the goal-scoring index would still have a slightly higher influence on the PPI final score because a higher total score can be obtained on goal-scoring subindex 4 in comparison to clean-sheet subindex 6. The second solution could be to rescale the awarded points on the goal-scoring index 4. Whereas the highest number of points on any subindex lie around ten, the goal-scoring points on subindex 4 can potentially be three times as high (as it is the case with Lewandowski). A consideration to get a more balanced index is to rescale the scored-goals subindex 4 points to have an upper limit that is similar to the assist subindex (maximum in BL 15/16: 14.19) and clean-sheet subindex (maximum in BL 15/16: 14.51) range. Rescaling of any subindex could make it necessary though to reweigh the subindices with respect to the final PPI score to get a fair, valid and adequate final performance score.

Another weakness of the PPI rating system might be that injured players, players in teams with frequent rotation of the starting squad or players in general that play in fewer games than others are disadvantaged because they get few points on the point-sharing and appearance subindices and indirectly on the other subindices. The PPI rating system is an accumulative rating system that rewards players for participating in as many games as possible. When players participate in a similar number of games per season as most players in the Top 15 PPI Table 3, comparison of player performances is valid because they have approximately equal chances to get a certain performance rating. However, when players participate in considerably less games, it is impossible for them to receive the same PPI score than a player with equal (or slightly lower) ability that participated in all league games. An ad hoc solution to this problem is to only compare players that played in a similar number of games. However, this solution is suboptimal as it limits the potential of the PPI rating system. A better solution would be to switch to a more adjustive style rating system in which players receive a default rating that is adjusted upwards if they play a lot and adjusted downwards only in the case that they participate in significantly less games than the other players.

4 Rating System Comparison

4.1 Other Rating Systems

4.1.1 Kicker Rating System

The Kicker is Germany's leading sports magazine in soccer and was founded in 1920. Kicker issues a paper magazine twice a week and has a frequently visited online platform, reaching both together about 3.4 million readers (Kicker-Sportmagazin, 2006). Since 9$^{\text{th}}$ May 1963, Kicker publishes performance ratings usually on the Monday after a Bundesliga weekend. The player ratings of each game are created by a small group of Kicker experts that are watching the game live in the stadium and asses the players' performances. The ratings are discussed again after the match with other Kicker experts and are adjusted in case of a major dispute (Traina, 2013). Although it is not known how exactly the Kicker ratings are developed, it is said that the ratings are a mix between statistics like shots on the goal, assists and pass accuracy, among others and evaluations of how important certain events were for the game outcome. For example, the game events of receiving a red card early in the game or causing a penalty by violent conduct influence the rating more negatively than receiving a red card in the 90$^{\text{th}}$ minute or causing a penalty due to a last minute tackle on a goal-scoring forward, respectively (Traina, 2013). The Kicker experts who rate a game rotate constantly to prevent that subjective impression and affiliations with a soccer team have an impact on the ratings. A player has to be for at least 30 minutes on the pitch to receive a Kicker rating. This requirement ensures that the ratings are reliable and are not based on a few player actions in a short time (Kicker Online, 2013). The Kicker applies an adjustive rating system in which every player starts off with a default rating of 3.5 (Jagodzinska, 2013). During the game, players receive positive (negative) points for actions that influence the match outcome positively (negatively), for example one point for an assist and two points for goalkeepers who concede no goals

(Kicker Online, 2013). Players receive two points for being nominated for the starting squad. Player positions also play a role in the rating process. For instance a forward player receives three points for scoring a goal but a defender receives five points, because defenders are much less likely to score a goal (Traina, 2013). After the points of a player have been weighed by the importance of the action with regard to the game, they are summed up to calculate the final rating. The summed up points have a linear association with the final rating were plus ten points equal the best final grade of a one and minus ten points the worst final grade of a six with intermediate steps of 0.5 (Kicker Online, 2013). This scale corresponds to the school grade scale in Germany. When players have received the same number of positive and negative points, resulting in a summed up score of zero, they receive a final rating of 3.5 which is equal to the start rating. This corresponds to a neutral impact on the game, neither helping nor hindering the team's play.

4.1.2 Bild Rating System

The daily newspaper Bild is the best-selling German tabloid with 1.8 million sold papers and a reach of 12.3 million readers. It was founded in 1952 and is since then issued by the Axel Springer media company (Bild Zeitung, 2003). On 9$^{\text{th}}$ May 1963, the same date as the Kicker magazine, Bild (in their Sunday issue called *"Bild am Sonntag"*) started to publish ratings for Bundesliga players' performances. Bild sends editors of their sport section to every game to come up with the player ratings. The Bild uses an adjustive rating system that gives every participating player a rating of four to start off with. Players depart from this initial rating according to their performance. Not much more is known about how the ratings come about in detail. A soccer magazine named *"11 Freunde"* published an interview with a Bild sport editor called Müller about the Bild ratings on 10$^{\text{th}}$ August 2009 (Köster, 2009). Müller claimed that the Bild ratings consider *"the entire performance, which means among others goals, assists and won/lost tackles"*. He continued explaining that *"in case a player performs very badly and loses 99% of his tackles but scores two goals, this player will receive a good rating"*. Which according to him meant, on the other hand, that *"a player that scores twice but sees a red card for spitting (on opponent players) will receive a bad rating"*. The best player rating of the Bild newspaper is a one and the worst a six, equal to the final ratings of the Kicker, but without the intermediate 0.5 steps. There has been critic on the Bild ratings from players, the media and supporters (Köster, 2009). In the era of Lother Matthäus, a famous German central midfielder playing for Bayern Munich and the national team, a Bild reporter (named Ruiner) admitted that, *"It's a giving and taking. If you reveal internal matters of the team you will receive a good rating even though you may have performed badly"* which partly explains why Matthäus received excellent ratings, even for some games in which he was a substitute player. Other critics pointed out that the

Bild sport editors have their favourite players that always receive good ratings and that bad team performances often result in the same low rating for every player of the team. In addition, they were criticized for departing, not only once, from their rating scale by rating players with a seven (e.g. for the entire German team after not surviving the group stage of the 2000 European championship in Portugal) which was seen as dubious and untrustworthy by other actors in Bundesliga soccer rating (Köster, 2009). In the comparison of the different rating systems, we will be attentive to any favourite player or team pattern and to whether any performance ratings lie outside of the one to six rating scale.

4.1.3 Comunio.de User Rating System

Comunio.de[3] is an online soccer manager game in which communities of users are forming teams by buying and selling Bundesliga players virtually. Community members receive points for their virtual team depending on the performance of the real players on the pitch. Each player of a community begins with some play money and/or an initial team of 15 players. Users can sell or buy players of other users (within their community) for a price that is determined by supply and demand. Points are earned when the users' players perform well, for instance when one of their players scores a goal. After every Bundesliga matchday, performance points of a user's team are summed up and added to the user's total points (Loschek, 2000). The performance points are determined by a performance rating of the spox.com website, a partner website of Comunio.de. Spox.com does rate player performances similar to Kicker and Bild on a one to six grade scale with intermediate steps of 0.5 and a start rating of 3.5 (Spox.com, 2015). At the end of the season, the user with the most points wins. Since Comunio.de was founded in 2000, it attracted a steadily increasing number of users with roughly 610,000 in 2013 (Comunio, 2008). Playing Comunio.de successfully asks for knowledge about Bundesliga players and teams in order to be able to predict how well they will play. Comunio.de users can be regarded as pseudo-experts. They are pseudo-experts, because in contrast to the experts that rate games for Kicker and Bild, the users are typically non-professionals that follow soccer events in part time, as their hobby. Comunio.de users can rate Bundesliga players after each matchday and the average rating of all users is published on the website spox.com. The players are rated on a one to six scale with one being the best and six the worst rating with intermediate steps of 0.5 (the same scale as spox.com applies for their own rating).

