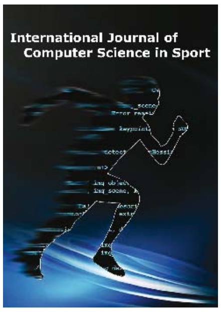
International Journal of Computer Science in Sport

Volume 20, Issue 1, 2021

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2021-0002



Comparing bottom-up and top-down ratings for individual soccer players

Lars Magnus Hvattum and Garry A. Gelade

Faculty of Logistics, Molde University College, , Norway

Abstract

Correctly assessing the contributions of an individual player in a team sport is challenging. However, an ability to better evaluate each player can translate into improved team performance, through better recruitment or team selection decisions. Two main ideas have emerged for using data to evaluate players: Top-down ratings observe the performance of the team as a whole and then distribute credit for this performance onto the players involved. Bottom-up ratings assign a value to each action performed, and then evaluate a player based on the sum of values for actions performed by that player. This paper compares a variant of plus-minus ratings, which is a top-down rating, and a bottom-up rating based on valuing actions by estimating probabilities. The reliability of ratings is measured by whether similar ratings are produced when using different data sets, while the validity of ratings is evaluated through the quality of match outcome forecasts generated when the ratings are used as predictor variables. The results indicate that the plus-minus ratings perform better than the bottom-up ratings with respect to the reliability and validity measures chosen and that plus-minus ratings have certain advantages that may be difficult to replicate in bottom-up ratings.

KEYWORDS: ASSOCIATION FOOTBALL; RANKING; PREDICTION; PLUS-MINUS; VALIDITY

Introduction

High level team sports, such as professional soccer, have become big business, and successful teams acquire significant revenues. As the performance of a soccer team is linked to the playing strength of the individual players, correctly assessing the quality of the players on a team has important repercussions. As large quantities of data become more easily available (Pappalardo et al., 2019b), there is now an increasing interest in assessing player contributions using sports analytics.

Traditionally, the evaluation of players is a manual task performed by experts. However, recent academic literature argues that subjective player evaluations tend to be biased and inaccurate. Kausel, Ventura & Rodríguez (2019) provided an example of outcome bias, where the evaluation of players by reporters is highly linked to the outcome of a penalty shoot-out, even for players not participating in the shoot-out. Tiedemann, Francksen & Latacz-Lohmann (2010) argued that experts do not pay proper attention to the performance of defenders and midfielders, and pointed out that different specialists tend to disagree in their performance evaluations. Vilain & Kolkovsky (2016) also claimed that offensive and spectacular actions tend to be overvalued. Szymanski (2000) showed a different type of bias, where during the years 1978-1993, the wage bills of clubs showed that black players were discriminated.

As a means to reduce the bias introduced by subjective player evaluations, while simultaneously being able to evaluate a large set of players, teams may instead consider data-driven player ratings. These are sometimes considered as more objective when used to evaluate players, but if they are to be used for scouting players, for negotiating player contracts, or for selecting a starting line-up, the decision makers must be aware of any inherent biases of the ratings arising from how they are calculated. Studies that seek to pin-point strengths and weaknesses of different rating paradigms are therefore provide valuable information to decision makers in soccer.

One of the first contributions to data-driven player ratings in soccer was made by McHale, Scarf & Folker (2012). In their system, the rating is expressed as a weighted sum of six components. The first component is based on detailed data from each match, including passes, tackles, crosses, dribbles, blocks, clearances, and cards. The other five components are based on higher level measures, including the number of minutes played, the goals scored, the number of assists, and the number of clean sheets.

A rating system based on observing detailed actions made by each player, then assigning a value to each action performed, and finally aggregating these values for a given player to produce a final rating, can be described as a bottom-up rating system. In these, the final rating is a sum of many small contributions. The first component of the rating system of McHale et al. (2012) is of this type.

Another type of rating system is based on observing only the performance of the team as a whole and then dividing this performance between the players of the team. This can be described as a top-down rating system. The five last components of the rating system of McHale et al. (2012) roughly follow this philosophy of calculation. Other rating systems for soccer players developed since have either followed a pure top-down strategy or a pure bottom-up strategy.

A class of top-down ratings is known as plus-minus (PM) ratings. A basic plus-minus rating consists of accumulating the goal differential obtained from the perspective of a given player (Hvattum, 2019). Schultze & Wellbrock (2018) proposed a PM rating based on simple calculations, making adjustments to control for the importance of goals as well as the strength of the opposing team. However, since the work by Sæbø & Hvattum (2015), most PM ratings for soccer are based on using a regression model to adjust for the quality of teammates and opponents when estimating how a player has contributed towards the goal difference obtained by the corresponding team. With similar models, Sittl and Warnke (2016) used a hierarchical fixed effects model, while Vilain & Kolkovsky (2016) used an ordered probit regression model and Matano et al. (2018) used a Bayesian framework. Regularized multiple linear regression models have been used frequently for PM ratings, with examples provided by Sæbø & Hvattum (2019), Kharrat et al. (2020), Pantuso & Hvattum (2020), and Hvattum (2020).

Wolf et al. (2020) proposed a different top-down rating, based on Elo ratings, where player ratings are updated after each match based on the difference between expected scores and actual scores. While no direct comparison has been made between player ratings based on Elo and PM ratings, by being able to exploit information about starting line-ups, both ratings appear to be better than an Elo team rating at predicting the outcomes of matches, as shown by Wolf et al. (2020) and Arntzen & Hvattum (2020), respectively.

Pappalardo et al. (2019a) discussed a system where the performance of a player is modelled by a feature vector describing the behavior of the player in a match based on observed events. These features are then weighted and summed to form a bottom-up rating. Decroos et al. (2019) introduced a bottom-up rating known as VAEP (valuing actions by estimating probabilities). It assigns a value to any situation that can appear on the pitch, based on event data, and calculates a rating for each player based on how the player changes the situation by making actions. Van Roy et al. (2020) also considered a similar bottom-up rating based on a concept known as expected threat (Singh, 2021).

While the abovementioned rating systems attempt to create a single value that describes the overall performance of a given player, other ratings have been developed for specific parts of a player's contribution. For example, Gyarmati & Stanojevic (2016) ranked players based on the intrinsic quality of their passes, McHale & Relton (2018) evaluated players based on their ability to succeed with difficult passes, and Bransen et al. (2019) evaluated players with respect to chance creation from passes. This builds on a large body of research focusing on passes, such as Chawla et al. (2017) who classified passes using spatiotemporal data and Power et al. (2017) who considered the risk and reward of passes.

Another direction of research is the measurement of team performance, which in turn can have an influence on the development of player ratings. One example is the concept of expected goals (xG), as discussed by Macdonald (2012). This is commonly accepted as a better measure of team performance than the actual goal difference observed, and was exploited by Kharrat et al. (2020) in the context of PM ratings. Link et al. (2016) presented the concept of dangerousity, which can be used to analyze individual actions in soccer, a concept that is also found in VAEP by Decroos et al. (2019).

