Quarterback Evaluation in the National Football League using Tracking Data

Matthew Reyers and Tim B. Swartz *

Abstract

This paper evaluates quarterback performance in the National Football League. With the availability of player tracking data, there exists the capability to assess various options that are available to quarterbacks and the expected points gained resulting from each option. The options available to a quarterback are based on considering multiple frames during a play such that a current option may evolve into new options over time. Our approach also considers the possibility of quarterback running options. With tracking data, the location of receivers on the field and the openness of receivers are measurable quantities which are important considerations in the assessment of quarterback options. Machine learning techniques are then used to estimate the probabilities of success of the passing options and the estimated expected points gained from the options. The estimation procedure also takes into account what may happen after a reception. The quarterback's observed execution is then measured against the optimal available option.

Keywords : expected points, machine learning, model validation, player tracking data.

*M. Reyers is an MSc candidate and T. Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Both Reyers and Swartz have been partially supported by the Natural Sciences and Engineering Research Council of Canada. The authors thank two reviewers whose detailed comments helped improve the manuscript.

1 INTRODUCTION

The National Football League (NFL) is the top revenue league in sport (Amoros 2016) with an average team revenue of $453,000,000 in the 2017 season (Gough 2018). Despite the big money nature of the NFL, football analytics trails some of the other "big" professional sports including basketball (the National Basketball Association), soccer (major European leagues) and baseball (Major League Baseball). For a survey of some of the work that has been done in sports analytics, see Albert, Glickman, Swartz and Koning (2017).

The analytics landscape in the NFL is beginning to change as Next-Gen-Stats' player tracking data was introduced in 2016 and was made available to all 32 NFL teams in 2018. Player tracking data is detailed spatio-temporal data where the locations of each player on the field are recorded 10 times per second. This type of data leads to analytics opportunities that were previously unthinkable in the era of boxscore data. Subsets of the data have been released by the NFL in a yearly competition known as the Big Data Bowl (https://operations.nfl.com/the-game/big-data-bowl/) which is an analytics event held in conjunction with the NFL Scouting Combine. The availability of the player tracking data has led to recent research in NFL analytics and includes Burke (2019), Chu et al. (2020), Deshpande and Evans (2020), Yam and Lopez (2020) and Yurko et al. (2020).

A traditional NFL statistic is the quarterback passer rating. The passer rating (Zilavy 2018) is a complex formula which does not yield a straightforward interpretation; the formula provides a minimum rating of 0 and a maximum rating of 158.3. It is a curious fact that amongst the top 12 quarterbacks in 2019 (according to passer rating), all 12 of these quarterbacks played for the 12 NFL playoff teams. An immediate reaction to this observation is that a team must have an outstanding quarterback in order to make the playoffs. Another explanation is that a quarterback's rating is highly dependent on the quality of his team. This is the motivation for our research; we attempt to introduce a quarterback metric which is less dependent on the performance of one's teammates. It is possible that there are some excellent quarterbacks who play on weaker teams and their

worth is not fully appreciated.

To investigate quarterback performance, we use the player tracking data previously mentioned. On each passing down, we identify options that are available to the quarterback. This is a novel investigation as it requires the enumeration of options (both passing and running) prior to the quarterback's actual decision, and even options that may have eventuated after his decision. The options available to a quarterback are based on considering multiple frames during a play such that a current option may evolve into new options over time. Then, each of these options are assessed an EPV (expected point value) depending on the state of the game. Further, probabilities of the successful execution of these options are estimated. These components then allow us to formulate a metric which compares actual outcomes versus optimal options. An advantage of this approach is that the quarterback metric is less dependent on one's teammates - some quarterbacks will have better options than other quarterbacks, and the metric is formulated such that a quarterback's performance is only compared against his available options. Another advantage is that the metric takes running into account unlike the traditional quarterback passing rating. The running abilities of quarterbacks such as Russell Wilson, Lamar Jackson, Deshaun Watson and Michael Vick have been a great benefit to their respective NFL teams.

In terms of completing a catch, there have been various investigations (Burke 2019, Deshpande and Evans 2020) and vendors (Next Gen Stats Team 2018) that provide probabilities of pass completion. These approaches typically look at the circumstances at the time of the catch (e.g. the location of defenders, where the ball is thrown, etc.). We emphasize an important novelty of our approach where completion probability is assessed *at the time that the ball is thrown* - this involves more uncertainty as it is unclear how the play will develop downfield.

Burke (2019) uses neural networks to predict the targeted receiver. The covariates chosen are different from ours and consideration is only given to passing options at the instant that the pass is made. In assessing quarterbacks, Burke (2019) uses expected yards; we believe that EPV is a more relevant measure in football since it incorporates

context. For example, gaining 7 yards on first down and 10 yards is a much better outcome than gaining 7 yards on third down and 10 yards. Unlike Burke (2019), we also introduce some stochasticity in our approach; this facilitates error assessment. Importantly, our approach also takes into account non-designed quarterback runs and interceptions; intercepted passes form a critical component of the outcome of games.

