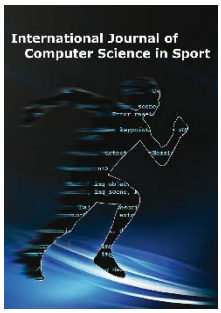
International Journal of Computer Science in Sport

Volume 18, Issue 1, 2019

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

sciendo

DOI: 10.2478/ijcss-2019-0001



A comprehensive review of plus-minus ratings for evaluating individual players in team sports

Lars Magnus Hvattum

Faculty of Logistics, Molde University College, Molde, Norway

Abstract

The increasing availability of data from sports events has led to many new directions of research, and sports analytics can play a role in making better decisions both within a club and at the level of an individual player. The ability to objectively evaluate individual players in team sports is one aspect that may enable better decision making, but such evaluations are not straightforward to obtain. One class of ratings for individual players in team sports, known as plus-minus ratings, attempt to distribute credit for the performance of a team onto the players of that team. Such ratings have a long history, going back at least to the 1950s, but in recent years research on advanced versions of plus-minus ratings has increased noticeably. This paper presents a comprehensive review of contributions to plus-minus ratings in later years, pointing out some key developments and showing the richness of the mathematical models developed. One conclusion is that the literature on plus-minus ratings is quite fragmented, but that awareness of past contributions to the field should allow researchers to focus on some of the many open research questions related to the evaluation of individual players in team sports.

KEYWORDS: RATING SYSTEM, RANKING, REGRESSION, REGULARIZATION

Introduction

Rating systems, both official and unofficial ones, exist for many different sports. Individual sports such as tennis and golf are good examples where official rating lists are published. Although well established, these rating systems have been found to be lacking in terms of being able to predict future results (McHale and Forrest, 2005, McHale and Morton, 2011). A similar situation holds true for team sports, where rating systems attempt to measure the quality of teams. For example, in soccer, the national teams are covered by the FIFA ranking, which was found to perform poorly with respect to predicting future results (McHale and Davies, 2007). Currently, the FIFA ranking system have been updated and replaced by a variant of the Elo rating system (Elo, 1978). Elo ratings were originally developed for chess, but had been unofficially adapted to soccer (Hvattum and Arntzen, 2010, Stefani and Pollard, 2007), along with other, similar, alternatives (Constantinou and Fenton, 2013, Lasek, Szlávik, and Bhulai, 2013).

Creating ratings for individual players in team sports is a significantly harder challenge: while the task of creating a meaningful rating system for teams is a challenge in itself, several new issues arise when trying to identify the contribution of a single player to the whole of the team. As noted by McHale, Scarf, and Folker (2012), interactions between the individual players and their teammates, as well as interactions between the individual players and the opposing players, both make the performance of a player difficult to isolate. In some team sports, different players have different roles. However, these roles may change during a player's career, or even during a single match, making it difficult to know which attributes of a player are essential when attempting to create a player rating.

Ratings have several applications. Official ratings may be used for seedings in tournaments, or even as a qualification criterion to enter a tournament (McHale et al., 2012). National federations may find individual player ratings useful for consideration of post season awards, or for picking national teams (Hass and Craig, 2018). Clubs may find player ratings useful for identifying and recruiting talented or undervalued players (Hass and Craig, 2018, McHale et al., 2012), evaluating transfer fees (Sæbø and Hvattum, 2015), and preparing for salary negotiations (Macdonald, 2012a). Coaches or managers may use player ratings as support to assess the value of a player's contributions to a team (Sæbø and Hvattum, 2019), to compile team rosters (Pantuso, 2017), or to determine which players of the opposing team to focus on in an upcoming match (Schultze and Wellbrock, 2018). Sports fans may follow ratings for entertainment purposes, to evaluate how well their favorite player or team is performing, and possibly use this as a starting point for friendly debates (Hass and Craig, 2018, McHale et al., 2012). Gamblers and bookmakers may use player ratings to improve predictions for financial gains (Engelmann, 2017).

Subjective player evaluations tend to be biased. Vilain and Kolkovsky (2016) pointed out that offensive and spectacular player actions tend to be overvalued, whereas defensive and less spectacular player actions are largely undervalued. Tiedemann, Francksen, and Latacz-Lohmann (2010) also claimed that specialists' assessments of top players were paying insufficient attention to the performance of certain player roles, while noting that specialists tended to disagree on performance evaluations. Szymanski (2000) presented evidence that certain irrelevant player attributes influenced the players' wages. In other words, there is evidence that subjective opinions about players are biased. At the same time, good player ratings can be useful for fans, club management, coaches, soccer associations, the media, and the players themselves (Tiedemann et al., 2010).

This paper focuses on a particular type of player ratings, known as plus-minus ratings. To characterize plus-minus ratings, it may be beneficial to distinguish between what can be labelled as bottom-up ratings and top-down ratings. Bottom-up ratings are calculated based on individual player actions: if a player performs an action that is associate with a positive outcome for the team, that player's rating is increased, while the rating of other players, who were not involved in the play, are unaffected. While bottom-up ratings focus on the performance of each player, ignoring the eventual result for the team, top-down ratings do the opposite. In top-down ratings, the idea is that the performance of the team as a whole is observed, and the credit for this performance is distributed on individual players being involved in the match, no matter which actions they performed.

Plus-minus ratings are top-down ratings, and attempt to assess the impact of a player by comparing the performance of the team with and without the player. Many variants of this concept have appeared, mostly to evaluate players in basketball and ice hockey, but gradually also for other sports. This paper aims to provide a thorough survey of relevant literature. This literature is to some extent characterized by non-peer reviewed work, but there is an increasing number of peer reviewed publications presenting valuable contributions.

The remainder of this paper is structured as follows. In the three next sections, literature on plus-minus ratings for basketball, ice hockey, and then soccer is discussed. After this follows a section considering plus-minus ratings for other sports. The third last section summarizes the literature and provides an analysis of publication patterns. Some suggestions for future work on plus-minus ratings are provided in the penultimate section, before concluding remarks are given in the final section.

Plus-minus for basketball

Consider a player in a basketball team and a given match. Count the number of points scored by the player's team while the player is in the game, and then subtract the number of points scored by the opposing team during the same time intervals. The resulting number will hereafter be referred to as the basic plus-minus rating of the player. Such ratings have unofficially been calculated for basketball players in the NBA since 2003, and starting in 2007 the NBA officially updated each player's individual plus-minus rating during live play (Winston, 2009).

Several weaknesses of the basic plus-minus ratings were known from the beginning (Thomas, Ventura, Jensen, and Ma, 2013). One such weakness is that basic plus-minus ratings ignore the quality of a player's teammates and opponents. The consequence is that poor players on good teams are overrated, while good players on poor teams are underrated. Adjusted plus-minus ratings use multiple linear regression to resolve this issue, and initial developments were made in the context of basketball.

Dan Rosenbaum was likely the first to publish details of what is now known as an adjusted plus-minus rating, in an article posted on the web site 82games.com (Rosenbaum, 2004). His method involved splitting each match into observations where no substitutions are made, and then running the following multiple linear regression:

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j X_{ij} + \varepsilon_i \quad (1)$$

where the dependent variable Y_i is taken as the difference of home team points per possession and away team points per possession for the observation. The independent variable X_{ij} is 1 if player j is playing at home, -1 if player j is playing away, and 0 otherwise. Then, β_0 measures

the average home court advantage, and β_j for $j = 1, \ldots, K$ measures the difference between player j and a set of reference players. This set of reference players refers to all players with less than 250 minutes of playing time in the two seasons of data used in the calculations, and the reference players would not appear in the regression. The difference between the observed value of the dependent variable and the corresponding value derived from the independent variables, in other words the error term, is written as ei. Rosenbaum (2004) assigned a weight to each observations, such that an observation in the second season of data is weighted twice as high as an observation in the first season, and such that crunch time observations have higher weights and garbage time observations have lower weights.