The purpose of this paper is to examine the relationship between top-down ratings and bottom-up ratings. In particular, three experiments are performed using the bottom-up rating system of Decroos et al. (2019) referred to as VAEP and the top-down rating system of Pantuso and Hvattum (2020) which is a PM rating.

Since bottom-up ratings exploit more detailed data than top-down ratings, the bottom-up ratings should in principle be more reliable and should more accurately reflect the true playing strengths when based on data from the same number of matches. The first experiment is designed to test if this is true for VAEP and PM. While bottom-down ratings require more detailed data, it is typically easier to find data for calculating top-down ratings. The second experiment aims to test whether the estimation of playing strengths based on top-down ratings can be improved, relative to bottom-up ratings, by including observations from a larger set of matches. Again, the experiment is conducted using VAEP and PM. The third experiment focuses on combining top-down ratings and bottom-up ratings to create a hybrid rating system, and to test whether this can lead to ratings that more accurately reflect the true capabilities of the players. The basis for the hybrid rating system is the framework of PM, but using information from VAEP calculations.

While the authors are unaware of any other direct comparisons of top-down and bottom-up ratings for soccer players, the tests reported here may be considered as an extension of the research presented by Gelade and Hvattum (2020). They considered the same PM ratings, and examined their relationship to simple event-based player performance statistics. Their comparisons also indicated that the PM ratings were superior to the information contained in the player performance statistics, with a weak indication that PM ratings could be improved by taking into account data regarding the number of saves per 90 minutes as well as the number of key passes made per 90 minutes. Van Roy et al. (2020) compared VAEP and another bottom-up rating, noting that we lack an understanding of the differences, both conceptually and in practice, of metrics that evaluate actions. Comparisons between top-down and bottom-up ratings are also uncommon for other team sports. However, Engelmann (2011) performed a study where an ad hoc bottom-up rating based on modelling basketball games as a finite state machine performed better than a PM rating.

The rest of this paper is structured as follows. The next section describes the VAEP ratings, the PM ratings, and two variants of a hybrid rating. The data and the metrics used to evaluate the ratings are discussed the subsequent section. Then follows a section presenting the main results and discussions, related to three different experiments. The last two sections contain our conclusions and a discussion of the limitations of the study, respectively.

Description of rating systems

This section first presents the VAEP ratings, then the PM ratings, and finally two hybrid ratings.

VAEP ratings

The original exposition of the VAEP ratings was by Decroos et al. (2019). Decroos et al. (2020) also presented the main ideas, with some updated calculations and discussions, while Decroos (2020) provided a comprehensive treatment of the subject matter.

The rating of a player p is calculated relative to a time frame T, such as a single match, a full season, or several seasons. The rating is expressed as

$$rating(p) = \frac{90}{m} \sum_{a_i \in A_p^T} V(a_i), \tag{1}$$

where A_p^T is the set of actions performed by player p within the time frame, m is the number of minutes played by player p during T, and $V(a_i)$ is the value of action a_i. The latter is given by the following expression:

$$V(a_i) = \Delta_{t_i}^{SCORE}(S_i) - \Delta_{t_i}^{CONCEDE}(S_i), \tag{2}$$

where S_i is the game state after action a_i has been executed, t_i is the team of the player that performed action a_i and

$$\Delta_{t_i}^{SCORE}(S_i) = P_{t_i}^{SCORE}(S_i) - P_{t_i}^{SCORE}(S_{i-1}), \tag{3}$$

$$\Delta_{t_i}^{CONCEDE}(S_i) = P_{t_i}^{CONCEDE}(S_i) - P_{t_i}^{CONCEDE}(S_{i-1}),$$

given that S_{i-1} is the game state that immediately precedes S_i, that is, the game state before action a_i was performed.

To derive VAEP ratings, one therefore has to first estimate, for a given game state and team t, the values of $P_t^{SCORE}(S_i)$ and $P_t^{CONCEDE}(S_i)$, which are defined as the probabilities that team t will score or concede a goal, respectively, within the next $n = 10$ actions.

These probabilities are estimated through machine learning, using observations with labels that simply describe whether or not a goal was scored or conceded in the n actions following a given observed game state. The features that describe a game state S_i are calculated by considering the current action and the k immediately preceding actions: a_{i-k}, a_{i-k+1}, ..., a_{i-1}, a_i. Then, 21 different features are defined as given in Table 1.

Table 1: Features used in VAEP models

ID	Feature description	Defined for
1	Type of action	Each action
2	Result of action	Each action
3	The body part used in the action	Each action
4,5	The two coordinates for the start location	Each action
6,7	The two coordinates for the end location	Each action
8	Time elapsed since the start of the game	Each action
9,10	Distance and angle to goal for start location	Each action
11,12	Distance and angle to goal for end location	Each action
13,14	Distance covered by action in each direction	Each action
15	Distance between end and start location	Each pair of consecutive actions
16	Time elapsed between end and start	Each pair of consecutive actions
17	Whether the ball changed possession	Each pair of consecutive actions
18	Goals scored by the team in possession before action	Final action
19	Goals conceded by the team in possession before action	Final action
20	Goal difference as seen by the team in possession before action	Final action
21	Manpower differential	Final action

The feature with ID 21 has been added in our implementation of VAEP and was not originally suggested by Decroos et al. (2019). The definitions of features with ID 18 and ID 19 in Table 1 also differ slightly from the definitions given by Decroos et al. (2019), where the goals scored and conceded were stated to be obtained after the final action.

As in the original VAEP rating we use $k = 2$, meaning that the last three actions are considered when defining the game state. In some situations, an action a_i does not have any preceding actions, for example if a_i is the kick-off. In that case, preceding actions are replaced by dummy actions located at the middle of the pitch. As the data set used in this study is created in a different way than the data set used by Decroos et al. (2019), the distributions of different types of actions differ. Examples of differences are in the calculation of carries, as well as in the classification of corners. In a later section we demonstrate that the ratings obtained are in any case very similar to those presented by Decroos et al. (2019).

The model is estimated using XGBoost (Chen & Guestrin, 2016). Although Decroos et al. (2019) reported slightly better results with CatBoost (Prokhorenkova et al., 2018), we failed to apply this method to our data set due to running out of available memory.

By noting that the ratings presented by Decroos et al. (2019, 2020) use a cut-off based on the number of minutes played, and by observing very volatile ratings for players with few minutes played, we are motivated to make a change in the formula for calculating final player ratings. As an alternative formula for the rating, instead of dividing by the number of minutes played, m, we divide by $\max\{m, M\}$, where M is a parameter to be decided:

$$rating(p) = \frac{90}{\max\{m, M\}} \sum_{a_i \in A_p^T} V(a_i). \tag{4}$$

PM ratings

PM ratings are based on the idea of distributing credit for the performance of a team onto the players of the team (Hvattum, 2019). We consider a variant of PM where this distribution is done by solving a large unconstrained quadratic program, with variables representing player ratings. For the full mathematical details of the rating system we refer to (Pantuso and Hvattum, 2020), while noting that the PM ratings have also been discussed by Gelade and Hvattum (2020) and Arntzen and Hvattum (2020). An attempt at improving this version of the PM ratings was presented by Hvattum (2020), but with modest success.