In Section 2, we begin with a broad overview of our approach. The idea is that we identify options that are available to a quarterback, we assign value to these options, and we then compare the actual results versus the optimal options. This results in a quarterback rating that is less dependent on teammates. Recall that better teammates typically generate better options and we only consider the options that are available. For example, a receiver who is fast and able to quickly change directions will likely provide better quarterback passing options. On the other hand, highly accomplished defensive backs will likely reduce quarterback passing options. In Section 3, details of the procedure are provided. We describe the player tracking data, propose covariates and then use machine learning algorithms to determine the probabilities associated with the options available to the quarterback. The probabilities are then validated against holdout data. In Section 4, the methods are applied and ratings are obtained for NFL quarterbacks. The ratings generally agree with popular opinion although they reveal some surprises; there are some quarterbacks held in high esteem who are not rated so highly, and vice-versa. We conclude with a short discussion in Section 5.

2 OVERVIEW OF THE APPROACH

Consider a particular quarterback and all of his passing and running options on a play that was not a designed run. For the ith play, the quarterback executes a decision at time t_i. For the time interval $t \in (0, t_i + \epsilon]$, we consider all $j = 1, \ldots, n_i$ options that were available to the quarterback. An option is defined as either a non-designed quarterback run or a potential pass to an eligible receiver (i.e. wide receiver, running back, tight end, halfback, fullback as coded in the dataset). The players are followed throughout the

time sequence where each combination of player and time increment results in an option. No doubt, inferences become more difficult for larger values of ϵ since players alter their patterns once $t > t_i$. For example, players tend to slow down once a pass is initiated and they realize that they will not be active in the play. In Section 4, we set a small window $\epsilon = 0.5$ seconds.

We denote p_{ij} as the probability that the jth option on the ith play is a completion where all running plays are treated as completions. Recall that the options $j = 1, \ldots, n_i$ are identified over all frames of the tracking data. The quantity p_{ij} is an unknown parameter which we estimate by $\hat{p}_{ij}$ using machine learning methods. We let G_{ij} denote the corresponding expected points gained from the successful execution of option j on play i. Similarly, we define $\mathrm{G}_{ij}^{(U)}$ as the expected points gained from the unsuccessful execution of option j on play i. Note that unsuccessful execution of option j implies that $\mathrm{G}_{ij}^{(U)} < 0$. Since all unsuccessful pass options result in no yards gained, we further define $\mathrm{G}_{i}^{(U)} = \mathrm{G}_{i1}^{(U)} = \cdots = \mathrm{G}_{in_i}^{(U)}$. Expected point values were developed by Yurko, Ventura and Horowitz (2019) and take into account both field position and down. We have utilized the *nflscrapR* software (Horowitz, Yurko and Ventura 2020) to evaluate EPV. For example, suppose that your team is faced with first down and 10 yards at your own 20 yard line. The EPV is 0.40, indicating that on average a team will gain 0.4 points on the set of possessions following this state. Your team then completes a 6 yard pass and is faced with second down and 4 yards at your own 26 yard line. The EPV from the updated state is 0.69, and therefore the expected points gained from the completed pass is $G = 0.69 - 0.40 = 0.29$.

Taking into account the completion probability p_{ij} of the jth option on play i, the value from making the optimal decision is therefore given by

$$Y_i = \max \left[(1 - \hat{p}_{i1}) G_i^{(U)} + \hat{p}_{i1} \hat{\mathrm{G}}_{i1} \ , \ldots, \ (1 - \hat{p}_{in_i}) G_i^{(U)} + \hat{p}_{in_i} \hat{\mathrm{G}}_{in_i} \right] \tag{1}$$

where each term in (1) is an estimated expected value. In (1), we emphasize that we have used the notation $\hat{G}$ since we need to estimate the yards gained after a potential

completion in determining EPV.

Now, corresponding to play $i = 1, \ldots, N$ for the given quarterback, we can calculate the actual expected points gained A_i. This is obtained by taking the difference between the EPV value before and after the play. We therefore propose the quarterback metric

$$Q = \left(\frac{\sum_{i=1}^{N} A_i}{\sum_{i=1}^{N} Y_i} \right) 100\% \ . \tag{2}$$

Unlike the NFL passer rating, we note that Q is interpretable. Recall that the sum $\sum_i Y_i$ denotes the maximal EPV gained by an average quarterback (in terms of execution) who is always making the best decisions. Therefore, a score of Q represents the execution percentage relative to this hypothetical quarterback. As a measure of quarterback performance, the metric Q combines both the fundamental elements of decision making and execution. And again, we emphasize that a feature of the metric (2) is that the basis of comparison involves the options that are available to the quarterback. Different quarterbacks have different options, and we might expect quarterbacks on better teams to have better options. There is flexibility in choosing N so that it corresponds to the entire season, a segment of the season or even particular games.