Rosenbaum (2004) noted that the resulting ratings are noisy, presumably due to relatively large standard errors for the parameters of the regression model. A three step procedure was proposed that leads to less noisy ratings: First, the adjusted plus-minus ratings are regressed on a set of game statistics, such as shot attempts, assists, rebounds, turnovers, and fouls, measured per 40 minutes. In a model to do this, each observation corresponds to one player, the dependent variable is the adjusted plus-minus rating of that player, and the independent variables are the player's game statistics. This provides estimates of how different actions tend to be correlated with players' plus-minus ratings. Second, these estimates are used to calculate so-called statistical plus-minus ratings, based on the game statistics of each player. Third, an overall rating is calculated for each player as a convex combination of the two previous ratings, while weighing the ratings so as to minimize the standard error of the overall rating. Rosenbaum (2005) used a weight of 60 % for adjusted plus-minus and 40 % for statistical plus-minus when presenting updated results based on additional seasons of data.

While Rosenbaum (2004) may have been the first to publish a detailed description of adjusted plus-minus ratings, the first version of adjusted plus-minus ratings is likely to have been developed by Jeff Sagarin and Wayne Winston (Kubatko, Oliver, Pelton, and Rosenbaum, 2007), under the name WINVAL.Winston (2009) provided a simplified firsthand explanation of WINVAL. While most of the details of the method are omitted, an interesting variant is described. The variant is referred to as WINVAL Impact ratings, and replaces the points differential, Y_i, with the change in win probability, as calculated based on the current score and the time remaining of the match. Versions of WINVAL are known to have been in use by the NBA team Dallas Mavericks (Ilardi and Barzilai, 2008).

According to Ilardi (2007), adjusted plus-minus ratings were developed independently at least three times: by Rosenbaum (Rosenbaum, 2004), by Sagarin andWinston (Winston, 2009), and by Steve Ilardi for the college basketball team Kansas Jayhawks. The method of Ilardi was also described in an internet article at 82games.com, outlining one particular extension compared to the method of Rosenbaum. The extension is to explicitly model offensive and defensive ratings for each player. Assuming that a match is split into segments where no substitutions are made, each segment contributes to two observations: one for the offense of the home team, and one for the offense of the away team. The model can be stated as:

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j^O X_{ij}^O + \sum_{j=1}^{K} \beta_j^D X_{ij}^D + \beta_{K+1} X^H + \varepsilon_i \qquad (2)$$

where Y_i is now taken as the points per possession for the team playing offense. The independent variable X_{ij}^O takes the value 1 if player j is playing on offense, and 0 otherwise, whereas X_{ij}^D is -1 if player j is playing defense, and 0 otherwise. The last independent variable, X^H, takes the value 1 if it is the home team that is playing on offense.

Ilardi and Barzilai (2008) also highlighted the importance of trying to reduce the level of noise in the ratings obtained when using the method of ordinary least squares to estimate the parameters of the multiple linear regression model given above. That is, since some players tend to appear mostly in the same lineups, and the number of observations per player per season is quite low, the model suffers from high degrees of multicollinearity. To reduce the noise when estimating ratings for a given season, Ilardi and Barzilai (2008) used several seasons of data, but with lower weights for the past seasons so as to still obtain ratings that are relevant for the current season.

An important contribution to plus-minus ratings was made by Sill (2010). Given concerns about the accuracy of adjusted plus-minus ratings, Sill identified overfitting in addition to multicollinearity as a reason for noisy ratings. While earlier contributions tackled this by using more seasons of data and removing players with few minutes played, Sill (2010) proposed to use regularization and cross validation to improve the ratings. Using the method of ordinary least squares, the parameters of model (1) would be found by minimizing

$$\sum_i \left(Y_i - \beta_0 - \sum_{j=1}^{K} \beta_j X_{ij} \right)^2 \tag{3}$$

A Bayesian technique of regularization, known as ridge regression or Tikhonov regularization, implies that one instead minimizes

$$\sum_i \left(Y_i - \beta_0 - \sum_{j=1}^{K} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=0}^{K} (\beta_j)^2$$

This corresponds to using a Gaussian prior distribution, stabilizing the estimated ratings, both in cases with few observations and when multicollinearity is present (Thomas et al., 2013). Cross validation can be used to determine suitable values for the regularization parameter λ. Sill (2010) found that the resulting ratings performed better when using regularization, and that it became much less useful to have additional seasons of data or to remove players with few minutes played.

Engelmann (2011) compared a regularized adjusted plus-minus rating to an ad hoc player rating based on modelling basketball games as a finite state machine. The latter is able to use information about individual player actions, and was shown to perform better when predicting point differentials. This is one of few studies that have compared the performance of plus-minus ratings with that of ratings based on more detailed data.

Omidiran (2011) started with the realization that adjusted plus-minus ratings are calculated by minimizing the expression (3), and suggested that this objective function could be extended with additional terms multiplied by a set of regularization weights. In particular, he suggested to add terms to 1) penalize deviations of ratings from a median rating, 2) align ratings with the players' box scores while 3) penalizing the magnitude of weights for each box score statistic, and 4) force the sum of ratings to be close to zero. Box score statistics cover simple summaries of actions performed on the player level, such as assists, blocks, or steals. For a particular set of regularization weights, a coordinate descent algorithm was used to estimate parameters, while cross validation was used to find the best set of regularization weights.

Basic plus-minus ratings were used by Okamoto (2011). Considering each player separately, he counted the number of wins and losses for the team in games where the player had a positive or

negative plus-minus, respectively. This allowed the calculation of odds ratios, giving the probability of the team winning as a function of whether a given player obtained a positive or negative plus-minus score. This analysis was extended to take into account up to two players at a time, and to take into account that some players may be more important in home games or away games.

Fearnhead and Taylor (2011) presented an adjusted plus-minus model to calculate both offensive and defensive ratings, as first suggested by Ilardi and Barzilai (2008). However, they used a Bayesian framework where the offensive and defensive ratings were assumed drawn from separate Gaussian distributions. While Sill (2010) and Omidiran (2011) used cross-validation to determine the regularization parameter, Fearnhead and Taylor (2011) used maximum likelihood to estimate the hyperparameters of the prior Gaussians, having the effect that the distributions of offensive and defensive ratings can take means different from 0:

$$\beta_j^O \sim \mathcal{N}(\mu_O, \sigma_O^2),$$
$$\beta_j^D \sim \mathcal{N}(\mu_D, \sigma_D^2).$$

Using several seasons of data to support the estimation of player abilities in a single season, Fearnhead and Taylor (2011) also allowed player strengths to vary over time. This was achieved by using the previous season's rating, shrunk towards the prior mean, as the basis for calculating a new rating based on current season data. Their analysis also included using multiple linear regression, with the resulting player ratings as the dependent variable and various individual game statistics as independent variables, as suggested by Rosenbaum (2004).