The calculations of PM ratings start by splitting each match into segments of time, such that new segments start whenever a player is sent off or a substitution is made. In other words, within a segment the set of players appearing on the pitch is constant. For each segment, a goal difference is calculated as the difference in the number of goals scored by the home team and the away team, in favor of the home team.

The underlying principle of the PM ratings is to find player ratings such that the sum of the player ratings of the home team minus the sum of the player ratings of the away team is as close as possible to the observed goal difference, where the latter is normalized to represent a goal difference per 90 minutes of playing time. To ensure that the ratings match the observed goal differences as closely as possible, the difference is squared for each segment. The sum of the squared differences is then minimized to obtain the PM ratings. Other implementations of PM ratings have used different observed values instead of goal differences, such as the difference in xG or a calculated change in win probabilities (Kharrat et al., 2020), while other types of performance measures such as dangerousity values (Link et al., 2016) have yet to be tested.

This does not immediately lead to reasonable player ratings, and a range of additional factors is therefore included to better model the link between ratings and observed goal differences. First, the home team has a natural advantage, which is represented using a separate variable. Given that the size of the home field advantage can vary among competitions, a single variable is included for each competition in the data set. Second, some segments have missing players due to red cards being given by the referee. This is handled by including additional variables corresponding to the missing players, while remaining player ratings are scaled so that their sum represents the average rating for a full team.

Third, the data may include matches from a wide variety of different league systems and divisions, spanning many years. To better deal with this, the player ratings are split into several components: Each type of competition is given a separate rating component. The rating of a player is extended by the average of the ratings for competitions in which the player has appeared. Furthermore, the rating of a player is not assumed to be constant over the whole data set, but rather to be a function of the player's age. This is modelled as a piecewise linear function which is estimated together with the ratings by introducing corresponding age adjustment variables.

To obtain more stable ratings, at the expense of a bias, the variables are subject to an L2 regularization scheme, also known as ridge regression or Tikhonov regularization. This implies that all variables are shrunk towards zero. However, for variables representing the main player rating component and for variables representing age adjustments, a different method is used. Player rating variables are, for each player, instead shrunk towards an average of the player rating variables for a set of similar players. The set of similar players is taken as those players that have played the most minutes together with the target player. For age adjustments, each adjustment is shrunk towards the average value of the adjustments of the immediately lower and higher age groups.

Finally, each segment is weighted by a factor that depends on the recency of the segment, the duration of the segment, and the game states. Taking into account all of the above results in an optimization problem with a number of variables slightly larger than the number of players in the data set and a number of squared terms in the objective function marginally larger than the number of segments plus the number of players.

The calculations of PM ratings can be summarized by the simplified expression

$$\min_{\beta} Z = \sum_{q} \left(w(q) \left(f^{LHS}(q) - f^{RHS}(q) \right) \right)^2 + \lambda \sum_{j} \left(f^{REG}(\beta_j) \right)^2, \tag{5}$$

where the first summation is over all match segments, q. The second summation is over all variables representing player components, home field advantage components, competition components, age components, and red card components. The terms of the second summation represent regularization terms, and are weighted by a parameter λ.

In the first summation, each segment is weighted by $w(q)$, and the expression has two main parts: First, $f^{RHS}(q)$ represents the observation that is made within the segment. For calculating PM ratings, this is taken as the goal difference accrued within the segment, as seen from the perspective of the home team. Second, $f^{LHS}(q)$ contains a mix of variables representing components that aim to explain how the observed goal difference came about.

Combined VAEP and PM ratings

When we test combinations of top-down ratings and bottom-up ratings, we keep the definition of $f^{LHS}(q)$ as indicated in the preceding section, and in more detail by Pantuso and Hvattum (2020). However, we change the definition of $f^{RHS}(q)$ to account for the VAEP scores obtained in a segment, rather than the goal difference observed.

Two variants of combining VAEP and PM are considered. In the first variant, the observations are simply based on the difference in accumulated VAEP scores for each of the two teams.

That is

$$f^{RHS}(q) = \sum_{p \in R_1} \sum_{a_i \in A_p^q} V(a_i) - \sum_{p \in R_2} \sum_{a_i \in A_p^q} V(a_i), \tag{6}$$

where A_p^q is the set of actions performed by player p within the segment q, $V(a_i)$ is the value of action a_i, and R_1 and R_2 are the sets of players appearing on the home team and away team, respectively. We refer to this variant as PM-VAEP.

The second variant aims to capture the dominance of a team over another. It considers, the scoring probability of the home team minus the scoring probability of the away team for the given game state S_i arising from each action a_i: $P_{R_1}^{SCORE}(S_i) - P_{R_1}^{CONCEDE}(S_i)$. Each of these differences in scoring probabilities are associated to the time at which the corresponding event takes place. A dominance function is defined by interpolating between the events. The dominance of the home team over the away team is then calculated by taking the integral of this function, from the beginning of a segment until the end of the segment. This variant is referred to as PM-P-INT.

The resulting ratings can capture some aspects of the VAEP, since the dependent variable in the calculations is based on VAEP scores, as well as some aspects of the PM, since the observed VAEP scores are credited to all the players on the pitch and not just the player performing the action.

Experimental setup

This section starts by describing the data used in the numerical experiments. Then the process for evaluating the validity of player ratings is outlined, before the calculations to assess reliability of ratings are given.

Data sets

The main data set comes from the top five leagues in Europe: the English Premier League, the German Bundesliga, the Italian Serie A, the Spanish Primera División, and the French Ligue 1. Ten seasons from 2009/10 to 2018/19 are covered, for a total of 18,047 matches with sufficient data quality to be used in the experiments described in this paper.

For the main data set, the information required to calculate bottom-up ratings is present. This involves, for each match, a list of actions performed during the match. Each action has information about the player performing the action, the match clock, the type of action performed, the outcome of the action performed, the location where the action started, the location where the action ended, and the body part used to execute the action. In addition, for each action the current numbers of goals scored by each team and the current manpower differential are noted.

Calculating top-down ratings requires less detailed data. In particular, it suffices to know which players are starting the match, the timing of goals scored, as well as the time and players involved for any substitutions and red cards. In addition, the top-down ratings examined here uses the birth dates of players as a part of the calculations.

For calculating top-down ratings, an additional set of data is obtained, that can be split into four parts. The first part comprises 3,620 additional matches from the 2007/08 and 2008/09 seasons of the top five leagues. The second part has 21,377 matches from the ten seasons covered in the main data set, but from the second tier of each league: the English Championship, the German 2. Bundesliga, the Italian Serie B, the Spanish Segunda División, and the French Ligue 2. The third part again covers the same seasons as the main data set but contains matches from the UEFA Champions League and the Europa League. This part has 6,553 matches. Finally, the fourth part has matches from three other leagues, namely the Turkish Süperlig, the Portuguese Primeira Liga, and the Dutch Eredivisie. This additional data set is only used in the second experiment to test whether top-down ratings can be improved by including more observations.