Figure 1 provides a plot of the change over time of the criterion $(1 - \hat{p}_{ij})G_i^{(U)} + \hat{p}_{ij}\hat{\mathrm{G}}_{ij}$ for five receivers during a passing play involving the New England Patriots in a match against the Kansas City Chiefs on September 7, 2017. We see that as the play develops, the criterion for the five players changes. Increases occur when a player becomes more open and as a player moves further downfield. On this play, Dwayen Allen was the intended receiver and the pass was incomplete. According to the criterion, passing to Allen appeared to be the correct decision.

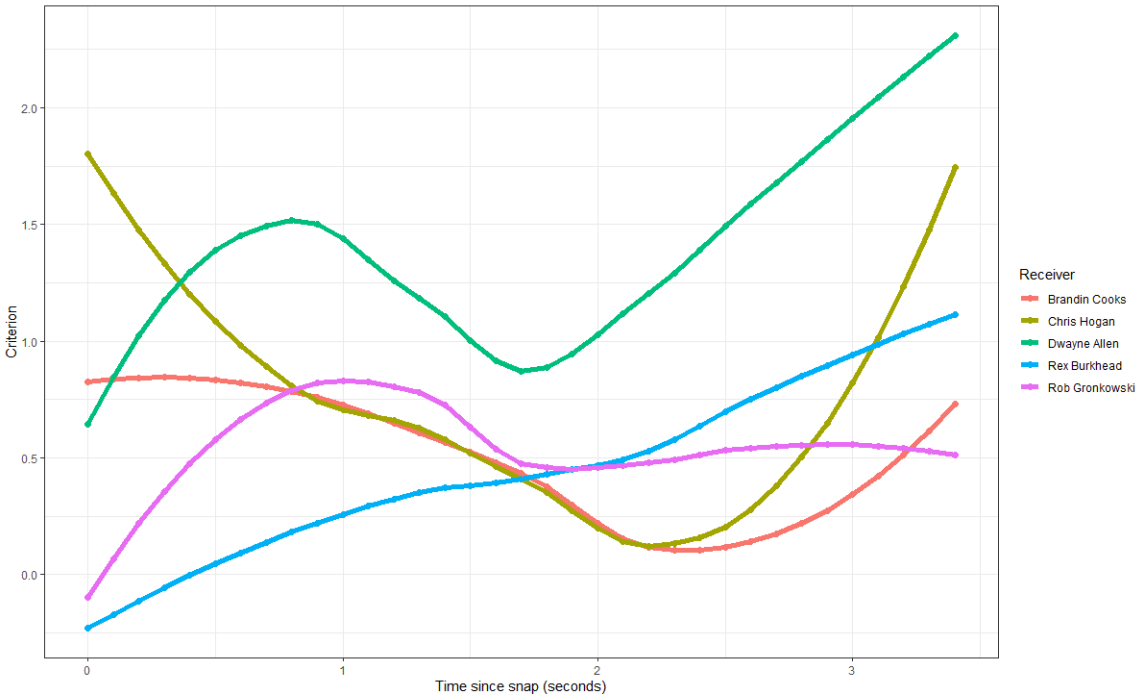

Figure 1: Example of the change in the criterion $(1-\hat{p}_{ij})G_i^{(U)} + \hat{p}_{ij}\hat{\mathrm{G}}_{ij}$ (smoothed) for five receivers from the New England Patriots during a passing play from September 7, 2017.

3 DETAILS OF THE APPROACH

3.1 Data

The data used in this investigation were provided by Next Gen Stats. Released in 2019, the data cover the first six weeks of the 2017 NFL season. This subset of the season includes five or six games per team, dependent on whether teams had been assigned a bye week. This leads to a total of 91 games for which there are 6960 passing plays. These plays were augmented with 252 non-designed quarterback runs and 452 sacks. After removing problematic tracking data, the cleaned dataset consists of 6727 plays of interest.

At a more granular level, each play contains measurements on 23 unique actors on the field: 11 offensive players, 11 defensive players and the football. Measurements for

each actor were recorded 10 times per second. The data were collected by Next Gen Stats and its partner organizations Zebra Technologies and Wilson Sporting Goods (see https://operations.nfl.com/thegame/technology/nfl-next-gen-stats/). Each player measurement includes detailed information about movement including velocity, direction, distance travelled since the last frame, acceleration, and position. Similar measurements are available for the football in the same format.