A fact that should be kept in mind is that the Comunio.de user rating might be influenced by the spox.com rating which is released immediately after a game. The

[3]We will refer to Comunio.de users as *"users"* or *"Comunio.de users"*

spox.com performance rating is listed just next to the button that the Comunio.de users use to submit their own rating. The spox.com performance rating is initially covered and has to be *"uncovered"* by a click, therefore it is difficult to determine how many Comunio.de users *"uncover"* the spox.com rating before they submit their own rating. To conclude, although the Comunio.de users can choose a rating between one and six without any suggestion of a default rating, they could be influenced by the (default) rating of spox.com that is initially set to 3.5 and departs from there depending on the performance of the player.

4.2 Comparison Part - Methods

4.2.1 Data Collection

Kicker Data Collection For the Kicker data collection we used the Kicker website as a starting point because it contains links with information about the 18 Bundesliga teams of the current season[4]. This website contains the link to the *"Kader"* (in English: *"squad"*) section of every team. On the *"Kader"* information website, all players of a team are listed with their name, position, number of appearances, goals and the Kicker end-of-season rating. The Kicker rating scraping script is online available on `https://github.ugent.be/jklaiber/ratingsystem/blob/master/KickerScraper.py`. The script was run on 12.06.2016 to retrieve the end-of-season Kicker ratings.

Bild Data Collection The Bild player ratings were obtained in a similar manner as the Kicker ratings. The Bild Bundesliga website shows the emblems of all current BL teams[5]. These team emblems *"contain"* the links to the individual team websites. At the bottom of each team website, a table is displayed which entails a tab that is called *"Kader"*. This tab lists information about all players. From this *"Kader"* tab, the player's name, position, appearances, goals and the end-of-season Bild rating were scraped. Information about 540 BL players was saved in a CSV file on 12.06.2016. The Bild rating scraper script is online available on `https://github.ugent.be/jklaiber/ratingsystem/blob/master/BildScraper.py`.

Comunio.de User Data Collection The ratings of the Comunio.de members cannot be accessed via comunio.de but instead from the spox.com website. In contrast to the Kicker and Bild ratings, there is no aggregated end-of-season rating available of the Comunio.de users. Instead, an overview page of every matchday has to be accessed to

[4] A screenshot of the Kicker.de overview page is accessible via the URL `https://github.ugent.be/jklaiber/ratingsystem/blob/master/data_collection_kicker.png`

[5] A screenshot of the Bild.de overview page can be reached via the URL `https://github.ugent.be/jklaiber/ratingsystem/blob/master/data_collection_bild.png`

scrape the ratings for each game during the season[6]. The overview pages of the different matchdays have not the same structured URL but most of the matchday overview URLs begin with `http://www.spox.com/de/sport/fussball/bundesliga/saison2015-2016/spieltag-...`, continue with the matchday number (for example '3') and end with `...-spieltag.html`. The different URLs to get to the matchday overview page are listed on `https://github.ugent.be/jklaiber/ratingsystem/blob/master/UserScraper.py` within the Comunio.de user rating script. On each matchday overview page, all players that played in a match are listed with their name, team, the Comunio.de user rating, the number of users who rated the player and the spox.com expert rating (which was not further used). On 23.06.2016, ratings for all players of the BL season 2015/2016 were obtained and the average Comunio.de user rating was calculated as the end-of-season Comunio.de user rating.

Software for the Rating Comparison The Kicker, Bild, Comunio.de and whoscored.com ratings were collected with the software *Python 2.7* similar to data collection in the first part. Also the software testing framework *"Selenium"* version 2.53 was used again to manoeuvre through the websites and partly *"Selenium"* and partly *"Beautiful Soup"* version 4.4.0 were applied to extract and process the web data. The data of the different performance ratings was standardized in terms of player names, team names and positions on the pitch (goalkeeper, defence, midfield, forward). The comparison of rating systems was done with the statistical software R (R Core Team, 2014), the script is available on `https://github.ugent.be/jklaiber/ratingsystem/blob/master/rating_comparison.R`.

4.3 Comparison Part - Results

4.3.1 Rating Distributions

In the results section of the first part, the rating distribution of the PPI end-of-season ratings was shown. Figure 3 shows the distributions of the Bild, Kicker, Comunio.de user and whoscored.com end-of-season ratings. Note that for the Bild, Kicker and Comunio.de users rating distributions, a lower rating indicates a better performance and for the whoscored.com ratings (as for the PPI ratings) a higher rating represents a better performance. All four different ratings in Figure 3 appear to approximate a normal distribution. The number of players that have been rated during the season varies slightly between rating systems because different criteria on when a player should receive a rating are applied. The same reasoning applies to the observation that the Bild, Kicker and whoscored.com

[6]A screenshot of the spox.com overview page can be accessed via the URL `https://github.ugent.be/jklaiber/ratingsystem/blob/master/data_collection_users.png`

ratings are on average based on the performance in approximately 20 games, whereas the Comunio.de user rating is only based on about 16 games. The Comunio.de user ratings are only given when the players spend more than 20 minutes on the pitch (otherwise they are not listed on the rating website) to get stable performance ratings which explains the fact that less games are rated on average. The mean of the number of Comunio.de users that participated in rating players was 60 (SD: 41). There are many players that were rated by a few users (median is 49) and some popular players that were rated by a lot more users.