Validity

Franks et al. (2016) discussed the evaluation of player ratings in general. One criterion that they mentioned is relevance, or how a rating is related to an outcome of interest such as the number of matches won. We refer to this as a test of the validity of player ratings: if ratings are accurate in assessing the capabilities of players, one should be able to accurately predict the outcome of a match based only on the ratings of the players involved.

The procedure used follows the description by Arntzen and Hvattum (2020). The matches in the full data set are divided into three parts, as shown in Table 2. The matches from 2009/10 up to and including 2014/15 are only used for calculating player ratings. Then, matches from the 2015/16 and 2016/17 seasons are considered in a sliding window fashion. That is, the matches are considered sequentially, sorted from oldest to newest. Before a match is played, the value of a covariate, x, is calculated as the average rating of players in the starting line-up of the home team minus the average rating of players in the starting line-up of the away team. After the match, the outcome of the match is recorded as $y = 1$ if the home team won, $y = 2$ if the game ended in a draw, and $y = 3$ if the away team won. Afterwards, the ratings for players involved in the match are recalculated, using all information up to and including the match.

Matches from the seasons 2017/18 and 2018/19 are handled similarly: covariate values x and outcomes y are recorded, and the ratings of players are updated after the match is played. In addition, the covariate value x is used to predict the outcome of the match before it is played, and the quality of the prediction is evaluated. The prediction is based on a regression model estimated using observations of x and y from past matches.

In particular, for predictions, an ordered logit regression model (Greene, 2012) is applied. This technique was suggested already by Dobson and Goddard (2001) in the context of soccer. The regression model provides a prediction in the form of probabilities z_j for each of the three possible outcomes $j = 1,2,3$ corresponding to home win, draw, and away win. Let $d_j = 1$ if j is the actual outcome, and $d_j = 0$ otherwise. We then evaluate the predictions through two proper scoring rules (Witten, Frank & Hall, 2011), known as quadratic loss,

$$L^2 = (d_1 - z_1)^2 + (d_2 - z_2)^2 + (d_3 - z_3)^2, \tag{7}$$

and informational loss,

$$L^I = -\log_2(d_1 z_1 + d_2 z_2 + d_3 z_3). \tag{8}$$

When the process has completed, we have recorded the quadratic loss and the informational loss for all matches in the 2017/18 and 2018/19 seasons, which amounts to 3,604 matches. To evaluate whether one set of predictions is better than another set of predictions, a two-tailed paired t-test is used. The P-value from this test is used to determine whether we can reasonably assume that the predictions are equally good, and we reject this notion and conclude that the quality of predictions differ when the P-value is small.

Table 2: The ten seasons in the main data set and how they are used in calculations for assessing the validity of ratings.

	Usage		
Season	**ratings**	**observations**	**predictions**
2009/10	Y*	N	N
2010/11	Y*	N	N
2011/12	Y*	N	N
2012/13	Y*	N	N
2013/14	Y*	N	N
2014/15	Y*	N	N
2015/16	Y	Y	N
2016/17	Y	Y	N
2017/18	Y	Y	Y
2018/19	Y	Y	Y

A minor adjustment of the process outlined above happens if a team uses a player in the starting line-up that has not previously played any matches. Such players do not have a current rating before the match is played, and therefore do not contribute towards the average rating of players on their team.

Table 2 highlights seasons 2009/10 through 2014/15 with an asterisk when it comes to their use to calculate ratings. The reason is that to calculate VAEP ratings, one must first estimate the probabilities for scoring and conceding conditional on the game state, P^{SCORE} and $P^{CONCEDE}$. This is a time-consuming task, and we have chosen to only estimate this model once, using the data from all the matches in the highlighted seasons. For the PM, on the other hand, the model used can be solved quickly, and is therefore updated after each match day.

Reliability

The second dimension along which we evaluate the player ratings is referred to as their reliability. The underlying idea is that the relative ratings of players should not depend on which data set is used to calculate the ratings. In other words, if the data set is randomly split into two halves, and player ratings are calculated for each half separately, then the rating calculated for a given player should be approximately the same for both halves.

To get an overall number indicating the similarity of the ratings in the two halves of the data set, the sample Pearson correlation coefficient is calculated for the set of players that appear in both rating lists. This is repeated twenty times, each time using a different random split of the data, and the average correlation coefficient is reported as an indication of how reliable the ratings are: if the average is close to 1, it means that the player ratings lead to very similar rankings, independent of the actual matches used to calculate ratings, whereas a value close to 0 means that the ratings appear almost random, and that the selection of matches used to calculate ratings is very important.

Again, some special considerations must be taken as a result of the VAEP relying on having estimates of the probabilities of scoring and conceding for the observed actions and game states. Therefore, seasons 2009/10 to 2014/15 are ignored, and the data set used to calculate ratings for reliability testing is limited to the seasons 2015/16 to 2018/19. This is somewhat biased against the PM ratings, as they are then based only on these four seasons, whereas the VAEP implicitly exploits the previous seasons as they are used to estimate the model for P^{SCORE} and $P^{CONCEDE}$.

When splitting the data set in two halves, one can choose either to split based on matches or on segments. In this work, the data set is split based on segments, as defined for the PM ratings. The results are very similar in the case that the data set is split on matches, but splitting on segments typically means that a higher number of players will appear in both halves of the data set, and thus can be considered in the calculations of the correlation coefficients.

Results and discussion

The first part of this section is devoted to a reproduction of the results reported by Decroos et al. (2019), to verify that our implementation of VAEP ratings is sufficiently close to the original. The next three parts address each of the three experiments presented in the introduction.

Reproduction of VAEP results

Subject to some differences in the set of leagues covered and the actions included, as well as minor differences in the features of the model, the tests reported by Decroos et al. (2019) are reproduced using seasons 2012/13 through 2017/18. The first four seasons are initially used to train a model that is then used to predict labels for the fifth season. Then, the model is retrained using the five first seasons, after which labels from the sixth season are predicted. Table 3 shows the frequencies of different actions in the data set considered. These differ noticeably from Decroos et al. (2019) in the presence of carries (23 %), which may be partly included as dribbles (9 %) by Decroos et al. (2019).

The obtained results for evaluating P^{SCORE} and $P^{CONCEDE}$ are similar to those presented by Decroos et al. (2019), but slightly better: Values for the area under the receiver operating characteristic curve (ROC AUC) are 0.788 for scoring and 0.767 for conceding, compared to 0.756 and 0.723 in (Decroos et al., 2019), with higher values indicating a better fit. The values of quadratic loss (Brier scores) are 0.0124 for scoring and 0.0039 for conceding, compared to 0.0139 and 0.0055 in (Decroos et al., 2019), with lower values representing better predictions.