The primary motivation of this investigation is the assessment of quarterbacks. However, there is a contextual aspect to the evaluation where it is well-known that teams have dramatically different styles depending on the circumstances of the game. For example, in "garbage time", a team will stop throwing the ball when they lead by an insurmountable margin. To address these less competitive situations, we use the win probability calculation in nflscrapR (Horowitz, Yurko and Ventura 2020), and we omit plays where the win probability falls outside the range $(0.1, 0.9)$. This further reduces the number of pass attempts and non-designed runs in our dataset to 5276.

3.2 Covariates

A core problem in the development of our methods involves the estimation of the success probability p_{ij} corresponding to the jth option on the ith play. Our modeling will attempt to capture the covariates that influence the completion of a pass attempt. Since the eventual goal concerns quarterback evaluation involving decision making, completion probability is assessed at the time the ball is released rather than when it arrives. Therefore, some variables that are relevant at the time when the ball arrives (e.g. receiver separation from defenders) will be estimated at the time of release.

Previous work (Next Gen Stats Team 2018) has explored the modeling of completion probability. Their work highlights the relationship between factors such as pass air distance, air yards, receiver separation, pass rush separation, and the speed of the quarterback at release. There are other covariates included in their modeling but these have not been publicly disclosed. Unfortunately, many of the modeling details remain proprietary

and cannot be reviewed.

Deshpande and Evans (2020) also model completion probability. They leverage a collection of factors including receiver separation from the nearest defender and from the ball, receiver movement vectors, and cumulative distance covered by the receiver during the game. These covariates are essentially doubled up, being measured both at the time of release and at the time of pass arrival. Their model which is based on Bayesian additive regression trees generates 90% prediction accuracy with respect to completions. For a completion (incompletion), the prediction is defined as accurate if the completion probability is greater (less) than 0.5.

We use similar covariates to the aforementioned work with a few additions. However, we emphasize that we only make use of covariates that were measurable at the time of the throw since we wish to focus on quarterback decision making. We now introduce the covariates such that for every play i and option j, there is a specified potential receiver. For the time being, we omit running options.

For convenience, in Table 1, we summarize the covariates that are subsequently introduced in subsections 3.2.1, 3.2.2 and 3.2.3.

Football	Receiver	Quarterback
air distance*	receiver separation	rush separation
yards downfield*	sideline separation	time from snap
	field ownership*	quarterback speed
		pocket indicator variable

Table 1: Covariates used in the analyses corresponding to subsections 3.2.1, 3.2.2 and 3.2.3. An asterisk indicates that the covariate is determined at the time of ball arrival as opposed to the time corresponding to the current frame.

3.2.1 Football covariates

The two football covariates that we consider are similar to those used in previous completion probability models. The first football covariate is air distance. Given the intended

receiver, we calculate the Euclidean distance that the football needs to travel. This is a measurement between final ball location and quarterback location at the time of the pass. Intuitively, longer passes have lower probabilities of completion. The final ball location is estimated using the receiver velocity and the ball velocity. We use a fixed value of 20 yards per second for all ball velocity calculations.

The second covariate is yards downfield. This is similar to air distance but only considers yardline distance. The covariate adds football context as the number of yards gained is relevant to scoring. Also, the probability of a pass completion may depend on the angle that the ball is thrown. For example, a pass 10 yards to the side of the quarterback typically has a higher pass completion probability than a 10 yard pass directly downfield.

3.2.2 Receiver covariates

Generally, the more open the receiver, the higher the completion probability. We attempt to characterize openness with three covariates. The first two are similar to those in other completion probability models whereas the remaining covariate is novel.

The first covariate is receiver separation from the nearest defender. This is obtained by calculating the minimum Euclidean distance between the receiver and all players on defence at the time that the pass is initiated.

A second covariate is the sideline separation distance at the time of release. A pass is complete only if the receiver establishes control of the ball inbounds and the sideline is used to mark the edge of the inbounds surface. If there is little space along the sideline, this reduces the completion probability.

Although receiver separation provides information on openness, we also introduce a field ownership metric which utilizes the positions and velocities of receivers and defenders. The resultant covariate extends the notion of receiver separation beyond the consideration of a single defender. The field ownership metric is adapted using ideas from Fernandez and Bornn (2018) which were developed for soccer. We begin by estimating the probability densities of the location of players at the time of ball arrival. The densities are based on kinesiological ideas such as the recognition that it is more difficult for players to change

directions at higher speeds. A team's ownership at a given location is then the sum of the individual densities for that team's players at that location. Influence at a given location is then calibrated on the interval $[0, 1]$ where a value of 0.5 is interpreted as equal location ownership by both teams. An owned cell by the offensive team is one for which influence > 0.5. We then define the field ownership covariate as the total influence of offensive owned cells within five yards of estimated ball arrival. We note that the threshold of five yards may be regarded as an assumption where it may make more sense to have larger thresholds for passes that are further downfield.