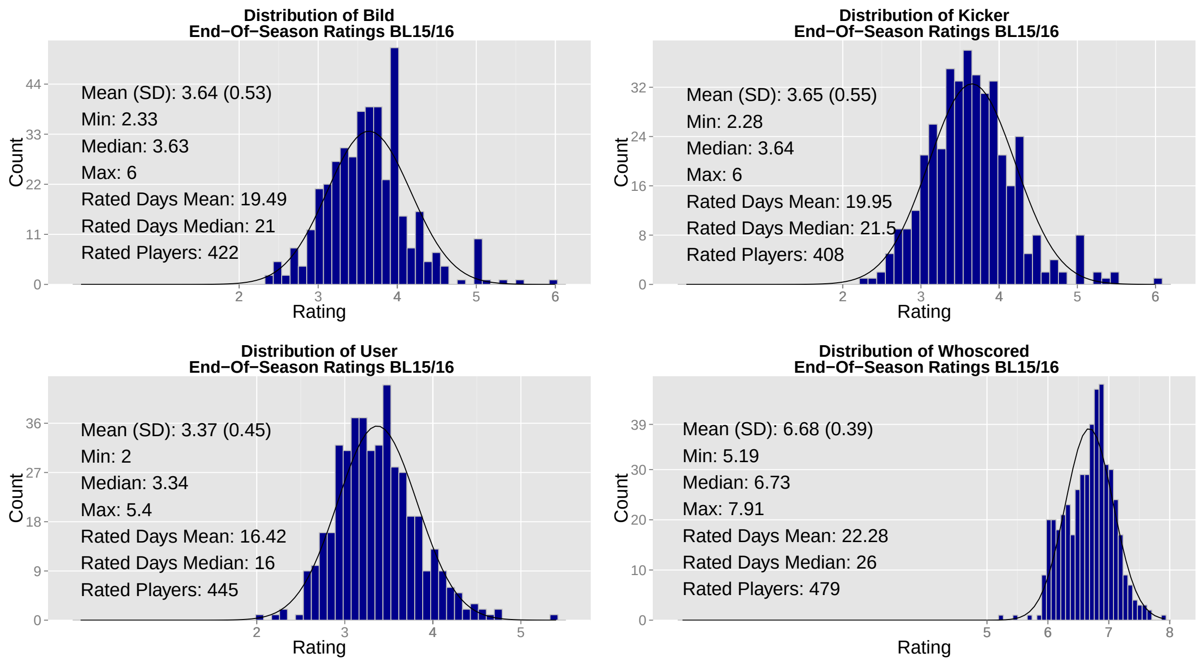

Figure 3: The distributions of Bild (top left), Kicker (top right), Comunio.de user (bottom left) and whoscored.com (bottom right) ratings for the Bundesliga season 15/16. Distribution characteristics are displayed as text on the plot and the black lines indicate the best-fitting normal curves.

The Bild and Kicker ratings (top panel) show a spike at rating four and 3.5, respectively. These spikes make sense because they show players that neither had a good nor a bad impact on the game. Among the Bild, Kicker and Comunio.de user ratings are no ratings below two, the good extreme of the one to six rating scale, whereas there are a few bad ratings that are close to six. The reason might be that a player who performed particularly bad and received a six is sent off to sit on the bench for the next games while a player who plays extra good will play in the next game and will on average get a worse rating than in the begin because of the regression to the mean. In conclusion this means that bad players stay with their bad rating because of a few ratings that are close to six and good players regress towards a rating of about two.

Figure 3 shows that the whoscored.com rating distribution is limited between five and eight. Theoretically the whoscored.com rating system is scaled to allow for values between one and ten, but as it appears, large regions of the scale are not used. We do not know how the whoscored.com ratings are formed, but most probably an average performance will result in a rating of seven, as the distribution mode at seven suggests. One can speculate that a major component of the whoscored.com rating is attributed to the performance of the team, similar to the PPI point-sharing subindex 2 or appearance subindex 3. This could mean that bad individual performances are absorbed by the team performance and do not result in a low rating as they should. One could argue that the limited usage of the scale is due to the aggregation of the end-of-season ratings which reduces the variance of the ratings. However, the distribution of single-game ratings show the same picture of a limited rating scale as the aggregated end-of-season ratings because the single-game ratings also only partly cover the lower end of the scale. The lowest single-game rating given in the BL season 15/16 was a 4.1. The statistic based performance rating algorithm that is applied by whoscored.com appears to be not using the lower end of their rating scale.

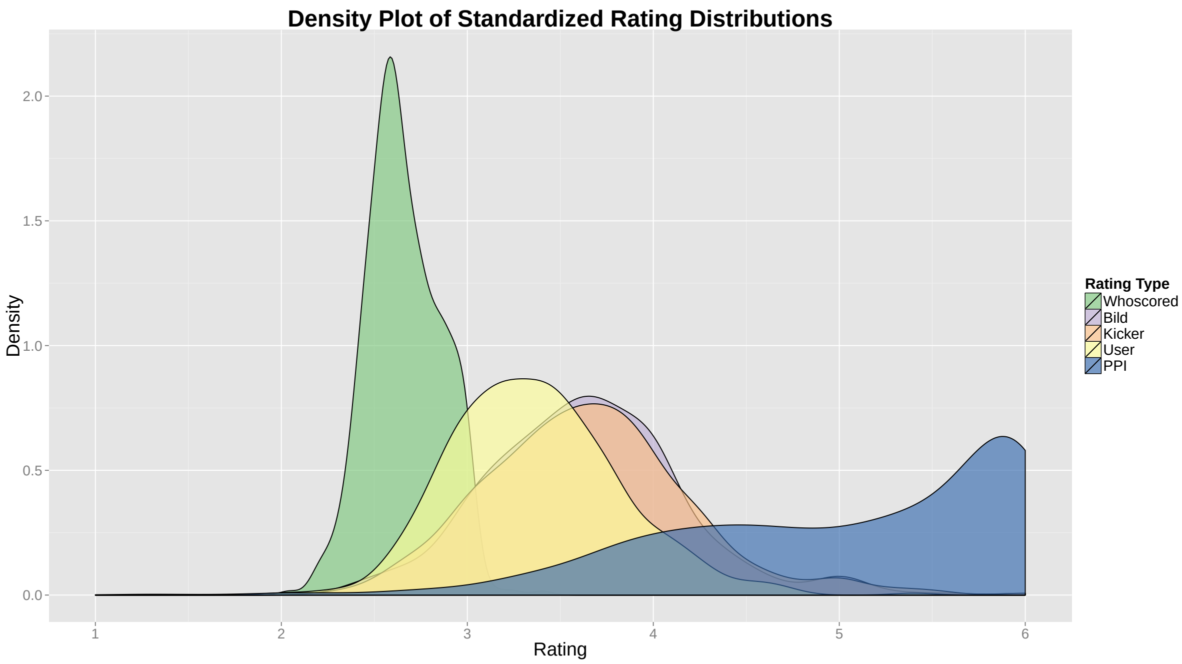

Figure 4: Density distributions of the Bild, Kicker, Comunio.de user, whoscored.com and PPI ratings standardized on a one (best rating) to six (worst rating) scale.

A standardized rating distributions figure was created to be able to better compare the distributions. All ratings were transformed to the one to six scale that the Bild, Kicker and Comunio.de systems use. The PPI does not have a theoretical upper limit and therefore we chose an upper limit of 800 PPI points which seemed to be plausible with regard to the best rated player of the 15/16 season with 751 points. Figure 4 shows the standardized rating density distributions. The Bild, Kicker and Comunio.de density distributions are nicely bell shaped and the Kicker and Bild distributions also overlay each other almost entirely which is an indication that their way of creating a performance rating is similar. The Kicker and Bild rating systems both depart from a fixed starting point and either move towards one if performance was good or to six if it was bad. The fact that sports media, with the Bild and Kicker as the most prominent representatives, are relatively coherent might also partly explain why similar ratings are given. The shape of the PPI distribution with respect to the other distributions is that players start off with zero points and have to play a lot and play good in order to decrease towards a good performance rating.