To document the differences in the underlying data, Table 4 and Figure 1 show the final goal of the match between Barcelona and Real Madrid on December 23, 2017. In the paper by Decroos et al. (2019), the two carries and the take on performed by Messi are combined into a single take on. Then, Figure 2 shows the top 10 players in English Premier League 2017/18 according to VAEP. It lines up well with the corresponding figure in Decroos et al. (2019), and we conclude that the reproduction of VAEP ratings is sufficiently accurate.

Table 3: Frequencies of actions in the data set from seasons 2012/13 to 2017/18.

Action	Frequency
Pass	52.56 %
Carry	22.98 %
Clearance	3.38 %
Throw-in	2.90 %
Interception	2.80 %
Take on	2.44 %
Tackle	2.41 %
Cross	1.75 %
Foul	1.69 %
Freekick pass	1.63 %
Ball touch	1.51 %
Shot	1.51 %
Keeper pickup	0.86 %
Crossed corner	0.53 %
Keeper save	0.38 %
Crossed freekick	0.19 %
Claim	0.16 %
Short corner	0.10 %
Punch	0.07 %
Freekick shot	0.07 %
Unsporting behavior	0.05 %
Penalty shot	0.02 %

Table 4: Evaluation of actions leading up to the goal illustrated in Figure 1.

ID	Time	Player	Action	P	V
1	94.07	Busquets	Pass	0.03	0.00
2	94.11	Messi	Pass	0.03	−0.01
3	94.13	Busquets	Pass	0.04	0.01
4	94.16	Messi	Carry	0.05	0.01
5	94.17	Messi	Take on	0.06	0.01
6	94.18	Messi	Carry	0.08	0.02
7	94.18	Messi	Pass	0.16	0.08
8	94.22	Vidal	Shot	1.00	0.84

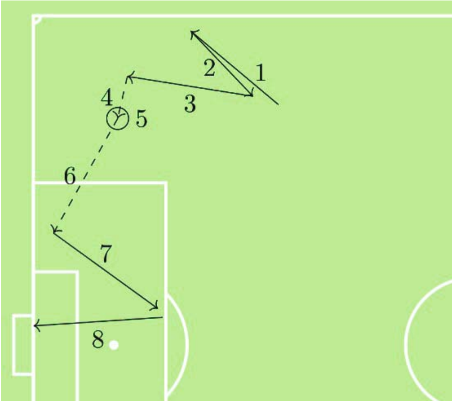


Figure 1: The attack leading up to Barcelona's final goal in their 3-0 win against Real Madrid on December 23, 2017

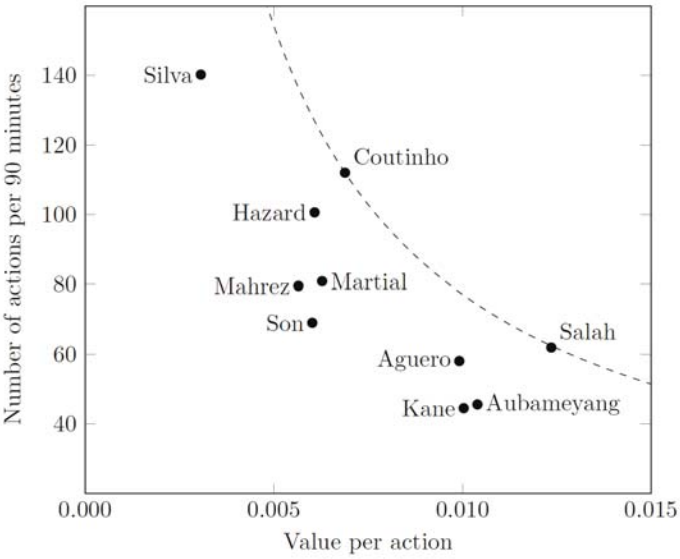


Figure 2: Top 10 players in the English Premier League 2017/18 according to VAEP.

Bottom-up ratings versus top-down ratings

The first experiment aims to test whether the validity and the reliability of VAEP is better than PM, given that the bottom-up ratings have access to more detailed data. Before evaluating this experiment, one parameter of the VAEP ratings was tuned: the value of M that governs how ratings are pushed towards zero for players with few minutes of recorded play. When tuning this parameter, the same data and setup was used as for performing the final tests. Figure 3 shows, in gray dots connected by a solid line, different values of M, ranging from 1 (bottom left) to 3,600. The best value was decided in terms of the performance on the validity criterion, corresponding to minimizing the quadratic loss of predictions based on the ratings, and the final value used was $M = 1{,}800$, indicated by a black dot.

As shown in Figure 3, the PM ratings outperform the VAEP ratings both in terms of reliability (the consistency of the ratings when evaluated on different data sets) and validity (the quality of predictions made based on ratings of players in the starting line-ups of a match). For validity, the figure shows the average quadratic loss of predictions from 3,604 matches. Smaller values are better, and the axis is turned so that better values are to the right. Regarding reliability, when splitting the segments of the last four seasons in the data set in two halves, there are on average around 4,500 players that obtain a rating from both halves. The PM and VAEP are relatively close in terms of reliability, but the VAEP requires a proper value for the parameter M, and is very unreliable when M is small.

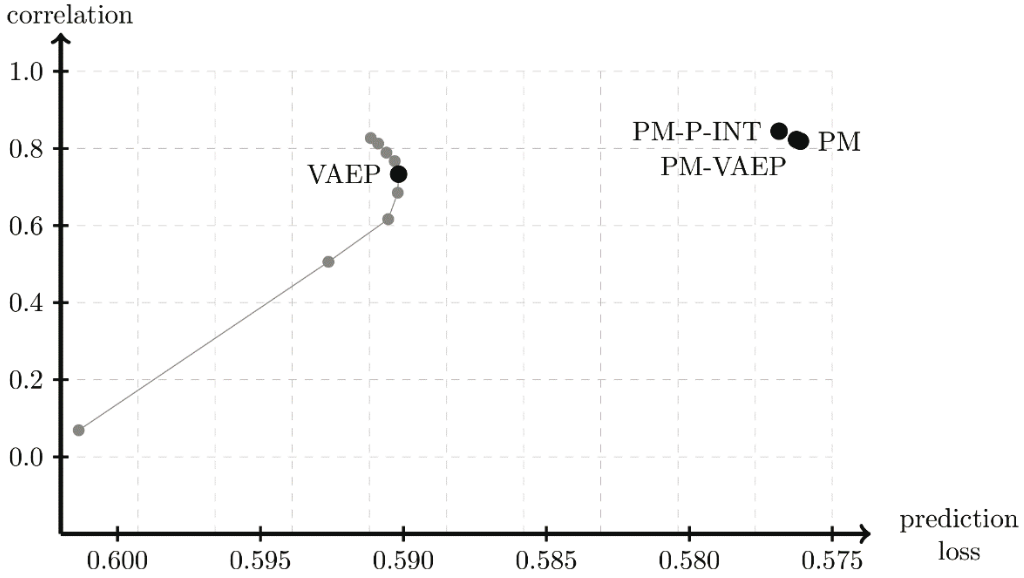


Figure 3: Evaluation of reliability and validity for top-down ratings, bottom-up ratings and two different hybrid ratings.