3.2.3 Quarterback covariates

The success of a passing play depends on more than just the receiver and his ability to get open. In addition, there is a reliance on the offensive line to provide ample time for the quarterback while also minimizing required quarterback movement. We aim to capture these notions via the four following quarterback covariates which are similar to existing covariates in the literature. Calculation of the covariates is done on a frame by frame basis to assess hypothetical passes.

We define the covariate rush separation as the Euclidean distance between the quarterback and the nearest defensive opponent. This accounts solely for physical closeness and does not consider the estimated time it takes the defender to reach the quarterback.

We also measure the time to throw covariate which is the time from the snap to the current observed frame. Generally, a quarterback is under more duress as time progresses.

The covariate quarterback speed is estimated from his change in position between the current frame and the frame observed 0.5 seconds prior. It is generally more difficult for a quarterback to complete a pass when he is moving faster.

Finally, the distance from the pocket covariate uses the positioning of the quarterback relative to a 7 yard by 7 yard square bordering the line of scrimmage. The covariate is set to 0 if the quarterback is within the pocket; otherwise it is the minimum distance for the quarterback to re-enter the pocket. The intuition is that it is easier to make a pass from the pocket.

3.3 Modeling and Estimation

3.3.1 Estimation of the p_{ij}'s

The estimation of the completion probabilities p_{ij} requires a statistical learning approach that is flexible (e.g. non-linear) to accommodate the non-linearity and multicollinearity of the covariates. We utilize a Stacking algorithm built on an ensemble of base learners including random forests, gradient boosting, general linear models, logistic regression, neural networks and naive Bayes. At the super learner level we incorporate a gradient boosting model. This treats the cross-validated predictions generated by the base learners as covariates (van der Laan, Polley and Hubbard 2007). Although there are many choices at the super learner level, we found that gradient boosting offers the best predictive performance for our problem (Naimi and Balzer 2017). Note that the prediction exercise is more challenging in our context where covariates were obtained at the time of the throw rather than at the time of arrival of the pass.

3.3.2 Estimation of yards gained after the catch

To model the yards gained after the catch, we restrict the dataset to the 3933 instances where the pass was completed. In addition to original covariates used in the completion probability model (Section 3.3.1), we introduce two new covariates that describe the presence of tacklers "downfield" where downfield encompasses all defenders beyond the receiver. The first covariate estimates the distance of the nearest downfield defender to the intended receiver at the time of ball arrival. This is based on the velocities of the two players and the average speed of a pass. The second covariate is the estimated number of defenders downfield at the time of ball arrival. With more separation from the nearest downfield defender and fewer tacklers downfield, there is an expectation of a greater number of yards after the catch.

We use the same class of base learners as in the completion probability model (Section 3.3.1) with slight modifications for a regression task rather than a classification task. We use a non-negative generalized linear model (GLM) as the super learner which combines

the base learners. For the test dataset, the root mean squared error corresponding to the fitted yards after the catch compared to the actual yards after the catch is 2.96 yards.

3.3.3 Estimation of yards gained from non-designed runs

Non-designed quarterback runs make up a small proportion of our observed plays (only 252 plays). Therefore, building a training and testing set to assess model fit would likely lead to overfitting. Instead, we treat the yards gained from non-designed quarterback runs as similar to yards gained after the catch, and we derive our estimates from the respective model. The root mean squared error corresponding to these plays is 3.99 yards.

3.3.4 Handling interceptions

Modeling thus far has considered a pass outcome as binary - either a completion or an incompletion. This was formulated with interceptions treated as incomplete passes. Although this is sensible from the perspective of estimating completion probability, it is inadequate to equate incompletions with interceptions in terms of EPV. Generally, an interception is far more damaging to the offensive team than an incompletion.

The introduction of interceptions complicates the simple formulation (1) involving the optimal expected points gained on the ith play. Denote $\hat{q}_{ij}$ as the estimated probability of an interception corresponding to passing option j on play i. Then equation (1) is modified by replacing the jth term $\hat{p}_{ij}\hat{G}_{ij}$ in (1) by

$$\hat{p}_{ij}\hat{G}_{ij}^{(\text{comp})} + \hat{q}_{ij}\hat{G}_{ij}^{(\text{int})}$$

where $\hat{G}_{ij}^{(\text{comp})}$ is new notation for the expected points gained from a completion and $\hat{G}_{ij}^{(\text{int})}$ is the expected points gained from an interception. Note that the expected points corresponding to an incompletion is constant across all options on a given play.

With our restricted dataset involving only 158 interceptions, it is challenging to estimate the probabilities q_{ij} of an interception with a comprehensive categorical model that includes completions, incompletions and interceptions. For this reason, we analyze

interceptions separately using the same approach and covariates as in the completion probability model of Section 3.3.1.