4.3.2 Top 20 Player Comparison

Table 4 displays the Top 20 players according to the different rating systems that participated in at least ten league games. There is some agreement among rating systems about which players performed well in the 15/16 season. In any rating system Lewandowski, Müller and Mkhitarayan are to find between the top ranked players. Lewandowski was the best goal scorer, Mkhitarayan the best player giving assists and Müller scored many goals and provided many assists. The teams Bayern Munich and Dortmund appear frequently in the Top 20 lists of all rating systems except in the Comunio.de user list. The Comunio.de users have only voted six Bayern Munich and two Dortmund players into the Top 20 in return for players from less successful teams like Bremen, Hamburg, Mainz or Stuttgart. The PPI is less diverse in terms of teams than the other systems because 13 out of the 20 players are either from Bayern Munich or from Dortmund. The two teams won many league points in the 15/16 season and finished with almost 20 league points ahead of any other team. The over representation of Bayern Munich and Dortmund players in the PPI is related to the point-sharing subindex 2 that translates league points directly into PPI points but also related to the other subindices as discussed in the first part of this study.

The Top 20 players can also be compared with regard to the player roles on the pitch. In theory one would expect that any player position should have equal chance to end up in the Top 20 under the condition that the rating system is not biased towards certain positions[7]. In order to test whether the rating systems are fair, we calculated

[7] The assumption is that any player position is equally likely to end up on a certain rank and that one player position group is not better than any other group on average (for example that on average all

	Bild Top 20			Kicker Top 20			PPI Top 20			Comunio.de User Top 20			Whoscored Top 20		
Rank	Name	Team	Role	Name	Team	Role	Name	Team	Role	Name	Team	Role	Name	Team	Role
1.	F. Ribéry	FCB	Mid	F. Ribéry	FCB	Mid	R. Lewandowski	FCB	For	R. Fährmann	S04	Gk	H. Mkhitaryan	BVB	For
2.	R. Lewandowski	FCB	For	H. Mkhitaryan	BVB	Mid	T. Müller	FCB	Mid	R. Lewandowski	FCB	For	Raffael	BMG	For
3.	T. Müller	FCB	For	J. Martnez	FCB	Def	PE. Aubameyang	BVB	For	M. Hitz	AUG	Gk	Caiuby	AUG	Mid
4.	A. Hahn	BMG	Mid	R. Lewandowski	FCB	For	H. Mkhitaryan	BVB	For	A. Christensen	BMG	Def	R. Lewandowski	FCB	For
5.	H. Mkhitaryan	BVB	Mid	A. Hahn	BMG	For	D. Alaba	FCB	Def	D. Costa	FCB	Mid	D. Costa	FCB	Mid
6.	J. Martnez	FCB	Def	R. Adler	HSV	Gk	M. Neuer	FCB	Gk	J. Kimmich	FCB	Def	K. Coman	FCB	Mid
7.	RR. Zieler	H96	Gk	M. Hitz	AUG	Gk	M. Hummels	BVB	Def	Vieirinha	VFL	Def	T. Müller	FCB	Mid
8.	M. Hitz	AUG	Gk	O. Baumann	HOF	Gk	Raffael	BMG	For	R. Adler	HSV	Gk	L. Stindl	BMG	For
9.	A. Robben	FCB	For	L. Karius	M05	Gk	S. Kagawa	BVB	Mid	Z. Junuzovic	WB	Mid	I. Gündogan	BVB	Mid
10.	M. Hummels	BVB	Def	RR. Zieler	H96	Gk	A. Vidal	FCB	Mid	D. Alaba	FCB	Def	A. Vidal	FCB	Mid
11.	K. Kampl	LEV	Mid	T. Müller	FCB	Mid	X. Alonso	FCB	Def	L. Karius	M05	Gk	EM. Choupo-Moting	S04	Mid
12.	I. Gündogan	BVB	Mid	X. Alonso	FCB	Mid	P. Lahm	FCB	Def	L. Hradecky	FF	Gk	M. Hummels	BVB	Def
13.	D. Costa	FCB	Mid	J. Boateng	FCB	Def	A. Christensen	BMG	Def	H. Mkhitaryan	BVB	For	G. Xhaka	BMG	Mid
14.	PE. Aubameyang	BVB	For	R. Fährmann	S04	Gk	L. Stindl	BMG	For	K. Coman	FCB	Mid	Thiago	FCB	Mid
15.	J. Kimmich	FCB	Def	T. Horn	FCK	Gk	S. Kalou	BSC	For	Raffael	BMG	For	J. Hector	FCK	Def
16.	T. Horn	FCK	Gk	B. Leno	LEV	Gk	A. Modeste	FCK	For	A. Robben	FCB	Mid	A. Meier	FF	Mid
17.	M. Ginter	BVB	Def	A. Robben	FCB	Mid	J. Matip	S04	Def	A. Kramaric	HOF	For	J. Matip	S04	Def
18.	L. Karius	M05	Gk	M. Neuer	FCB	Gk	J. Weigl	BVB	Def	F. Kostic	VFB	Mid	PE. Aubameyang	BVB	For
19.	M. Reus	BVB	Mid	D. Costa	FCB	Mid	Chicharito	LEV	For	H. Kiyotake	H96	Mid	P. Djilobodji	WB	Def
20.	A. Vidal	FCB	Mid	K. Casteels	VFL	Gk	M. Reus	BVB	For	I. Gündogan	BVB	Mid	I. Traoré	BMG	Mid

Table 4: Names, teams, and player roles (position) of the Top 20 players of BL season 15/16 according to the Bild, Kicker, PPI, Comunio.de user and whoscored.com rating system. The team name abbreviations are, FCB: Bayern Munich, BVB: Dortmund, BMG: Gladbach, VFL: Wolfsburg, WB: Bremen, LEV: Leverkusen, H96: Hannover, AUG: Augsburg, FCK: Cologne, M05: Mainz, HSV: Hamburg, HOF: Hoffenheim, S04: Schalke, FF: Frankfurt, VFB: Stuttgart, BSC: Berlin and the position abbreviations are, Gk: Goalkeeper, Mid: Midfield, Def: Defence and For: Forward.

Rating System	# Goalkeepers	# Defenders	# Midfielders	# Forwards	p - value
Null-Count (Proportion):	1.53 (0.08)	9 (0.45)	6.54 (0.33)	2.92 (0.15)	
Bild	4	4	8	4	0.038
Kicker	10	2	6	2	<0.001
PPI	1	7	3	9	0.007
Comunio.de User	5	4	7	4	0.013
Whoscored.com	0	4	11	5	0.027

Table 5: The null-count represents the number of players in the Top 20 for the four different player positions that we expect when the position distribution is the same in the Top 20 of a given rating system and among all players. This null hypothesis was tested by means of exact multinomial tests with the alternative hypothesis that the Top 20 position distribution of a rating system is not equal to the position distribution in the entire player population. The resulting p-values for these tests are shown in the leftmost column, a significance level of $\alpha = 0.05$ was applied. The table lists the player position counts for all five rating systems against which were tested.