While focusing on the visualization of match statistics, Sisneros and Van Moer (2013) considered a new variant of plus-minus ratings, based on the use of box scores. They first considered team level differentials of each box score statistic, then summed a team's per game statistic differential over all games played in a season in order to find the value of a season long differential. Each player was then assigned the percentage of the value of a statistic corresponding to the percentage contribution of the player towards that statistic.

In 2014, the media actor ESPN introduced their own version of adjusted plus-minus, called real plus-minus (Ilardi, 2014), of which details are mostly unknown. It is, however, clear that it is a mix of regularized adjusted plus-minus and statistical plus-minus (Engelmann, 2017). Then, Deshpande and Jensen (2016) published a paper introducing several new techniques for calculating adjusted plus-minus ratings. One of their main ideas is to use the change in win probabilities across shifts, instead of the score differential, as the dependent variable. This had not been considered since the work of Winston (2009). In addition to variables denoting the presence of players and the home court advantage, they also included team variables among the independent variables. This means that player ratings must be interpreted as the marginal contribution of a player relative to its team, and makes it difficult to compare players across teams.

Another contribution of Deshpande and Jensen (2016) is in the estimation of ratings. As in (Fearnhead and Taylor, 2011), a Bayesian linear regression is used. However, the authors chose to use independent Laplacian priors for each player component and team component. Furthermore, Markov Chain Monte Carlo (MCMC) simulation is then used to estimate the posterior distribution. To compare players, it was suggested to use the ratio of the posterior mean to the posterior standard deviation of the player's partial effect. This results in a measure that has relatively low correlation from one season to the next, but this correlation improves when using more seasons of data.

Engelmann (2017) reviewed several versions of adjusted plus-minus ratings for basketball. He discussed the use of additional independent variables in adjusted plus-minus models, for example to control for schedule related factors such as rest days, to control for the influence of coaching, to control for the current score difference, or to control for the age of players when using multiple seasons of data. Few results of such analyses have been presented in the scientific literature so far. While plus-minus ratings typically use the points scored (or the change in the points differential) as the dependent variable, the impact of players on other metrics can also be assessed. In particular, Engelmann (2017) mentions 4-factor regularized adjusted plus-minus, in which the dependent variable can be based on field goal attempts, rebounds, turnovers, or free throws.

Plus-minus for ice hockey

Basic plus-minus ratings originated in ice hockey, being in secret use by the NHL team Montréal Canadiens in the 1950s, and then popularized by the NHL coach Emile Francis in the following decade (Thomas et al., 2013). The NHL keeps track of plus-minus ratings for each player. At the time of writing, the record of NHL goes back to 1959/1960, while different authors have claimed that the official recording of plus-minus ratings started in either 1967 or 1969 (Fyffe and Vollman, 2002; Winston, 2009).

Whereas plus-minus ratings for basketball suddenly took a large step from their basic rendition to the adjusted plus-minus versions, in ice hockey some other techniques were discussed first. Fyffe and Vollman (2002) suggested to modify the basic plus-minus ratings by subtracting the expected plus-minus, corresponding to the goal difference for the player's team multiplied by the percentage of time that a player has been on the ice. Only even strength situations would be considered. Awad (2010a) made a similar suggestion, but instead subtracted the individual basic ratings of each teammate, weighted according to how much time they have spent on the ice together. In addition, instead of using goal differentials, he used weighted shots, adjusted for situation of play and the opponents. This corresponds to what is now known as expected goals (xG) in some literature (Rathke, 2017), providing for each shot the probability that similar shots end with a goal. In particular, Awad (2010a) calculated the values of weighted shots (xG) based on the distance from goal; whether the shot was a rebound; whether the shot was made in even strength, short handed, or in power play; whether the shot was made after an opponent give-away, the shot type; and the game score Awad (2010b).

While the adjusted plus-minus ratings were initially developed for basketball, between 2004 and 2009, it did not take long before the ideas surrounding adjusted plus-minus started to be adapted for ice hockey. First out was Witus (2008) who re-implemented the models of Rosenbaum (2004) in the context of ice hockey. Schuckers, Lock, Wells, Knickerbocker, and Lock (2011) developed one of the first new versions of adjusted plus-minus for ice hockey. The data used was constructed on a play-by-play basis. Unlike basketball, ice hockey involves one player per team dedicated to defend the goal. Noting that there is a large number of plays where goalies are not involved, goalies were therefore omitted from the regression model. In contrast to basketball, ice hockey is a low scoring sport. Therefore, instead of using the change in score as the dependent variable, Schuckers et al. (2011) used for each play the difference between the expected and the observed outcome, where the outcome is an indicator of whether or not a goal was scored within seconds following the play. The authors noted that there was a weak relationship between the players' basic plus-minus ratings and their adjusted plus-minus ratings.

Early contributions to adjusted plus-minus ratings for ice hockey, however, are mostly associated to the work by Brian Macdonald, outlined in five separate papers. Macdonald (2011a)

presented two different models. Only even strength situations are considered, and the dependent variable is measured in terms of goals per 60 minutes. In the first model, both offensive and defensive ratings are estimated directly, as in (Ilardi and Barzilai, 2008), and an overall rating for a player can be found by adding up his offensive and defensive rating. In this model, goalies are only considered when on defense. The second model estimates the overall rating directly, as in (Rosenbaum, 2004). To separate these ratings into offensive and defensive ratings, another linear regression model is used to estimate the contribution to the total number of goals per 60 minutes for each player. The ratings from the two models are averaged, resulting in smaller standard errors for the rating estimates.

Macdonald (2011b) presented two extensions of the model with offensive and defensive ratings. The first extension is to include two additional independent variables to represent whether or not the observation starts with a faceoff in either the defensive or offensive zone. This can be stated as follows:

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j^O X_{ij}^O + \sum_{j=1}^{K} \beta_j^D X_{ij}^D + \beta_{K+1} X^{OFF} + \beta_{K+1} X^{DEF} + \varepsilon_i \quad (4)$$

where Y_i is the number of goals per 60 minutes during the observation (for the team on offense), X_{ij}^O is one if player (excluding goalies) j is playing offense in observation i, and X_{ij}^D is one if player (including goalies) j is playing defense in observation i. The new indictor variables are X^{OFF} and X^{DEF}, which are equal to one if the observation starts with a faceoff in the offensive zone or defensive zone, respectively.

The second extension in (Macdonald, 2011b) is to calculate separate ratings for power play and shorthand situations. This is accomplished simply by using a separate model with independent variables corresponding to player ratings on offense and defense when playing shorthanded and in power play, thereby yielding approximately twice the number of variables compared to the even-strength model.

Macdonald (2012a) used ridge regression, as proposed by Sill (2010), to reduce the high errors in rating estimates. He noted that, besides multicollinearity and players with few minutes played, another reason for high errors in ice hockey is the low number of goals scored per match, compared to the much higher number of points scored in basketball. With this in mind, the paper also suggests to use other measures than the number of goals to represent the dependent variable. In particular, three additional models are presented that use 1) shots, 2) shots plus missed shots, and 3) shots plus missed shots and blocked shots, respectively. It was shown that the year-to-year correlation for the resulting ratings are higher when using ridge regression than when using the method of ordinary least squares, and that it is higher when using shots instead of goals. The year-to-year correlation is higher for even strength calculations than for power play and shorthand situations.

Macdonald (2012b) added a fifth option to calculate the dependent variable, based on the expected number of goals for the observation. The expected number of goals is derived from a linear regression model with independent variables based on goals, various shot metrics, zone starts, turnovers, faceoffs and hits. Finally, Macdonald, Lennon, and Sturdivant (2012) presented a sixth alternative for the dependent variable, calculating the probabilities of scoring from each shot encountered during each observed match segment. The probability of scoring from a shot (xG) is calculated using a logistic regression, and leads to similarities with the method of Awad (2010a).