Table 5 provides additional details to properly evaluate the experiment, by supplying both the quadratic loss and informational loss. The differences in both quadratic loss and informational loss between PM and VAEP are statistically significant, with a paired t-test indicating a P-value of less than 0.0001 for both loss functions. Furthermore, the effect size is relatively large when compared to similar types of studies, where the difference sin quadratic loss between good prediction methods and market predictions are typically less than 0.005 (Hvattum & Arntzen, 2010). We can therefore not conclude that VAEP ratings are better than PM ratings, as the evidence supports the opposite: in this case the top-down rating is better than the bottom-up rating.

We also tested the ordered logit regression model with two covariates: using both the PM and the VAEP ratings to calculate separate covariates. The difference between using only PM and using both PM and VAEP is not significant, with P-values of 0.17 and 0.09 for quadratic loss and informational loss, respectively. This means that there is only a weak indication that including information from the VAEP improves the predictions provided by PM alone.

Table 5: Quadratic and informational loss for predictions on 3,604 matches from the 2017/18 and 2018/19 seasons, using PM ratings, VAEP ratings, or both as covariates.

	Quadratic loss		**Informational loss**	
Covariates	**Avg.**	**Std.**	**Avg.**	**Std.**
PM	0.576	0.352	1.399	0.736
VAEP	0.590	0.316	1.427	0.645
PM, VAEP	0.575	0.352	1.397	0.735

The validity of ratings can be evaluated in other forms that through match forecasts. Tables 6 and 7 present the top 20 players ranked according to PM and VAEP, respectively, calculated using the full data set. A form of face validity can be tested from these tables, in that they should correspond to the best players as commonly perceived by experts. The Ballon d'Or is an annual award recognizing the best soccer players of the last year. The players in the rankings of Tables 6 and 7 that were also among the 30 nominees for the 2019 Ballon d'Or are highlighted in bold. For PM ratings, there are seven nominees in the top 20, whereas for VAEP ratings there are six. On the other hand, VAEP has the eventual winner of the 2019 Ballon d'Or, Lionel Messi, ranked as a clear number one. However, unlike this particular award, neither the PM nor the VAEP aim to assess the performance of players in a single year only, but rather across all the seasons in the provided data set.

Utilizing more data in top-down ratings

The second experiment simply intends to show whether top-down ratings such as the PM ratings can be improved by including observations from additional matches. As described in the section on data sets, an additional data set in four parts was prepared to test this. The reliability measure is not convenient when comparing different data sets, as the sets of players that are included will differ. Therefore, we only consider the validity measure in this part.

Table 8 reports the quadratic loss and the informational loss on the 3,604 matches used in prediction when using PM ratings only, and when including different sets of additional matches in the calculations of ratings. The table indicates a very weak trend that adding more data improves the quadratic loss and informational loss, and we again use paired t-tests to see if this improvement, relative to not adding any extra data, is statistically significant.

When adding data from two additional seasons, data from the second level of competitions, or data from other leagues, the improvement in prediction loss is not significant, with P-values well above 0.1. Adding data from the European cups gives an improvement of the prediction loss with a P-value of 0.04, for both the quadratic loss and the informational loss. Interestingly, adding all of the above data, even though giving the smallest loss values, still gives P-values that are rounded to 0.04. At this point, there is not sufficient evidence to claim that adding more data improves the quality of the obtained PM ratings.

Table 6: Top 20 players ranked according to their PM ratings as of July 2019

	Player	Country	PM
1	Ederson	BRA	0.296
2	Thomas Muller	GER	0.269
3	Kyle A. Walker	ENG	0.264
4	Thiago Alcantara	ESP	0.264
5	**Lionel Messi**	ARG	0.259
6	Neymar	BRA	0.258
7	**Bernardo Silva**	POR	0.252
8	**Mohamed Salah**	EGY	0.231
9	David Alaba	AUT	0.230
10	**Raheem Sterling**	ENG	0.223
11	Danilo	BRA	0.220
12	Toni Kroos	GER	0.219
13	Aymeric Laporte	FRA	0.215
14	**Sadio Mane**	SEN	0.212
15	James Rodriguez	COL	0.211
16	**R. Lewandowski**	POL	0.210
17	Ilkay Gundogan	GER	0.207
18	**Kun Aguero**	ARG	0.207
19	Fabinho	BRA	0.207
20	Jerome Boateng	GER	0.203

Table 7: Top 20 players ranked according to their VAEP ratings as of July 2019

	Player	Country	VAEP
1	**Lionel Messi**	ARG	0.897
2	**Kylian Mbappe**	FRA	0.687
3	Arjen Robben	NED	0.615
4	**C. Ronaldo**	POR	0.555
5	Franck Ribery	FRA	0.546
6	**Mohamed Salah**	EGY	0.534
7	James Rodriguez	COL	0.519
8	Neymar	BRA	0.513
9	Z. Misimovic	BIH	0.501
10	Jadon Sancho	ENG	0.486
11	**Eden Hazard**	BEL	0.485
12	Carlos Tevez	ARG	0.477
13	Julio Alvarez	VEN	0.468
14	**Kun Aguero**	ARG	0.467
15	Benito Raman	BEL	0.467
16	O. Dembele	FRA	0.453
17	Mohamed Zidan	EGY	0.452
18	Serge D. Gnabry	GER	0.452
19	Harry Kane	ENG	0.446
20	Dries Mertens	BEL	0.442

Table 8: Quadratic and informational loss for predictions on 3,604 matches from the 2017/18 and 2018/19 seasons, using PM ratings and additional data for ratings calculations.

Matches added		Quadratic loss		Informational loss	
Type	**Number**	**Avg.**	**Std.**	**Avg.**	**Std.**
None	0	0.576	0.352	1.399	0.736
Two seasons	3,620	0.576	0.352	1.399	0.736
Second level	21,377	0.575	0.354	1.397	0.741
European cups	6,553	0.575	0.351	1.396	0.732
Other leagues	8,798	0.576	0.352	1.397	0.736
All	40,348	0.574	0.353	1.394	0.738

Given that adding more data had a relatively small effect, a second test is performed where instead of adding data, existing data is gradually removed. Doing so, we can examine the effect on both PM and VAEP, with the caveat that the VAEP model is still built using all the data from seasons 2009/10 to 2013/14, as explained in the section on data sets.

In the original tests, six seasons of data were used only for calculating ratings, two seasons were used for ratings and initial observations for the ordered logit regression model, and the two final seasons were used also for recording predictions. Figure 4 shows the quadratic loss of predictions for the 3,604 matches in the last two seasons as well as the correlation coefficients of the reliability tests when removing up to nearly eight seasons worth of data, starting with the oldest matches.

Removing more than six seasons of data is very severe, as it means one must start creating observations for the ordered logit regression model without having any prior ratings, leading to the covariate having a value of zero. It is therefore a surprise to observe that the prediction loss does not increase drastically for the VAEP until almost 7.5 seasons of data are removed. For the PM ratings, the loss starts to increase after removing six seasons of matches, but even after removing almost all the matches from the eight seasons (the right-most observations correspond to removing all but the last two weeks of matches from the eighth season), the predictions using PM ratings are better than any predictions made using VAEP ratings.