the distribution of player position among all players of the 15/16 season and found that there are 8% goalkeepers[8], 45% defenders, 33% midfielders and 15% forwards. It was tested whether we find the same proportions of player positions in the Top 20 by means of exact multinomial tests. Table 5 shows the resulting p-values of these tests. Applying a significance level of $\alpha = 0.05$, all tests indicate that the distribution of player positions in the Top 20 of any rating system is significantly different to the player position distribution among all players *("null-distribution")*. We conclude that none of the rating systems is fair with regard to the distribution of player positions in the Top 20 according to these multinomial tests. The Bild ratings are closest to the null-distribution but too many goalkeepers and too few defenders ended up in the Top 20 to be considered fair. The human expert based rating systems Bild, Kicker and Comunio.de have many goalkeepers represented in the Top 20 with Kicker putting goalkeepers on half of all Top 20 ranks. In contrast, the statistic based systems PPI and whoscored.com are fair or even somewhat underrepresented with regard to the number of top ranked goalkeepers. Furthermore, Table 5 makes clear that too few defenders are in the Top 20 and too many forwards compared to the null-distribution. The PPI sticks out with nine out of 20 top ranks that are occupied by attackers. This is probably due to goal-scoring subindex 4 that discriminates positively in favour of goal-scoring forwards as discussed in the first part. The whoscored.com rating on the other hand seems to be in favour of midfielders with eleven top ranks given to midfielders at the expense of only four ranks given to defensive positions (goalkeepers plus defenders). The average rank of the four different player positions in the population of the Top 200 15/16 BL players are depicted in Figure 5 to get a more complete picture of the association between rank and position. The impression with regard to goalkeepers in the Top 20 is confirmed in Figure 5. The goalkeepers in the Kicker ranking have on average low ranks whereas the goalkeepers in the whoscored.com ranking have on average high ranks. The whoscored.com rating bars confirm that offensive forces (midfield and forwards) are ranked more favourable because of low bars at the expense of higher defensive forces bars. With regard to all ranking systems, we see that the average rank of defenders is slightly higher, of the midfielders equal and of the forwards slightly lower than the average rank position one would expect when there would be no association between position and rank. Fair which means that the ranks for a given player position are normally distributed with mean 100 (when the assumption stated in Footnote 7 applies). The green bars of the PPI ranking seem to indicate that it is the fairest ranking with respect to player positions because all green bars are relatively close to the 100 mark. Last but not least, Figure 5 suggests that all ranking systems have difficulties in ranking the performance of goalkeepers because they are either ranked too

goalkeepers do not perform better than all defenders).

[8]We calculated a share of 8% goalkeepers although we would expect to have 9% (1 / 11 = 0.09) because we only take players into consideration who participated in more than ten games.

low or too high. This indicates that it is indeed difficult to evaluate the special role of the goalkeepers, their influence on the team performance and on the match outcome because a goalkeeper's game is completely different to a field player's game. The lesson which we can retain from this is that a good rating/ranking system, no matter if expert or statistic based, must be able to rate the performance of the goalkeeper adequately.

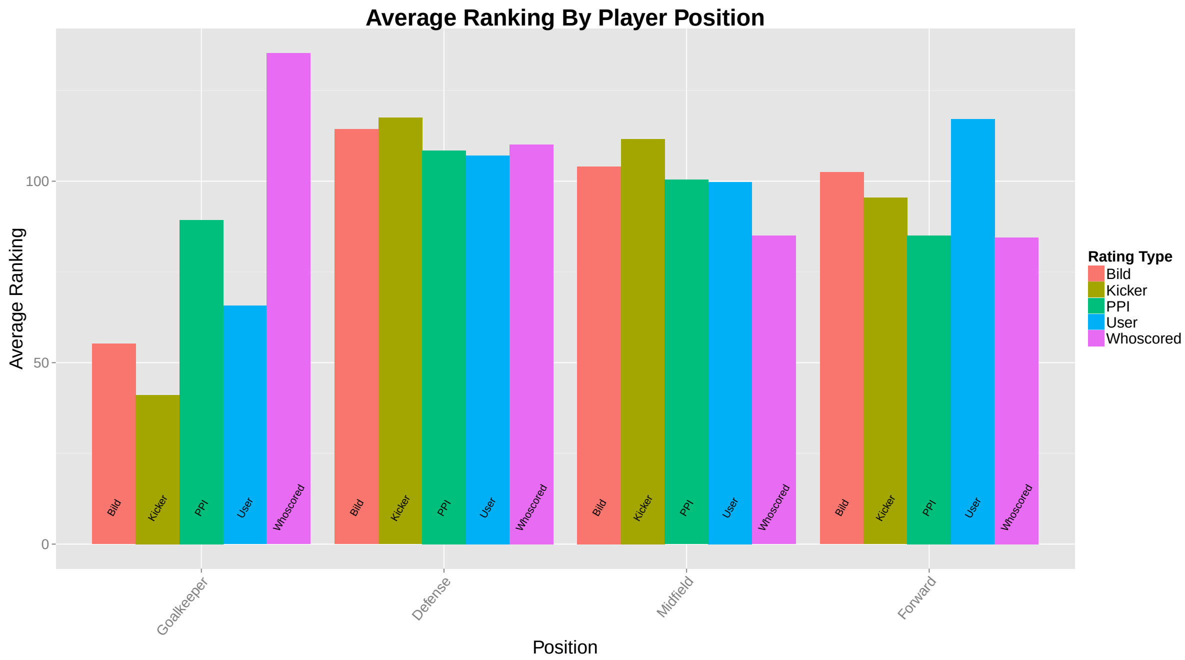

Figure 5: The figure shows the average rank of the four different player positions by the five rating systems in the population of the Top 200 BL players of the 15/16 season that have at least played in ten games. Note that the higher the bar, the worse is the average rank.

4.3.3 Team Captain Comparison

Figure 6 depicts the end-of-season ranks of the 18 players that were appointed as team captains at the beginning of the 15/16 season. The two captains, Lahm and Hummels, of the best performing teams in the 15/16 season ended up with good ranks and little disagreement between the five different rating systems. In contrast, there is no accordance between the ranks of Bender, the captain of Leverkusen which finished third in the league table. The green rectangle of the PPI ranking is far higher up than the other ranking symbols. The reason for this discrepancy is that Bender was injured for several games during the season, which means he received few points on the point-sharing and appearance subindex 2 and 3 of the PPI.