Schuckers and Curro (2013) extended the work from (Schuckers et al., 2011). They now used as the dependent variable the probability that a goal will be scored by the home team minus the probability that a goal will be scored by the away team, considering the 20 seconds following each event occurring within an observation. Further extensions comprise the inclusion of home-ice advantage, starting zone, and the use of ridge regression to estimate the regression coefficients. Players appearing for different teams were treated as separate players, and the regularization parameter of the ridge regression was set to minimize the differences of ratings for players when appearing for different teams.

Until this point, research on plus-minus ratings for ice hockey had involved testing many different versions of the dependent variable, with the conclusion that more stable ratings could be obtained by using more detailed information than just the number of goals scored for each team. However, all of the ratings involved the use of multiple linear regression, using either the method of ordinary least squares, or ridge regression. The three contributions discussed next outlined alternatives to the use of linear regression.

First, Gramacy, Jensen, and Taddy (2013) proposed to use logistic regression. The observations now correspond to goals scored, instead of segments of the match without substitutions. Hence, Y_i is taken as 1 if goal i was scored by the home team, and -1 otherwise. Using q_i to denote the probability that goal i was scored by the home team, the model can be stated as:

$$\log\left(\frac{q_i}{1-q_i}\right) = \sum_{j=1}^{M} \beta_j^T X_{ij}^T + \sum_{j=1}^{K} \beta_j^P X_{ij}^P \quad (5)$$

where X_{ij}^P are indicator variables for player appearances, and X_{ij}^T are other independent variables, which in the actual implementation included team indicators. Estimates for player ratings β_j^P are found using penalized likelihood maximization. That is, to avoid the known issues with players with few appearances and players that appear mostly together, the likelihood function is extended by regularization terms. The authors considered both the squared penalty (L2), as in ridge regression, which corresponds to placing a Gaussian prior on the player ratings, and an absolute value penalty (L1), as in lasso regression, corresponding to placing a Laplacian prior on the player ratings. Using the Laplacian prior to estimate player ratings, the authors found that few players are distinguishable from the average player of their team. MCMC was used to estimate the full posterior distribution. This allowed statements regarding the probability that a given player is better than another player on the same team.

Gramacy et al. (2013) also included an analysis of player-player interactions. That is, some pairs of players may obtain synergy effects, so that their performance together is better than just the sum of their individual performances. Despite the computational effort expended, the authors concluded that there was little evidence of significant player interaction effects. Also, including the few non-zero interactions found had a very limited effect on the individual player estimates.

Spagnola (2013) avoided the use of player fixed effects. Instead, two logistic regression models were used to find the relationship between the probability of a team scoring (or conceding) within a segment and several simple statistics, such as goals, shots, and hits. Then, player ratings were found by calculating average statistics over all shifts for a player and inserting those values into the estimated regression models. An accompanying study of players on one particular NHL team was performed to illustrate the approach.

Thomas et al. (2013) also departed from the well-established regime of using multiple linear regression. They modelled the scoring rate for each team as its own semi-Markov process, with hazard functions for each process that depend on the players on ice. This resulted in separate

offensive and defensive player ratings. Different shrinkage methods (regularization) were applied when estimating the player ratings, depending on the type of analysis performed. Only full-strength, five-on-five situations were considered. The authors used maximization of penalized likelihood to obtain initial parameter estimates, and then applied MCMC. This resulted in relatively large runtimes, up to 60 processorhours for 200,000 outcomes and 2,600 independent variables.

The study by Thomas et al. (2013) also included an attempt to identify player pair interactions. This was done by considering a subset of 1,000 player pairs, focusing on those with most minutes played together. Lasso regression was then employed to reduce the number of non-zero pairs, and finally MCMC was used to obtain the final results. In total, 221 of the player pairs had an effect not equal to zero when using this estimation technique. The authors noted that the method produced similar results for player effects as other approaches, including (Macdonald, 2011a) and (Gramacy et al., 2013), suggesting that there is sufficient information in the data to distinguish player ability at a grand level, despite different models.

Smith (2016) presented two plus-minus models. The first model is a standard multiple linear regression model estimated using the method of ordinary least squares. The second model is a logistic regression model estimated using lasso regularization, as in (Gramacy et al., 2013). For both models, observations are generated from shots taken, and the value of the dependent variable is based on the probability of scoring, conditional on the location of the shot. The independent variables consist of player presence variables, and the intercept is taken to represent the home ice advantage. These choices imply a criticism of the independent variables chosen in (Gramacy et al., 2013): it is argued that team effects, as included by (Gramacy et al., 2013) are unnecessary once the 10 players on the ice are identified. It is found that the model estimated using ordinary least squares performs inadequately, despite using shot data instead of goals, presumably due to multicollinearity. The author concludes that the regularized logistic regression works well, but that only half of the player estimates are non-zero.

Gramacy, Taddy, and Tian (2017) extended the logistic regression model from (Gramacy et al., 2013), while simultaneously giving a more detailed reasoning for some of their modelling choices. New independent variables are considered, including indicators for special-teams scenarios, covering power play and shorthanded situations, and indicators for whether the goalie is temporarily replaced by a skater. Interaction terms between players and seasons are added, so that player ratings can change between seasons, as well as similar terms for post-season play. There are also independent variables for team-season effects. When estimating the model parameters, the authors chose to penalize only the player effects. Furthermore, to decide on an appropriate penalty term, they use the corrected Akaike information criterion instead of cross validation.

In terms of results, Gramacy et al. (2017) do not find any evidence that any player's ability changes from regular season to post-season. However, they find ample evidence that player performance change between seasons. Additional tests are conducted where the dependent variable is changed from focusing on goals to focusing on shots, leading to a massive increase in the number of observations. This results in ranking lists with distinctly different players both at the top and at the bottom. The authors found this to be alarming for analysts relying on shot based statistics, based on the idea that only goal differentials are ultimately dictating the winners of games.

Plus-minus for soccer

While basic plus-minus ratings have been calculated for basketball players and ice hockey players for many years, soccer has never embraced this metric. Bohrmann (2011) considered an implementation of adjusted plus-minus as in (Rosenbaum, 2004), using one season of data. With an independent variable based on the numbers of goals scored, the conclusion was that no players had a rating significantly different from zero, and that the problem of multicollinearity was simply overwhelming. Hamilton (2014) also considered the adjusted plus-minus ratings, with the dependent variable equal to the observed goal difference in each segment, divided by the length of the segment and multiplied by 90, the length of a match. Using a single season of data from the English Premier League, ridge regression was applied to estimate player ratings. However, the conclusion was that the predictive ability of the resulting ratings was very weak. A major issue was pointed out in that a soccer match typically consists of only around six unique segments.

The first academic work on plus-minus ratings for soccer appears to be by Sæbø and Hvattum (2015), who developed an adjusted plus-minus rating based on a multiple linear regression model, as in Equation (1), estimated using ridge regression. Using observations based on segments without substitutions or dismissals, the observed goal difference within a segment is scaled by the number of minutes played. Additional independent variables are used to represent players that have been sent off with red cards. Up to four players can be sent off for each team, yielding eight such variables.