Due to the lack of data, it is also difficult to reliably estimate yards gained after an interception. Therefore, we assign no yards gained following an interception. Although this is an unrealistic assumption, we note that interceptions are rare events where the probabilities q_{ij} are small and do not affect Y_i in (1) greatly. With more data, yards gained after an interception could be better estimated with a larger dataset using the ideas from Section 3.3.2.

The same principles can be applied for the analysis of quarterback fumbles for non-designed runs in Section 3.3.3. Fumbles on non-designed quarterback runs are even more rare in our dataset with only a single occurrence resulting from 252 runs. For the time being, we omit the consideration of this possibility.

3.4 Validation

For the completion probability model (Section 3.3.1), we randomly split the data into a training set (85%) and a validation set (15%) where base learners and weights were determined using 10-fold cross-validation on the training data. Recall that a gradient boosting super learner was utilized. Model performance was then tested on the held-out validation data which generated an accuracy rate of 75.4% using $p = 0.5$ as the cutoff for classification. In addition, Figure 2 provides a plot of observed completion probability versus expected completion probability (ECP) where ECP is divided into 10 cells with roughly the same number of observations. The proximity of the points to the line $y = x$ suggests that the model is performing well.

For the yards after the catch model (Section 3.3.2), we again randomly split the data into a training set (85%) and a validation set (15%) where base learners and weights were determined using 10-fold cross-validation on the training data. Recall that a non-negative GLM super learner was utilized. Model performance was then tested on the held-out validation data which generated a root mean squared error of 3.64 yards. We

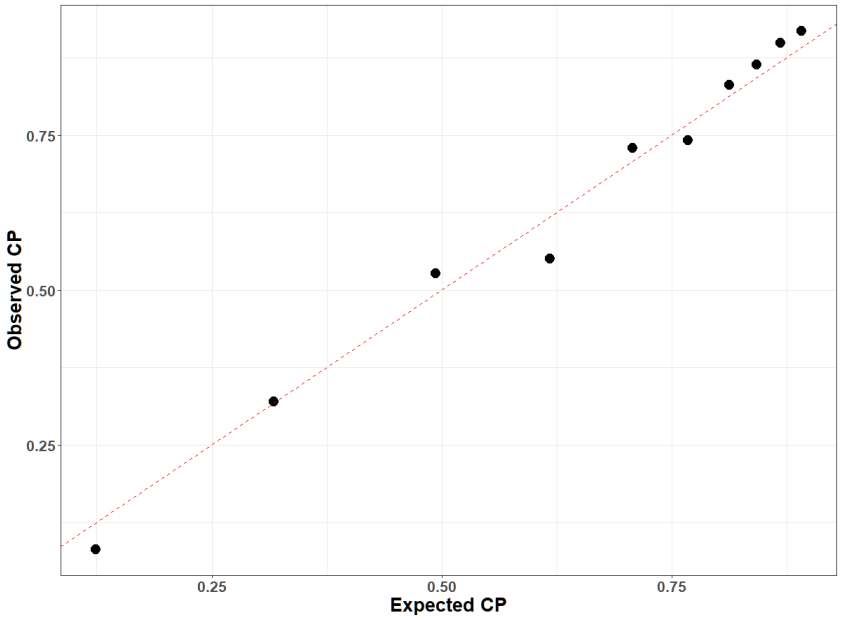

Figure 2: Scatterplot of observed completion probability versus expected completion probability based on binning the expected completion probabilities into 10 cells. The line $y = x$ is superimposed.

also report that 83% of the observations fell within five yards of the yards after the catch prediction.

4 RESULTS

Using the proposed models, we predict the completion probability and the yards gained after the catch for each option on all passing plays. Then using the EPV tables, this permits the calculation of the quarterback execution metric Q given by (2).

To provide some additional insight, we calculate Q under two conditions to highlight the impact of mobile quarterbacks through non-designed quarterback runs:

- Q_1: non-designed runs removed from the dataset
- Q_2: all potential passing plays (i.e. pass plays and non-designed runs)

It is important to emphasize the distinction between the metrics Q_1 and Q_2. First, we note that all designed quarterback runs have been removed from the dataset. The rationale is that these are planned running plays (and although potentially valuable), there is no decision making involved and the quarterback is taking the role of a running back. However, in Q_2, we do include non-designed quarterback runs (e.g. scrambles) as these are options that are available to the quarterback. All quarterback runs (i.e. designed and non-designed) are removed in the calculation of Q_1, and consequently, Q_1 is a measure that only considers the passing component available to quarterbacks.