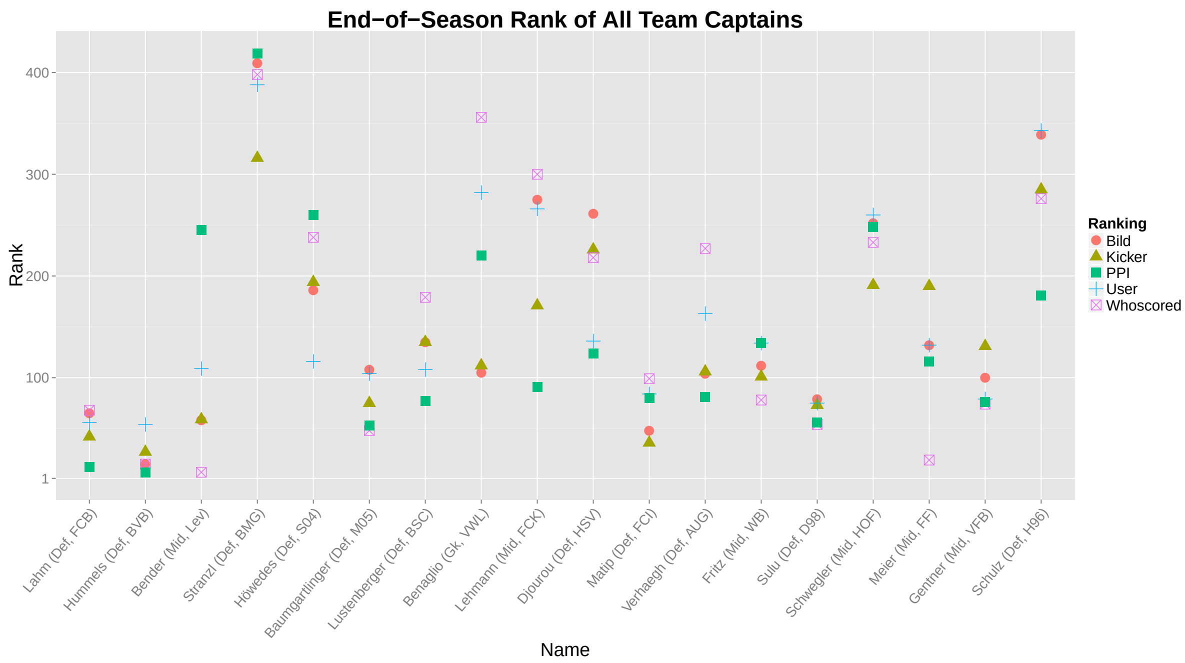

Figure 6: The end-of-season rank of the 18 team members that were named captain at the beginning of the 15/16 season with their name, position and team. Note that the lower the symbols are on the y-axis, the lower is the rank and the better was the performance of the team captain. The players are ordered according to the final position of their team in the league table. The team name abbreviations are, FCB: Bayern Munich, BVB: Dortmund, LEV: Leverkusen, BMG: Gladbach, S04: Schalke, M05: Mainz, BSC: Berlin, VFL: Wolfsburg, FCK: Cologne, HSV: Hamburg, FCI: Ingolstadt, AUG: Augsburg, WB: Bremen, D98: Darmstadt, HOF: Hoffenheim, FF: Frankfurt, VFB: Stuttgart, H96: Hannover and the position abbreviations are, Gk: Goalkeeper, Mid: Midfield, Def: Defence and For: Forward.

A similar situation to the one of Bender applies to Stranzl, who only played in four games during the season. The rating systems agree in placing him on an unflatteringly bad end-of-season rank. The ranking systems are in relative accordance for the Schalke and Mainz captains Höwedes and Baumgartlinger but in disagreement in terms of Lustenberger and Benaglio. The disagreement in the case of Benaglio can be explained with the favourable ranks goalkeeper in general receive in the Kicker and Bild system and unfavourable ranks that they receive in the whoscored.com system. The diverting ranks for Lehman, the team captain of Cologne, are not as clear to explain but are most likely due to him being a player who is not very outstanding on the pitch. As a midfielder, he did not score any goals and only provided two assists during the entire season. This could make him appear like a weak player, in particular for the expert based rating system that might favour outstanding, goal-scoring players more. However, the high score on the match outcome subindex 1 of the PPI suggests that Lehman is important for the Cologne game with many passes and indirectly with a lot of successful tackles and interceptions. The PPI can capture these valuable actions for the Cologne game which might go undetected by the other systems. The same story might play a role with regard to the ranks of Djourou, the Hamburg captain. For Matip, Fritz, Sulu and Schwegler there is relative agreement between rating systems. Verhaegh, team captain of Augsburg and defender, was ranked badly by whoscored.com which could be due to the defender bias of the whoscored.com ranking. He was ranked well by the other systems probably because he is a goal-scoring defender with seven goals in the 15/16 season. Finally, the disagreement about the ranks of Frankfurt's Meier between whoscored.com and the other rankings might also be due to giving offensive players favourable ranks by whoscored.com. Finally, there is agreement about the rank of Gentner but disagreement between PPI and the other rankings about Schulz, the captain of Hannover which finished last in the league table. The explanation of the disagreement between rankings in case of Lehman could also play a role here because, similar to Lehman, Schulz only scored two goals but might have been valuable to the team by organizing the Hannover game with his passes. An interesting observation is that the Kicker and Bild ranks of team captains is often similar, which suggests again that these rating systems might influence each other in one or another way.

4.4 Comparison Part - Discussion

In the rating comparison part we compared two human expert based rating systems, the Bild and Kicker ratings, one community based system, the Comunio.de user rating system and two statistic based systems, the whoscored.com and the self-constructed PPI system on different merits. The comparison between rating distributions was made with the averaged, end-of-season ratings whereas the Top 20 players and team captains were compared based on their ranks.

The comparison of rating distributions revealed that all ratings appear to be normally distributed except the PPI ratings. This was related to the fact that the PPI ratings initial position is a rating of zero whereas the other rating systems have an initial position somewhere in the centre of their respective rating scale. The rating distribution figures clearly showed that the distributions peak at their initial rating value except for the whoscored.com rating for which we do not exactly know what the default rating is. Furthermore it was very clear from the standardized ratings figure that the whoscored.com rating system uses only a small region of its scale, especially the bad rating region seems to be not used by the whoscored.com rating algorithm.

The comparison between the Top 20 ranked players revealed that foremost the Kicker, but also the Bild and Comunio.de user ratings favour goalkeepers but are unfavourable with regard to defenders. All rating systems were judged to be not fair with respect to the distribution of player positions (when the assumption of Footnote 7 applies that each position has the same chances of ending up on a certain rank) in the Top 20 and in the population of all players, yet the Bild ratings were closest to a fair distribution of player positions. We neither found any ratings outside of the scale nor pattern that would suggest that the Bild experts favoured any players or teams as has seemingly happened in the past. The whoscored.com rating system stuck out because it ranks offensive players better than defensive players. This picture was confirmed in the figure of the average rank by player position. All rating systems under examination have difficulties to rank the performance of goalkeepers adequately.

Finally we discussed the end-of-season ranks of the 18 team captains. The comparison of team captain ranks confirmed the pattern of preferring and disadvantaging players because of their playing positions that had already emerged from the Top 20 player comparison. We concluded that the PPI system is better than the other systems in incorporating *"less visible"* players that do not score or give assists but contribute to the performance of their team by increased passing game, crosses and indirectly by many interceptions and successful tackles. We also saw again that one of the major weaknesses of the PPI system is that it disadvantages players that have played in fewer games than other players.

5 Discussion

In the first PPI part of this thesis we built the PPI rating system model and rated the performance of Bundesliga players. We had the impression that the resulting PPI ratings are plausible and the division into subindices that capture different elements of the performance is sensible. However, we also argued that there is space for improvements. Firstly, the PPI system could be made less influential to the number of games that a player participated in. This would allow for rating comparisons between players that have played a different number of matches. Possible solutions for this problem were discussed. A promising solution is to change the PPI from being an accumulative to an adjustive rating system that allows for prior ratings which are adjusted according to the performance. Another weakness of the PPI is the non-variable shot effectiveness that does not take into account the saves, blocks and shots on/off target ratio. While a flexible shot effectiveness estimate would make the rating system more accurate (but with the danger of overfitting), it was also argued that it is technically not possible in the current implementation of the rating system. Moreover it became apparent that the goal-scoring subindex 4 of the PPI does either have to be rescaled or the subindex weights that determine the final PPI score have to be changed in order to reduce the large influence of goals on the PPI index.