As dividing soccer matches into segments based on substitutions and red cards provides few segments per match, several seasons of data must be used to provide reasonable player ratings. To facilitate using many seasons of data, Sæbø and Hvattum (2015) exponentially discounted older observations in the data set. The authors determined the appropriate discounting factor as well as the regularization parameter by simulating the league table of a hold-out season. Parameters were selected to minimize the mean squared error of predictions for the number of points per team. It was found that the ratings produced were useful when trying to explain observed transfer fees for players moving to the English Premier League.

Sæbø and Hvattum (2019) extended the model of Sæbø and Hvattum (2015) after realizing that the ratings did not sufficiently discriminate between players from different league systems or different divisions. To resolve this, it was suggested to add a component in the rating of each player corresponding to the competition (league) in which their current team plays. Kharrat, Peña, and McHale (2018) considered three different multiple linear regression models, with independent variables very similar to those of Sæbø and Hvattum (2019), but with a slightly different way to adjust for league differences. The three models differ in terms of the dependent variable. The first model uses the goal differential, as in (Sæbø and Hvattum, 2019). The second model uses the difference in scoring probabilities for shots during the segment (known as expected goals, xG) as in (Awad, 2010a) and (Macdonald et al., 2012). The third model uses the change in expected points for the home team minus the expected points for the away team, with three points for a win, one point for a draw, and zero points for a loss. This has similarities to using the change in win percentages, as in (Deshpande and Jensen, 2016) and (Winston, 2009).

Schultze and Wellbrock (2018) proposed a new plus-minus metric based on simple calculations, instead of following the trend of relying on regression models. Their adjustment to the basic plus-minus aims to control for the importance of goals as well as the opponents' strength. The metric is calculated on a per-minute basis. If goals are scored in a given minute, the ratings of players on the pitch are adjusted by adding the goals scored minus the goals conceded multiplied by a factor indicating the importance of the goal. The factor is equal to 1 if the goal changes the

outcome of the match, and a fraction of that if it only increases or decreases the winning margin. In addition, for every minute a player is appearing, his rating is adjusted by the difference in win probabilities for each team divided by the total number of minutes in the match. The win probability is calculated based on pre-match betting odds. This has the effect of reducing the ratings for players on good teams and increasing the ratings of players on bad teams, and makes it difficult to compare players from different teams.

Sittl and Warnke (2016) and Warnke (2017) described a novel model for calculating plus-minus ratings. In a hierarchical fixed effects model, observations were created for each player for each match. The dependent variable is the goal difference from the perspective of the player when he is on the pitch. This goal difference is then assumed to be a linear function of a player fixed effect, a team fixed effect, a coach fixed effect, an opponent team fixed effect, and some time varying characteristics of the player/match combination. The time varying characteristics included league and season effects, the home field advantage effect, the age (and age squared) of the player, and the number of dismissals for each team. For some of these characteristics, interaction terms were also included. The model was estimated using weighted least squares dummy variable regression, with observations weighted based on the number of minutes played. Sittl and Warnke (2016) also considered models similar to (Sæbø and Hvattum, 2015), but concluded that it was valid to split team productivity into a team component (long-term effect), a coach component (medium-term effect), and player components (short-term effect).

Another novel model was proposed by Vilain and Kolkovsky (2016). They formulate an ordered probit regression model and use penalized maximum likelihood estimation to derive player ratings. Each match is taken as two observations, one focusing on the home team and one on the away team, and the outcome is measured as the number of goals scored for the focal team. In ordered regression, the outcome is linked to independent variables through a latent continuous variable. This latent variable is assumed to depend linearly on variables representing the offensive ratings of players on the focal team and the defensive ratings of players on the opposing team, as well as a variable representing the home field advantage. Hence, Vilain and Kolkovsky (2016) calculated both offensive and defensive ratings, but in such a way that defenders do not have an offensive rating, that forwards do not have a defensive rating, and that goalkeepers do not have any rating.

Matano, Richardson, Pospisil, Eubanks, and Qin (2018) considered a Bayesian framework for calculating adjusted plus-minus ratings. The model has independent variables only for representing the presence of players, and the observations are scaled based on their duration. However, instead of assuming that all player ratings should be shrunk towards zero, the prior for each player's rating is centered around a value based on a subjective score. The scores used are taken from a popular video game where each soccer player is given a rating. The authors show that the resulting ratings are better at predicting game results than the ratings from a basic multiple linear regression model estimated using ridge regression. In addition, both these versions of adjusted plus-minus ratings produce better game results predictions than just using the video game ratings.

Plus-minus for other sports

While plus-minus ratings have been in widespread use in ice hockey and basketball for some time, and while there is an increasing amount of relevant research being performed in the context of soccer, similar ideas are rare to find for other team sports. Volleyball is an exception, although only considered in a single study so far. Hass and Craig (2018) first described how to deal with the fact that a primary data collection tool failed to properly identify all the players on the court

at all given times. After compensating for this, they tested seven different plus-minus variants. These included the basic plus-minus, basic plus-minus per 50 plays, two versions of adjusted plus-minus based on Bayesian logistic regression as by Gramacy et al. (2013), and three versions of frequentist logistic regression. The frequentist regression models were estimated using ridge regression, lasso regression, and elastic net regression, respectively. The penalty terms of elastic net regression are given as a combination of the penalty terms of lasso regression and ridge regression. The observations corresponded to points scored. The study only calculated ratings for players on one team, but noted that the resulting ratings were found to be believable by the team's staff with few exceptions. Some additional details of the methods can be found in (Hass, 2017).

The term "plus-minus" has also been used in other settings, but without qualifying for consideration in this survey. Guryashkin (2012) wrote an article about a plus-minus inspired metric for boxing. The metric is based on calculating the percentage of punches landed for each boxer, and then finding a plus-minus score by taking the difference in percentages between the two boxers. As boxing is an individual sport, this metric does not capture the essence of plus-minus ratings, as defined in this survey. Similarly, some metrics proposed for American football have names that include "plus-minus", yet do not comply with the underlying philosophy of plus-minus ratings. One example is the "receiving plus-minus" from (Barnwell, 2009), measuring the number of passes caught by a receiver compared to what the average receiver would catch under similar circumstances. A second example is the "passing plus-minus" from (Kacsmar, 2016), measuring the number of passes completed by a quarterback compared to what an average quarterback would have completed given the same locations.

As there have been few contributions to plus-minus ratings in sports outside of basketball, ice hockey, and soccer, it may be useful to point out that these three main sports are quite different in aspects of great importance when considering plus-minus ratings. Table 1 summarizes some key characteristics for different sports. For example, in basketball there are typically several scoring events for both teams per segment of a match. In ice hockey, scoring events are much rarer, on the order of 10 minutes between goals. As players typically spend only about 30-60 seconds on the ice before being substituted, this means that hockey consists mostly of segments with zero goals scored (Thomas et al., 2013). For soccer, there are even fewer goals per match, but in addition there are only around six segments per match.

Table 1: Key characteristics of different sports

	Segments per match	Goals per segment	Players per team	Always even strength
Basketball	High	High	5	Yes
Ice hockey	Very high	Very low	5+1	No
Soccer	Very low	Low	10+1	No
Volleyball	High	Medium	6	Yes

The team structure also varies between sports. While basketball and volleyball have the same number of players on each team throughout each match, both ice hockey and soccer are influenced by player dismissals. In ice hockey these are temporary, but in soccer they last until the end of the match. For these latter sports, it is therefore necessary to take the manpower differential into consideration when estimating ratings. In soccer, attempts have been made to directly estimate the effect of missing players, while in ice hockey it has been more common to estimate separate ratings for short-hand and power play situations. In each sport, different players may have different roles. This is particularly evident in ice hockey and soccer, where a

single player is dedicated to defending the goal. Soccer is unique in that there is around twice as many players on each team as for the other sports.