In Table 2, we report the statistics Q_1 and Q_2 for the 29 quarterbacks who had at least 100 potential passing plays in the first six weeks of the 2017 NFL season. The statistic Q_1 corresponds to only passing options whereas the statistic Q_2 incorporates both passing and running. One of our first observations from Table 2 is that there is some disagreement between Q_1 and the NFL Passer Rating. If we look at the six teams who had quarterbacks with passer ratings exceeding 100, we observe that these teams had fast starts in 2017. Specifically, after the first six weeks of the season, Kansas City was 5-0, Philadelphia was 5-1, New England was 4-2, New Orleans was 3-2 and the LA Rams were 4-2. This is again suggestive that the NFL Passer Rating is largely a function of team success. On the other hand, our statistic Q_1 incorporates performance with decision making. We see that the top quarterback according to the passing statistic Q_1 is Dak Prescott with $Q_1 = 44.5$ and at the bottom of the list is DeShone Kizer with $Q_1 = 24.5$. With Dak Prescott, the interpretation of the statistic Q_1 is that over the first six weeks of the 2017 NFL season, when only considering passing plays, his EPV contribution was 44.5% of the hypothetical quarterback who made optimal decisions on every play. We also observe that Q_1 does not correlate strongly with the NFL Passer Rating ($r = 0.51$).

When we look at the overall quaterback rating Q_2 in Table 2 which includes non-designed runs, we observe that Russell Wilson had the greatest increase in Q_2 over Q_1. This corresponds to the widespread opinion that Russell Wilson has great value as a scrambling quarterback. It is probably surprising to many football fans to see that Eli Manning's Q_2 statistic also suggests that he makes valuable runs. We bear in mind that

we have a limited dataset, and in the first six weeks of the 2017 season, Eli Manning ran only three times with a total gain of 21 yards. Generally, the differences between Q_1 and Q_2 are not great, and this demonstrates that passing (decision making and execution) remains the fundamental contribution of quarterbacks.

The calculations obtained in Table 2 were based on the allowance of an additional $\epsilon = 0.5$ seconds from the time of the release of the pass or until the quarterback had passed the line of scrimmage in a non-designed run (see Section 2). Although this is a feature of the methods, we need to be sensitive to the reality that prediction of potential player actions beyond $\epsilon = 0$ seconds becomes increasingly difficult for larger ϵ. We therefore repeated the analyses in Table 2 using $\epsilon = 0$ seconds and found that the sample correlation using the two timeframes was $r = 0.99$ for Q_1 and $r = 0.99$ for Q_2.

We have previously mentioned that although our statistics attempt to investigate quarterback performance, the statistics do not completely eliminate contributions from teammates and the effect of the opposing team. To investigate this, we are interested in the opportunities available to the 29 quarterbacks in our dataset. We measure opportunity for a quarterback by $\bar{Y}$ which is an average of Y_i in (1) taken over the plays available to the quarterback. In our dataset, Philip Rivers had the greatest opportunity $\bar{Y} = 0.473$ whereas Case Keenum had the least opportunity $\bar{Y} = 0.344$. More generally, opportunity appears to be a topic worthy of greater investigation in sports analytics. The blog post by Lopez (2020) considers opportunity in the context of NFL running backs.

5 DISCUSSION

In the NFL, the quarterback is generally regarded as the most important player on a team. The quarterback touches the ball on every offensive possession and his decision making is critical to team success. Yet, the way that quarterbacks are evaluated in the media is not nuanced. Generally, their assessment is determined by box score statistics.

This paper attempts to use the rich potential of spatio-temporal data to evaluate quarterbacks at a deeper level. The player tracking data used in this analysis considers the

locations and velocities of all players on the field in increments of 0.1 seconds. With this wealth of information, we develop interpretable statistics that are based on what a quarterback actually did compared to what they might have done. The statistics use machine learning techniques for the primary purpose of predicting what might have happened had the quarterback chosen a different option. We are not suggesting that our statistics ought to become the standard for quarterback evaluation. Rather, we suggest that they provide a nuanced view involving decision making where quarterbacks on weaker teams are provided a more balanced appraisal.

Although we believe that Table 2 is interesting, we recognize that it is based on only six weeks of available data during the 2017 regular season of the NFL. The main purpose of the paper is to explore the possibilities involving quarterback evaluation. Accordingly, there are both limitations and potential future research directions associated with our work.

One limitation that we do not know how to resolve is that quarterbacks are sometimes limited in their freedom to make decisions. Therefore, it is not genuine that all options evaluated by our statistic Q in (2) are realistic options. It may be the case that coaches provide experienced quarterbacks more leeway in decision making than inexperienced quarterbacks. Therefore, it might be argued that the statistics developed in this paper are also a function of coaching. We also note that we have not provided standard errors associated with the statistics. With larger datasets, this may be remedied by some sort of bootstrapping procedure.