The second part of this thesis about the comparison of rating system, in particular between statistic and expert based systems, showed that there is a clear distinction in the distribution of ratings between the accumulative rating system (PPI) and the adjustive systems (all other systems). Whereas the adjustive system distributions resemble a normal distribution, the accumulative system distribution peaks in the bad rating area and decreases continuously towards the good rating area. The other statistic based system, the whoscored.com ratings, was found to make suboptimal use of the full rating scale. In terms of the applied type of rating system and therewith the applied rating scale, the expert based systems are superior to the statistic based system in our opinion. However, the rating system type and rating scale are not factors that are tied to either expert or statistic based systems but could be used in any of the two.

The comparison of the Top 20 players of each rating system revealed that the players and teams that ended up in the Top 20 are plausible with regard to their performance. No distinction was found between statistic and expert systems but between the Comunio.de user and the other systems. The Comunio.de user rating system turned out to be less relying on team performance and more on individual performance because many players of less well performing teams made it into the Top 20, which was not the case in the other rating systems. In terms of player positions all rating system are biased towards specific positions because none of them reflect the general distribution of player positions in the Top 20. It became clear that it is difficult for statistic based as well as expert based systems to evaluate the performance of goalkeepers adequately. In our opinion, the PPI way of assessing goal keeper ability by means of the clean-sheet component (and a variable shot effectiveness in a prospective improved version) is the most promising way of rating the goal keeper performance appropriately.

The most valuable lesson from the comparison of team captains might be that the two statistic based systems, in particular the PPI, are better suited than expert systems to rate players appropriately who are important to the game but do not stand out through goals, assists or other visible actions. The rating of the Cologne team captain Lehman showed this exemplary. The PPI's match outcome subindex 1 is suited to detect passes, crosses and dribbles and reflect them in the performance rating.

Further Suggestions Along with the improvements that could be made with regard to the present version of the PPI, completely new features could be introduced to improve the rating system. New features are of course strictly related to the purpose and application that a rating system is meant to have. The PPI system might be fine to obtain end-of-season performance ratings, but to be able to get a performance rating at any point during the season a strength of schedule adjustment component should be implemented. A strength of adjustment component would correct for the strength of the opposing team and would increase the ratings of players that had to play against strong teams so far in the season and decrease ratings for players that had an easy league schedule. This PPI extension could take place via an additional factor in match outcome subindex 1 or via an additional seventh subindex that controls for opposing team strength. Other adjustments could be interesting as well as for example a league adjustment. If the aim would be to construct a PPI rating system that allows comparing performance between players in different European leagues, a strength of league adjustment should be added. Such a strength of league adjustment could be built based on data from UEFA Champions and Europa league games in which teams of all European leagues compete with each other. An all European PPI rating system would on top of that make it possible to identify talents early because their performance could be rated and compared appropriately, irrespectively of the league they play in at the moment.

The opinion a human expert forms about the performance of a player does not come out of the blue. The experts might observe certain behaviour or player actions and mentally adjust the performance rating accordingly. For further research it would be interesting to try to detect and capture the data and accompanying statistics that are the foundation of the human expert opinion. In a first attempt to see if research in this direction could be fruitful, we downloaded the Kicker ratings for every matchday and related them by a linear regression to the player action statistics (Appendix Table 6). The four player positions have different roles on the pitch and therefore we made four separate regressions. The results look promising and as intuitively expected. It appears that the models can explain a great share of variance in the Kicker ratings. In specific, the Kicker rating of goalkeepers seems to be positively associated with the number and accuracy of saves, interceptions and negatively with the number of big errors and committed fouls. For the other player positions there is a clear association between goals and assists and the Kicker rating. However, there are statistics that influence the defender, midfield and forward ratings differently. The defenders' rating is related to the number of big errors, yellow and red cards, and clearances whereas the forwards' rating depends on the number of times that they are fouled, the dribbles and shots on the target. The rating of midfield players appears to depend on their pass ability, how accurate key passes and passes through the opponent defence are. The venture of trying to capture the (subjective) human expert opinion in terms of data/statistics shows to be interesting and worthy to be continued. It shows that the human expert ratings are less *"subjective"* as one might think because there are to a great extent based on observable behaviour and statistics, but interpreted with the experiences that the experts have. In future research one could attempt to build a statistic based rating model that resembles the rating of human experts as good as possible in order to find out how exactly humans form an expert opinion about performance.

All in all, it seems that the expert based systems, that are predominantly used to rate the performances of players in the Bundesliga, are well suited and relatively fair. Nevertheless, the introduction of a weekly updated statistic based performance rating system would be an enrichment to the Bundesliga, for its supporters, the media and the clubs. Also football manager games like Comunio.de could benefit greatly from applying a statistic based performance rating system to allocate performance points during the manager game. The performance points allocation would be done in a comprehensible, objective and transparent way and lead to fewer discussions among the Comunio.de users about a fair performance points allocation. The PPI in its current form cannot yet be the new statistic based system for the Bundesliga but is very promising if a few improvements are made. The change to an adjustive type system, the reweighing of the goal-scoring and clean-sheet subindices, a more variable shot effectiveness and a strength of schedule adjustment are improvements that should be applied to make the PPI a suitable weekly performance rating system for the German Bundesliga.

References

Bild Zeitung. (2003). *Bild (Zeitung) — Wikipedia, the free encyclopedia.* Retrieved from https://de.wikipedia.org/wiki/Bild_(Zeitung) ([Online; accessed 16-April-2016])

Comunio. (2008). *Comunio — Wikipedia, the free encyclopedia.* Retrieved from https://de.wikipedia.org/wiki/Comunio ([Online; accessed 16-April-2016])

Dunning, E. et al. (1999). The development of soccer as a world game. *Sport matters: sociological studies of sport, violence and civilisation.*, 80–105.

Jagodzinska, G. (2013, July). *Bundesligakicker und "kicker"-Noten — Suite 101.* Retrieved 2016-04-16, from http://suite101.de/article/bundesligakicker-und-kicker-noten-a123175

Keller, J. B. (1994). A characterization of the poisson distribution and the probability of winning a game. *The American Statistician*, *48*(4), 294–298.

Kicker Online. (2013, July). *Bewertungsgrundlagen — Kicker Online.* Retrieved 2016-04-16, from http://www.kicker.de/games/pro/startseite/artikel/367762/

Kicker-Sportmagazin. (2006). *Kicker-Sportmagazin — Wikipedia, the free encyclopedia.* Retrieved from https://de.wikipedia.org/wiki/Kicker-Sportmagazin ([Online; accessed 16-April-2016])

Köster, P. (2009, August). *Montag ist Zeugnistag — 11Freunde.* Retrieved 2016-04-16, from http://www.11freunde.de/artikel/notengebung-im-fussball

Loschek, F. (2000, August). *Rules — Comunio.* Retrieved 2016-04-16, from http://www.comunio.de/rules.phtml

McHale, I. G., Scarf, P. A. & Folker, D. E. (2012). On the development of a soccer player performance rating system for the english premier league. *Interfaces*.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Schultze, S. R. & Wellbrock, C.-M. (2015). 'b-fasst': A statistical and spatial analysis system to evaluate player performance in soccer. *Available at SSRN 2643041*.