To give an indication of how good the ratings still are: when training the ordered logit regression model on a randomly generated covariate with zero information content, the average quadratic loss becomes 0.645, and the average informational loss becomes 1.540. Regarding the reliability, as indicated in Figure 4b, when removing seven or more seasons, the data set used in the splitting procedure is reduced, and the number of players appearing in both halves is reduced. This influences the correlation calculations as shown.

Taking all of these observations into account, it seems that the second experiment does not support the idea that more data in top-down ratings is necessarily helpful: Even the top-down ratings require much less data than was prepared for the planned experiments.

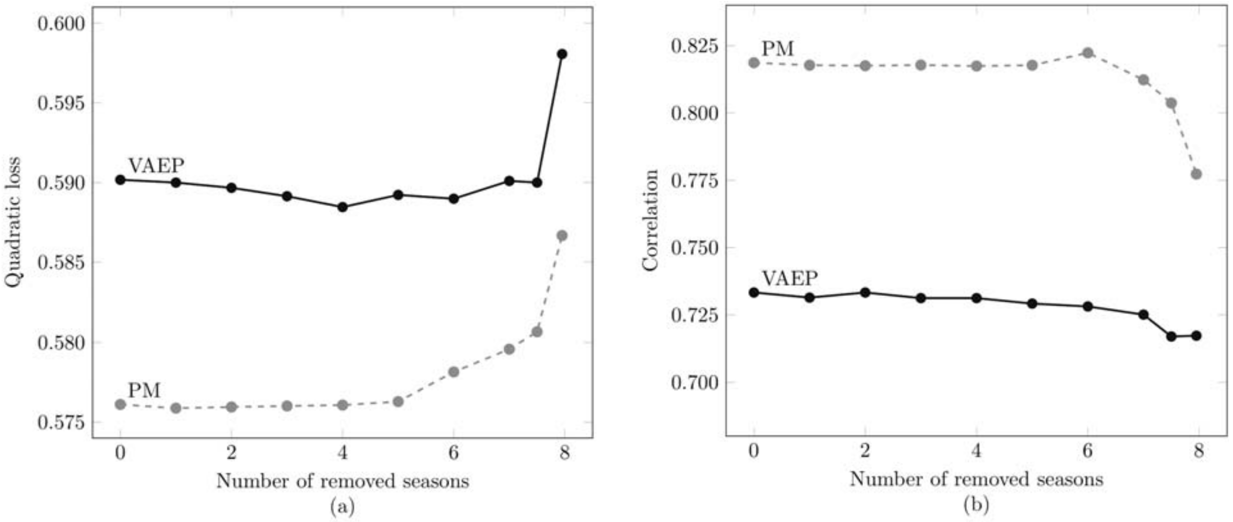


Figure 4: The effect of removing a number of seasons from the data set on the prediction loss (a) and split-correlations (b) of ratings.

Combining bottom-up and top-down ratings

The third experiment aims to test whether a hybrid rating system built on the principles of top-down and bottom-up ratings can lead to improved player evaluations. Table 7 already gave an indication that combining PM ratings and VAEP ratings would perhaps not be a big improvement. Table 9 shows results from similar tests where covariates in the ordered logit regression model used for predictions are calculated using the two variants of hybrid ratings, PM-VAEP and PM-P-INT.

The results show that neither PM-VAEP nor PM-P-INT provide an improvement over PM. This is also shown in Figure 3, where it is clear that the hybrids perform very similarly to the simpler PM rating. To explain this, we look at the relationship between the observed goal differences in a segment and the observed values used for $f^{RHS}(q)$ in the hybrid ratings.

Table 9: Quadratic and informational loss for predictions on 3,604 matches from the 2017/18 and 2018/19 seasons, using the two variants of combined top-down and bottom-up ratings investigated.

	Quadratic loss		Informational loss	
Covariates	**Avg.**	**Std.**	**Avg.**	**Std.**
PM-VAEP	0.576	0.351	1.399	0.733
PM-P-INT	0.577	0.357	1.400	0.747

It turns out that the correlation between the observed values of goal differences and the metrics used in PM-VAEP and PM-P-INT is high. Considering the weighted values, $w(q)f^{RHS}(q)$, the correlation between observations based on goal differences and based on VAEP is 0.93. For PM-P-INT, the correlation with values from PM is slightly less, but still 0.76. Part of the explanation for this is that VAEP scores aggregated over a segment is very close to the goal difference observed, simply by the fact that a goal leads to $P_{t_i}^{SCORE}(S_i) \approx 1$, since the outcome of the action and the type of action is included in the feature set.

To emphasize that the ratings produced by PM, PM-VAEP, and PM-P-INT are similar, the paired t-tests for the prediction loss show no differences between the predictions, with P-values upwards of 0.48, for any combination of methods. In other words, the experiments conducted show that hybrid ratings are not an improvement over the pure ratings.

Limitations

There are some aspects of the rating systems that are worthwhile to discuss in relation to the conclusions above. Although VAEP was outperformed by PM in our experiments, it has certain advantages that were not explored. For one, the VAEP ratings can be calculated for each match separately, allowing that the performance of a player is traced from one match to another. The flip-side of this is that this performance is likely to be noisy, which presumably shows up through the VAEP ratings being less reliable than the PM ratings.

Inherently, the VAEP ratings do not compensate for the quality of the opposition or the contributions from the teammates. This means that VAEP ratings could be bad at differentiating between players appearing on different teams, while still being valuable at determining which players are better contributors on a given team. PM ratings, on the other hand, may benefit by being able to correctly assess the difference in quality between different teams. However, it is not necessarily true that the PM ratings are as good at determining which players on a team are most important. That being said, Arntzen and Hvattum (2020) showed that PM ratings do provide useful information not captured by team ratings such as the Elo rating.

Furthermore, interaction effects between players are not taken into account by either type of rating. Bransen & Van Haaren (2020) used VAEP ratings as a means to investigate the chemistry between pairs of players, which could be used to improve the VAEP ratings of individual players. No similar work has been done using PM ratings for soccer, but both Gramacy et al. (2013) and Thomas et al. (2013) introduced player interaction effects in PM ratings for ice hockey.

The main limitation of the current work is perhaps the assumption that it is reasonable to assess the quality of ratings by the predictions of future match outcomes that can be made based on ratings of players in the starting line-ups. Other ways to evaluate the validity of ratings may be more favorable for bottom-up ratings.

Both VAEP and the variants of PM ignore the length of game stoppages. That is, when calculating ratings per 90 minutes, the assumption is that play is always continuous from the kick-off of the first half until half-time and from the kick-off of the second half until the final whistle. The assumption is that game stoppages are fairly uniformly distributed throughout all matches, which may not always be the case.