Summary of literature

The initial research on plus-minus ratings was to a large degree published in channels without peer review, with many contributions simply being described in internet blog posts. Table 2 lists all the contributions that have not been subjected to peer review. This also includes PhD dissertations and Master theses, as well as some working papers that may be going through peer review as this text is being written. Table 3 lists the contributions to plus-minus ratings that have been published in peer-reviewed journals, books, and conference proceedings.

Figure 1 shows how the number of contributions to plus-minus ratings per year has evolved in the last two decades. In particular from around 2010, the amount of work has increased, and many new ideas have been introduced in the short time since. The figure also shows that the number of contributions presented without peer-review is relatively stable, however, the nature of these may have changed slightly, with an increase in the number of theses and working papers to compensate for a decrease in the number of blog posts.

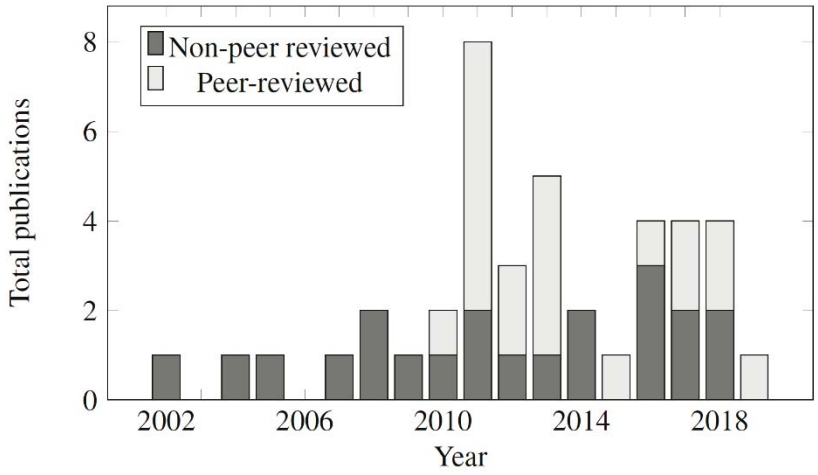


Figure 1: Number of publications per year, split into peer reviewed and non-peer reviewed contributions.

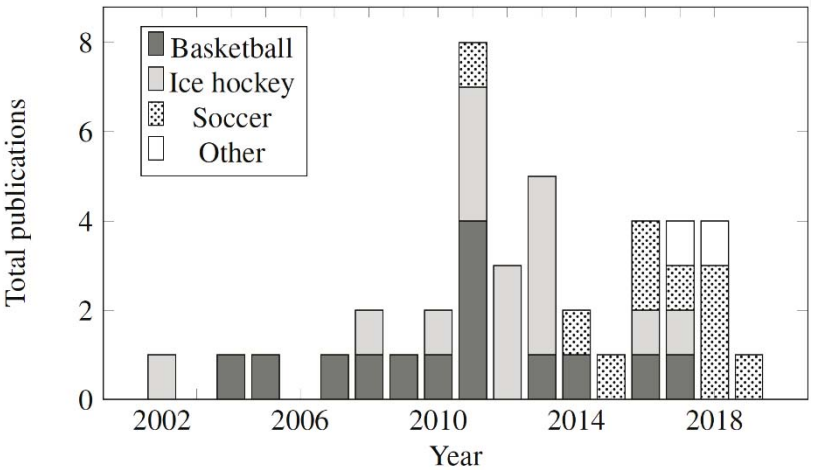


Figure 2: Number of publications per year, split into types of sport.

Table 2: Publications without peer review presenting plus-minus ratings.

Authors	Year	Sport	Performance Indictor	Statistical Model
Awad	2010a	ice hockey	net weighted shots	simple calculations
Bohrmann	2011	soccer	net goals	linear regression, least squares
Fyffe and Vollman	2002	ice hockey	net points	simple calculations
Hamilton	2014	soccer	net goals	linear regression, ridge
Hass	2017	volleyball	point won or lost	logistic regression, multiple types
Ilardi	2007	basketball	net points	linear regression, least squares
Ilardi	2014	basketball	points scored	unknown
Ilardi and Barzilai	2008	basketball	points scored	linear regression, least squares
Kharrat et al.	2018	soccer	net goals, net xG, win % change	linear regression, ridge
Macdonald et al.	2012	ice hockey	weighted shots (xG)	linear regression, ridge
Matano et al.	2018	soccer	net goals	Bayesian linear regression, Gaussian
Rosenbaum	2004	basketball	net points	linear regression, least squares
Rosenbaum	2005	basketball	net points	linear regression, least squares
Schuckers et al.	2011	ice hockey	net score	linear regression, least squares
Smith	2016	ice hockey	net weighted shot (xG)	linear regression, lasso
Spagnola	2013	ice hockey	goal scored or conceded	simple calculations with logistic regression
Sittl and Warnke	2016	soccer	net goals	hierarchical fixed effects, weighted least squares
Vilain and Kolkovsky	2016	soccer	goals scored	ordered probit regression, penalized likelihood
Warnke	2017	soccer	net goals	hierarchical fixed effects, weighted least squares
Winston	2009	basketball	net points, win % change	linear regression, least squares
Witus	2008	ice hockey	net and total points	linear regression, least squares

Table 3: Peer-reviewed papers presenting plus-minus ratings.

Authors	Year	Sport	Performance Indictor	Statistical Model
Deshpande and Jensen	2016	basketball	win % change	Bayesian linear regression, Laplacian
Engelmann	2011	basketball	net points	linear regression, ridge
Engelmann	2017	basketball	net points, various measures	linear regression, ridge
Fearnhead and Taylor	2011	basketball	points scored	Bayesian linear regression, Gaussian
Gramacy et al.	2017	ice hockey	goal scored or conceded	logistic regression, lasso
Gramacy et al.	2013	ice hockey	goal scored or conceded	logistic regression, lasso
Hass and Craig	2018	volleyball	point won or lost	logistic regression, multiple types
Macdonald	2011a	ice hockey	goals scored	linear regression, least squares
Macdonald	2011b	ice hockey	goals scored	linear regression, least squares
Macdonald	2012a	ice hockey	goals or shots	linear regression, ridge
Macdonald	2012b	ice hockey	exp. number of goals	linear regression, ridge
Okamoto	2011	basketball	net points	simple calculations
Omidiran	2011	basketball	net points	convex optimization, cross validation
Sæbø and Hvattum	2015	soccer	net goals	linear regression, ridge
Sæbø and Hvattum	2019	soccer	net goals	linear regression, ridge
Schuckers and Curro	2013	ice hockey	net probability of scoring	linear regression, ridge
Schultze and Wellbrock	2018	soccer	net goals	simple calculations
Sill	2010	basketball	net points	linear regression, ridge
Sisneros and Van Moer	2013	basketball	various measures	simple calculations
Thomas et al.	2013	ice hockey	time to next goal	hazard functions

In Figure 2 the contributions per year are broken down into different sports. It mainly illustrates that most early work happened within basketball, quickly followed by work within ice hockey, whereas most contributions within soccer are more recent. The only other sport to have been considered is volleyball.