Spox.com. (2015, September). *So entstehen die Comunio-Noten — Spox.com.* Retrieved 2016-04-16, from http://www.spox.com/de/sport/fussball/bundesliga/comunio/1509/Artikel/noten-bewertungskriterien.html

Stefani, R., Pollard, R. et al. (2007). Football rating systems for top-level competition: a critical survey. *Journal of Quantitative Analysis in Sports*, *3*(3), 1–20.

Szczepański, Ł. (2008). Measuring the effectiveness of strategies and quantifying players' performance in football. *International Journal of Performance Analysis in Sport*, *8*(2), 55–66.

Traina, D. (2013, January). *Wie entstehen die Noten bei Kicker und Sportal? — Ligainsider.* Retrieved 2016-04-16, from http://www.ligainsider.de/blog/wie-entstehen-die-noten-bei-Kicker-und-sportal/

Appendix

A Flow Diagrams of Relation Between Different Software Functions

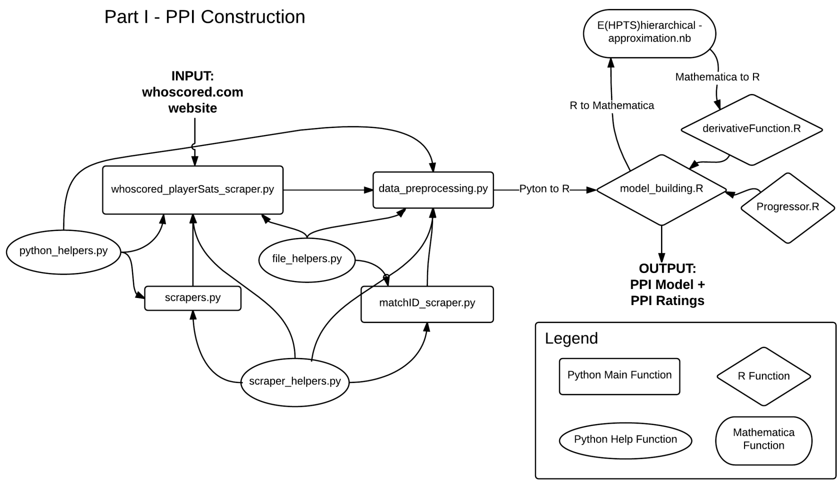

Figure 7: The figure depicts the relation between the software scripts from the input of the whoscored.com website to the output of the PPI model and the PPI ratings for the BL season 15/16.

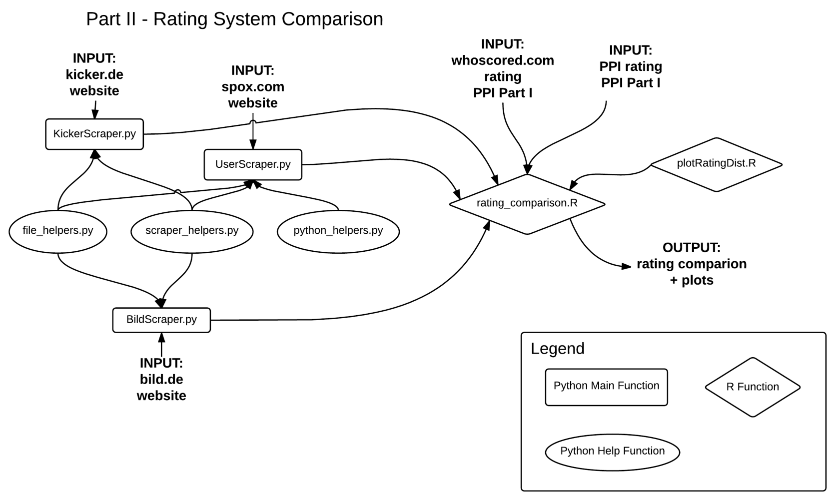

Figure 8: The figure depicts the relation between the software scripts from the input of the kicker.de, bild.de, spox.com (Comunio.de), whoscored.com websites plus the PPI ratings to the output of the rating system comparison and the related plots.

B List of Input Variables

Variable	Variable Type	Description
Assists	Count	Number of played assists.
Big Chance Miss	Count	Number of missed important chances.
Big Clearances	Count	Number of important clearances.
Big Error	Count	Number of big errors made by player.
Clean Sheet	Binary	True when clean sheet result was obtained, otherwise false.
Clearances	Count	Number of clearances.
Crosses	Count	Number of medium to long range passes from a wide area of the field towards the centre of the field.
Cross Success Accuracy	Percentage	Percentage of successful crosses with regard to all crosses.
Date	Date	The date on which the game took place.
Dispossessed	Count	Number of ball dispossessions by opponent.
Dribbles	Count	Number of dribbles.
Duel Aerials Won	Count	Number won duel aerials.
Fouls Committed	Count	Number of committed fouls.
Fouls Given	Count	Number of times the player was fouled.
Goals Scored	Count	Number of goals scored.
Interceptions	Count	Number of interceptions.
Key Passes	Count	Number of important passes played.
Long Ball Pass	Count	Number of long passes played.
Long Ball Pass Success Accuracy	Percentage	Percentage of successful long ball passes with regard to all long ball passes.
Matchday	Nominal	Matchday 1 to 34 on which the game took place.
Match Identifier	Nominal	A match identification number.
Minutes Played	Count	Number of minutes the player was on the pitch.
Name	Nominal	Name of the player.
Offside	Count	Number of times player was caught offside.
Passes	Count	Number of played passes.
Pass Success Accuracy	Percentage	Percentage of successfully played passes.
Pass Through Ball	Count	Number of played passes through opponent's defence.
Pass Through Ball Success Accuracy	Percentage	Percentage of successfully played passes through opponent's defence.
Pitch	Nominal	Value *"home"* for home played games and *"away"* for games played away.
Points Won	Score of 0, 1 or 3	Number of obtained league points for this game.
Position	Nominal	The player role: goalkeeper, defence, midfield or forward.
Rating	Score from 0 to 10	Rating of the website whoscored.com.
Red Card	Binary	True if player received a red card, false otherwise.
Rematch Identifier	Nominal	The match identification number of the rematch.
Result	Nominal	Either *"win"*, *"defeat"* or *"tie"* depending on score.
Saves	Count	Number of goalkeeper saves.
Saves Success Accuracy	Percentage	Percentage of saves with regard to all on target shots.
Score	Nominal	Final score of the match.
Score Team	Count	Number of goals that the team scored.
Season	Nominal	Indicates from which season the game is.
Shots	Count	Number of (on and off target) shots.
Shots Blocked	Count	Number of blocked shots when the shot was on target.
Shots on Target	Count	Number of on target shots.
Substituted	Binary	True if player was taken in or out of the game, otherwise false.
Tackles Won	Count	Number of won tackles.
Team	Nominal	Name of the player's team.
Touches	Count	Number of ball touches during the game.
Turnover	Count	Number bad ball control incidents.
Yellow Card	Count of 0, 1 or 2	Number of yellow cards.

Table 6: List of available player and team variables (not all of the variables are used for constructing the PPI). Variables are measured with regard to a single game.

UNIVERSITEIT GENT

Faculty of Sciences

Soccer Player Performance Rating Systems for the German Bundesliga

Jonathan D. Klaiber

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Christophe Ley

Department of Mathematical Statistics

Academic year 2015–2016