As a further limitation, the study considered the original form of VAEP ratings. However, Decroos (2020) discussed an update of VAEP based on a new definition of events, where each event is always completed successfully, without interruption. The usefulness of this is exemplified by a pass that is unsuccessful: the mistake could either be due to the player who initiated the action (the pass not reaching a teammate) or the player who received the pass (but failed to take control of the ball). It remains untested whether this updated VAEP (atomic-VAEP) is more reliable and better suited towards rating players with the aim of predicting outcomes of future matches.

Conclusions

This paper has compared a bottom-up rating (VAEP) with a top-down rating (PM) for individual players in soccer. Little work has previously been done in analyzing the differences between these two types of ratings, and three different experiments were performed in this paper. First, it was shown that bottom-up ratings are not necessarily better than top-down ratings at assessing player’s abilities. In fact, according to the experiments conducted, PM ratings were more reliable and produced better predictions for future match outcomes than VAEP ratings. Second, it was shown that no more than a few seasons worth of match data is necessary to provide reasonable ratings, when considering them as used for predicting future match outcomes. Third, producing hybrid rating systems to exploit properties of both types of rating systems is not trivial, and two different variants examined here failed to outperform the simpler PM ratings.

References

Arntzen, H. & Hvattum, L.M. (2020). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, forthcoming.

Bransen, L., Van Haaren, J. (2020). Player chemistry: striving for a perfectly balanced soccer team. ArXiv: 2003.01712v1.

Bransen, L., Van Haaren, J., & van de Velden, M. (2019). Measuring soccer players’ contributions to chance creation by valuing their passes, *Journal of Quantitative Analysis in Sports*, *15*, 97–116.

Chawla, S., Estephan, J., Gudmundsson, J., & Horton, M. (2017). Classification of passes in football matches using spatiotemporal data. *ACM Transactions on Spatial Algorithms and Systems*, 3, Article 6.

Chen, T. & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY , USA, pages 785–794.

Decroos, T. (2020). Soccer analytics meets artificial intelligence: learning value and style from soccer event stream data. PhD Dissertation, KU Leuven, Belgium.

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: valuing player actions in soccer. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY , USA, pages 1851–1861.

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2020). VAEP: an objective approach to valuing on-the-ball actions in soccer (Extended Abstract). In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), pages 4696–4700.

Dobson, S. & Goddard, J. (2001). *The Economics of Football*. Cambridge University Press, Cambridge.

Engelmann, J. (2011). A new player evaluation technique for players of the National Basketball Association (NBA), Proceedings of the MIT Sloan Sports Analytics Conference.

Franks, A., D'Amour, A., Cervone, D., & Bornn, L. (2016). Meta-analytics: tools for understanding the statistical properties of sports metrics. *Journal of Quantitative Analysis in Sports*, 12, 151–165.

Gelade, G.A. & Hvattum, L.M. (2020). On the relationship between +/− ratings and event-level performance statistics. *Journal of Sports Analytics*, 6, 85–97.

Gramacy, R., Jensen, S., & Taddy, M. (2013). Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports*, 9, 97–111.

Greene, W. (2012). Econometric Analysis. Pearson, Harlow, England, 7th edition.

Gyarmati, L. & Stanojevic, R. (2016). QPass: a merit-based evaluation of soccer passes. In *KDD 2016 Workshop on Large-Scale Sports Analytics*.

Hvattum, L.M. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport*, 18, 1–23.

Hvattum, L.M. (2020). Offensive and defensive plus-minus player ratings for soccer. *Applied Sciences*, 10, 7345.

Hvattum, L.M. & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26, 460–470.

Kausel, E.E., Ventura, S., & Rodríguez, A. (2019). Outcome bias in subjective ratings of performance: Evidence from the (football) field. *Journal of Economic Psychology*, 75, 102132.

Kharrat, T., Peña, J., & McHale, I. (2020). Plus-minus player ratings for soccer. *European Journal of Operational Research*, 283, 726–736.

Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. PLoS ONE, 11, e0168768.

Macdonald, B. (2012). An expected goals model for evaluating NHL teams and players. Proceedings of the 2012 MIT Sloan Sports Analytics Conference.

Matano, F., Richardson, L., Pospisil, T., Eubanks, C., & Qin, J. (2018). Augmenting adjusted plus-minus in soccer with FIFA ratings. ArXiv:1810.08032v1.

McHale, I.G. & Relton, S.D. (2018). Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research*, 268, 339–347.

McHale, I., Scarf, P., & Folker, D. (2012). On the development of a soccer player performance rating system for the English Premier League. *Interfaces*, 42, 339–351.

Pantuso, G. & Hvattum, L.M. (2020). Maximizing performance with an eye on the finances: a chance constrained model for football transfer market decisions. *TOP*, forthcoming.

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019a). PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology*, 10, 59.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019b). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6, 236.

Power, P, Ruiz, H., Wei, X., & Lucey, P. (2017). Not all passes are created equal: objectively measuring the risk and reward of passes in soccer from tracking data. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 1605–1613.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, pages 6639–6649.

Sæbø, O. & Hvattum, L. (2015). Evaluating the efficiency of the association football transfer market using regression based player ratings. In: NIK: Norsk Informatikkonferanse, Bibsys Open Journal Systems, 12 pages.

Sæbø, O. & Hvattum, L. (2019). Modelling the financial contribution of soccer players to their clubs. *Journal of Sports Analytics*, 5, 23–34.

Schultze, S. & Wellbrock, C. (2018). A weighted plus/minus metric for individual soccer player performance. *Journal of Sports Analytics*, 4, 121–131.

Singh, K. (2019). Introducing expected threat. https://karun.in/blog/expected-threat.html. Accessed: 2021-01-25.

Sittl, R. & Warnke, A. (2016). Competitive balance and assortative matching in the German Bundesliga. Discussion Paper No. 16-058, ZEW Centre for European Economic Research, Mannheim.

Szymanski, S. (2000). A market test for discrimination in the English professional soccer leagues. *Journal of Political Economy*, 108, 590–603.

Thomas, A., Ventura, S., Jensen, S., & Ma, S. (2013). Competing process hazard function models for player ratings in ice hockey. *The Annals of Applied Statistics*, 7, 1497–1524.

Tiedemann, T., Francksen, T., & Latacz-Lohmann, U. (2011). Assessing the performance of German Bundesliga football players: a non-parametric metafrontier approach. *Central European Journal of Operations Research,* 19, 571–587.

Van Roy, M., Robberecths, P., Decroos, T., & Davis, J. (2020). Valuing on-the-ball actions in soccer: a critical comparison of xT and VAEP. AAAI-20 Workshop on AI in Team Sports. (https://ai-teamsports.weebly.com/uploads/1/2/7/0/127046800/paper11.pdf

Vilain, J. & Kolkovsky, R. (2016). Estimating individual productivity in football. http://econ.sciences-po.fr/sites/default/files/file/jbvilain.pdf, accessed 2019-08-03.

Witten, I., Frank, E., & Hall, M.A. (2011). Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, 3rd edition.

Wolf, S., Schmitt, M., & Schuller, B. (2020). A football player rating system. *Journal of Sports Analytics*, 6, 243–257.