In total, 41 contributions related to plus-minus ratings were identified in this survey. Out of these, 20 were published after peer review. Table 4 shows the most important publication channels where this work has been published. Two of the channels stand out, with the Journal of Quantitative Analysis in Sports containing seven papers, and the Proceedings of the MIT Sloan Sports Analytics Conference containing six papers.

Table 4: Outlets for peer-reviewed work on plus-minus ratings, ranked by the number of papers.

Publication channel	Papers	Years
Journal of Quantitative Analysis in Sports	7	2011–2018
Proceedings of the MIT Sloan Sports Analytics Conference	6	2010–2013
Journal of Sports Analytics	2	2018–2019
Handbook of Statistical Methods and Analyses in Sports	2	2017
The Annals of Applied Statistics	1	2013
IEEE Vis Workshop on Sports Data Visualization	1	2013
NIK: Norsk Informatikkonferanse	1	2015

Although not evident from the discussion of the literature in the previous sections, it seems that the literature on plus-minus ratings is somewhat fragmented. That is, most papers on plus-minus ratings cite very few other papers with contributions to plus-minus ratings. Although 41 contributions were identified in Tables 2 and 3, the average number of citations in those contributions to previous work on plus-minus ratings is only 3.8, and the median number of citations to other relevant work is 2. From this, one may be led to believe that most researchers working on plus-minus ratings are not well informed about previous contributions in the area, which makes this survey an important contribution in itself.

Knowing that this literature contains relatively few citations to other relevant contributions, there are still five papers that have received ten or more citations from later publications. These are listed in Table 5, and may thus be said to constitute the most central work on plus-minus ratings. Two of them have indeed been pivotal for the development of advanced plus-minus ratings: Rosenbaum (2004) was the first to publicly discuss the use of multiple linear regression to calculate adjusted plus-minus ratings, whereas Sill (2010) was the first to introduce regularization techniques to stabilize ratings.

Table 5: Most central work on plus-minus ratings, measured by citations. The table includes all publications with at least 10 citations from the papers listed in Tables 2 and 3.

Authors	Year	Citations	Contribution
Rosenbaum	2004	23	Multiple linear regression
Ilardi and Barzilai	2008	13	Both offensive and defensive ratings
Sill	2010	13	Regularization and cross validation
Macdonald	2012a	13	Shots as dependent variable
Macdonald	2011a	12	Early journal paper on plus-minus for ice hockey

Future work on plus-minus ratings

The initial research on plus-minus ratings relied on the use of multiple linear regression and ordinary least squares estimation. Starting with (Sill, 2010) it has become clear that better ratings can be found by either using a form of regularization or by using a Bayesian modelling technique. However, despite this important insight, there are still many open questions regarding how to best calculate plus-minus ratings. In the following, some of these open questions are discussed.

To compare different techniques for calculating plus-minus ratings, it is important to have a framework for evaluating the quality of the resulting ratings. This has rarely been problematized in the literature of plus-minus ratings. Franks, D'Amour, Cervone, and Bornn (2016) discussed how to compare different player metrics from a general perspective. They developed evaluation methods based on three criteria: 1) stability, 2) discrimination, and 3) independence. These relate to whether ratings are consistent over time, whether they are able to differentiate between players, and whether they can be shown to measure something that is not captured by other ratings. One of the metrics illustrated by Franks et al. (2016) is the basic plus-minus rating in ice hockey, which unsurprisingly is shown to score badly with respect to discrimination. The authors also mention a fourth criterion: relevance. It is suggested that this could be, for example, a description of the meaning and value of the metric, or a summary of the relationship between the metric and an outcome of interest, such as wins or revenue generated.

There is at this point no common method to measure the quality of different plusminus ratings. s most contributions have been made independently, many have resorted to compare their plus-minus ratings with measures based on transfer fees, player salaries, player awards, or ratings by journalists. Others have used objective measures, such as the ability of ratings to predict future results or future goal differences. This may then allow comparisons with predictions from betting market odds, which forms a useful benchmark. Nevertheless, it remains an open question to determine the best way of measuring the quality of plus-minus ratings.

Given a suitable method to compare different plus-minus ratings, some more detailed questions arise. For example, what is the best choice of the dependent variable, what is the best choice of defining what an observation is, and how does the answer to these questions depend on the characteristics of the particular sport studied? For the choice of observation, past research has proposed to use segments of a match without substitutions or dismissals, goals, or shots. In the case of segment-based observations, the dependent variable has typically been either the change in score, the change in win probability, or the difference in accumulated quality of shots or chances produced. In the case of observations based on goals or shots, the dependent variable is naturally chosen to represent which team has produced the goal or shot.

Another important question when specifying a plus-minus model seems to be which independent variables to include. There seems to be some types of variables that are statistically significant but has little effect on the actual player ratings calculated. An example of this is the home field advantage variable for soccer, which is commonly found to be positive for the home team. However, since most players have a similar amount of appearances both away and at home, the effect tends to cancel out and ratings found without compensating for home advantage are equally good. Other variables may be expected to have a larger effect on the player ratings calculated, such as age variables for models with data spanning several seasons. A somewhat controversial choice is whether or not to include team effects, and if included, how to determine their effect on the player ratings. It is likely that novel types of independent variables may still be found and included in plus-minus models, both to improve the player ratings, but also to analyze the effects of the specific variables themselves.

Finally, an open question is to determine the limits of the capabilities of plus-minus ratings. Considering plus-minus ratings as top-down ratings, that only require relatively rudimentary data from each match, it remains to see how well they compare to bottom-up ratings that exploit detailed event and tracking data from each match. Very few comparisons have been made between the two types of ratings yet, and it is unclear whether they end up measuring different aspects of players, and as such may exhibit a certain degree of independence. That is, will bottom-up ratings dominate plus-minus ratings in the end, or do they have distinct uses? In the latter case, it may turn out that hybrid rating systems can exploit the advantages of both types of rating.

Concluding remarks

Objective player ratings have multiple areas of use, ranging from entertainment purposes to being used as decision making support in high-stakes player transfers. Plus-minus ratings are based on the idea that the performance of a team can be distributed onto individual players, without knowledge of the exact player actions taking place within matches. The basic form of plus-minus ratings are known to have several deficiencies, but an increasing amount of research has lately provided insights into how particular types of regression models and estimation techniques can be used to overcome some of these inadequacies. Research into improved player rating models, and plus-minus ratings in particular, appears to be increasing.

This paper has presented 41 contributions to plus-minus ratings, out of which 20 are peer-reviewed publications. These have discussed plus-minus ratings for individual players in ice hockey (15 contributions), basketball (14 contributions), soccer (10 contributions), and volleyball (2 contributions). Key papers were identified as (Rosenbaum, 2004), first describing the use of regression to find plus-minus rating, (Sill, 2010), first using regularization when estimating ratings from a regression model, and the work of Macdonald (2011a, 2012a) who popularized the use of adjusted plus-minus ratings in ice hockey.

The literature review highlights that many contributions to plus-minus ratings are made independently, or at least without explicit references to much of the existing work: a typical contribution to plus-minus only cites two other contributions to plus-minus ratings. However, there is no need to reinvent plus-minus ratings separately for each sport, or several times for the same sport: certain key properties of the ratings can be understood by examining existing literature, and there are plenty of open research questions that deserve attention.

Acknowledgements

The author is grateful for the insightful comments and suggestions from two anonymous reviewers. Their comments helped to improved an initial version of the paper.

References

Awad, T. (2010a). Delta with teammate adjustments – DeltaSOT. http://www.hockeyprospectus.com/ puck/article.php?articleid=454, accessed 2018-09-05.

Awad, T. (2010b). Plus-minus and Corsi have a baby. http://www.hockeyprospectus.com/ puck/article.php?articleid=436, accessed 2010-10-24.

Barnwell, B. (2009). Receiving plus/minus, part I. https://www.footballoutsiders.com/stat-analysis/2009/receiving-plusminus-part-i, accessed 2018-09-19.

Bohrmann, F. (2011). Problems with an adjusted plus minus metric in football. http://www.soccerstatistically.com/blog/2011/12/28/problems-with-an-adjusted-plusminus-metric-in-football.html, accessed 2018-09-13.

Constantinou, A. & Fenton N. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 9, 37–50.

Deshpande, S. & Jensen, S. (2016). Estimating an NBA player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, 12, 51–72.

Elo, A. (1978). The Rating of Chessplayers, Past and Present. New York: Arco Publishing.

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

Engelmann, J. (2017). Possession-based player performance analysis in basketball (adjusted +/– and related concepts). In: Albert, J., Glickman, M., Swartz, T., & Koning, R., eds., Handbook of Statistical Methods and Analyses in Sports, Boca Raton: Chapman and Hall/CRC, 215–228.

Fearnhead, P. & Taylor, B. (2011). On estimating the ability of NBA players. *Journal of Quantitative Analysis in Sports*, 7, https://doi.org/10.2202/1559-0410.1298.

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.

Fyffe, I. & Vollman, R. (2002). Improving plus-minus. http://www.hockeythink.com/research/plusmin.html, accessed 2011-12-19.

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.

Gramacy, R., Taddy, M., & Tian, S. (2017). Hockey performance via regularized logistic regression. In: Albert, J., Glickman, M., Swartz, T., & Koning, R., eds., Handbook of Statistical Methods and Analyses in Sports, Boca Raton: Chapman and Hall/CRC, 287–306.

Guryashkin, I. (2012). Mayweather measures up with greats. http://www.espn.co.uk/boxing/story/_/id/7780088/floyd-mayweather-jr-measuresboxing-greats, accessed 2018-09-19.

Hamilton, H. (2014). Adjusted plus/minus in football - why it's hard, and why it's probably useless. http://www.soccermetrics.net/player-performance/adjusted-plusminus-deep-analysis, accessed 2018-09-13.

Hass, Z. (2017). Division of credit modeling for team sports with an emphasis on NCAA volleyball. Ph.D. thesis, Purdue University.

Hass, Z. & Craig, B. (2018). Exploring the potential of the plus/minus in NCAA women's volleyball via the recovery of court presence information. *Journal of Sports Analytics*, 4, 285–295.

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

Ilardi, S. (2007). Adjusted plus-minus: An idea whose time has come. http://www.82games.com/ilardi1.htm, accessed 2018-08-31.

Ilardi, S. (2014). The next big thing: real plus-minus. http://www.espn.com/nba/story/_/id/10740818/introducing-real-plus-minus, accessed 2018-09-05.

Ilardi, S. & Barzilai, A. (2008). Adjusted plus-minus ratings: new and improved for 2007-2008. http://www.82games.com/ilardi2.htm, accessed 2018-08-31.

Kacsmar, S. (2016). 2015 passing plus-minus. https://www.footballoutsiders.com/index.php?q=stat-analysis/2016/2015-passingplus-minus, accessed 2018-09-19.

Kharrat, T., Peña, J., & McHale, I. (2018). Plus-minus player ratings for soccer. ArXiv:1706.04943.

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3, article 1.

Lasek, J., Szlávik, Z., & Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, 1, 27–46.

Macdonald, B. (2011a). A regression-based adjusted plus-minus statistic for NHL players. *Journal of Quantitative Analysis in Sports*, 7.

Macdonald, B. (2011b). An improved adjusted plus-minus statistic for NHL players. Proceedings of the MIT Sloan Sports Analytics Conference.

Macdonald, B. (2012a). Adjusted plus-minus for NHL players using ridge regression with goals, shots, Fenwick, and Corsi. *Journal of Quantitative Analysis in Sports*, 8.

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

Macdonald, B., Lennon, C., & Sturdivant, R. (2012). Evaluating NHL goalies, skaters, and teams using weighted shots. ArXiv:1205.1746.

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. & Davies, S. (2007). Statistical analysis of the FIFA world rankings. In: Koning, R. & Albert, J., eds., Statistical Thinking in Sport, Boca Raton, FL: Chapman and Hall, 77–90.

McHale, I. & Forrest, D. (2005). The importance of recent scores in a forecasting model for professional golf tournaments. *IMA Journal of Management Mathematics*, 16, 131–140.

McHale, I. & Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27, 619–630.

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.

Okamoto, D. (2011). Stratified odds ratios for evaluating NBA players based on their plus/minus statistics. *Journal of Quantitative Analysis in Sports*, 7, Article 5, article 5.

Omidiran, D. (2011). A new look at adjusted plus/minus for basketball analysis. Proceedings of the 2011 MIT Sloan Sports Analytics Conference.

Pantuso, G. (2017). The football team composition problem: a stochastic programming approach. *Journal of Quantitative Analysis in Sports*, 13, 113–129.

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, 12(2proc), S514–S529.

Rosenbaum, D. (2004). Measuring how NBA players help their teams win. http://www.82games.com/comm30.htm, accessed 2018-08-31.

Rosenbaum, D. (2005). Defense is all about keeping the other team from scoring. http://82games.com/rosenbaum3.htm, accessed 2018-09-28.

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.

Schuckers, M. & Curro, J. (2013). Total hockey rating (THoR): a comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. Proceedings of the MIT Sloan Sports Analytics Conference.

Schuckers, M., Lock, D., Wells, C., Knickerbocker, C., & Lock, R. (2011). National Hockey League skater ratings based upon all onice events: an adjusted minus/plus probability (AMPP) approach. http://myslu.stlawu.edu/˜msch/sports/LockSchuckersProbPlusMinus113010.pdf.

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

Sill, J. (2010). Improved NBA adjusted +/– using regularization and out-of-sample testing. Proceedings of the 2010 MIT Sloan Sports Analytics Conference.

Sisneros, R. & Van Moer, M. (2013). Expanding plus-minus for visual and statistical analysis of NBA box-score data. In: Proceedings of IEEE Vis Workshop on Sports Data Visualization.

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.

Smith, G. (2016). A shot quality adjusted plus-minus for the NHL. Master's thesis, University of Toronto.

Spagnola, N. (2013). The Complete Plus-Minus: A Case Study of The Columbus Blue Jackets. Master's thesis, University of South Carolina.

Stefani, R. & Pollard, R. (2007). Football rating systems for top-level competition: A critical survey. *Journal of Quantitative Analysis in Sports*, 3, Article 3, article 3.

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. (2010). Assessing the performance of German Bundesliga football players: a non-parametric metafrontier approach. *Central European Journal of Operations Research*, 19, 571–587.

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.

Warnke, A. (2017). Essays on Gender Differences in Training, Incentives and Creativity, Survey Response, and Competitive Balance and Sorting in Football. Ph.D. thesis, University of Freiburg.

Winston, W. (2009). *Mathletics*. Princeton, New Jersey: Princeton University Press.

Witus, E. (2008). Offensive and defensive adjusted plus/minus. http://www.countthebasket.com:80/blog/2008/06/03/offensive-and-defensive-adjustedplus-minus/, accessed 2009-03-31.