Another limitation of our approach is that we have not completely separated the performance of a quarterback from the performance of his teammates. For example, in the calculation of Q_1 and Q_2, really good receivers who catch a ball that others would not catch helps to inflate Q_1 and Q_2. Perhaps the metrics might better be interpreted as measures of a team's performance in the passing game. We also ought to keep in mind that maximizing EPV gained is not always a quarterback's objective. Particularly in late game situations, we believe that maximizing win probability is a more realistic criterion. One could also imagine situations where maximizing first down probability is the most

important criteria.

For future research, we see various potential enhancements and extensions. First, a greater exploration of ϵ outlined in Section 2 could be investigated. Recall that ϵ is the amount of time that we consider after a pass attempt to assess alternative quarterback options.

We also see the possibility of the consideration of additional statistics. For example, instead of the observed statistics A_i in (2), we could replace it with its expected value $\mathrm{E}(A_i)$ which is explored in Reyers (2020). The idea is that whereas A_i represents the actual EPV gained on the ith play, $\mathrm{E}(A_i)$ is the expected value taken over the population of quarterback throws given the passing decision. In this case, Q provides a greater reflection of decision making rather than a combination of decision making and execution. For illustration, in Table 2, Prescott, Cousins and Winston were ranked the top three quarterbacks, respectively according to Q_1. Using $\mathrm{E}(A_i)$, Reyers (2020) reports that the top three quarterbacks are Cousins, Rivers and Newton, respectively. Alternatively, execution-based metrics could be obtained by considering the differences $A_i - \mathrm{E}(A_i)$ which describe the excess of observed performance over expected performance. Another avenue for future work is the consideration of player specific traits. Currently, for example, the completion probability model is based on the concept of an average receiver. A quarterback's decision making may change depending on the quality of a potential receiver.

6 REFERENCES

Albert, J.A., Glickman, M.E., Swartz, T.B. and Koning, R.H., Editors (2017). *Handbook of Statistical Methods and Analyses in Sports*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton.

Amoros, R. (editor), (2016). "Which professional sports leagues make the most money", *howmuch.net*, retrieved March 18/20 at https://howmuch.net/articles/sports-leagues-by-revenue

Burke, B. (2019). "DeepQB: Deep learning with player tracking to quantify quarterback decision-making and performance", *MIT Sloan Analytics Conference*, retrieved May 18/20 at www.sloansportsconference.com/wp-content/uploads/2019/02/DeepQB.pdf

Chu, D., Reyers, M., Thomson, J. and Wu, L.Y. (2020). "Route identification in the National Football League", *Journal of Quantitative Analysis in Sports*, 16, 121-132.

Deshpande, S. and Evans, K. (2020). "Expected hypothetical completion probability", *Journal of Quantitative Analysis in Sports*, 16, 85-94.

Fernandez, J. and Bornn, L. (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *12th Sloan Sports Analytics Conference*, retrieved May 14, 2020 at www.sloansportsconference.com/wp-content/uploads/2018/03/1003.pdf

Gough, C. (2018). "National Football League (NFL) - Statistics and facts", *Statista*, retrieved March 18/20 at https://www.statista.com/topics/963/national-football-league/

Horowitz, M., Yurko, R. and Ventura, S.M. (2020). "nflscrapR: Compiling the NFL play-by-play API for easy use in R", R package version 1.8.3, https://github.com/maksimhorowitz/nflscrapR.

Lopez, M. (2020). "Don't running backs matter", *StatsbyLopez*, retrieved April 21/21 at https://statsbylopez.netlify.app/post/don-t-running-backs-matter/

Naimi, A. and Balzer, L. (2017). "Stacked generalization: An introduction to super learning", retrieved June 26/20 at http://www.jstatsoft.org/v61/io8/.

Next Gen Stats Team (2018). "Next Gen Stats introduction to completion probability", *NGS Photo Essays*, retrieved May 1/20 at www.nfl.com/news/story/0ap3000000964655/article/nextgen-stats-introduction-to-completion-probability

Reyers, M. (2020). "Quarterback evaluation in the National Football League", *MSc project in the Department of Statistics and Actuarial Science, Simon Fraser University*, retrieved November 15/20 at http://stat.sfu.ca/content/dam/sfu/stat/alumnitheses/2020/Reyers-Matthew-MSc-Project.pdf

van der Laan, M.J., Polley, E.C. and Hubbard, A.E. (2007). "Super Learner", *UC Berkeley Division of Biostatistics Working Paper Series*, Working Paper 222, retrieved June 26/20 at https://biostats.bepress.com/ucbbiostat/paper222.

Yam, D.R. and Lopez, M.J. (2020). "What was lost? A causal estimate of fourth down behavior in the National Football League", *Journal of Sports Analytics*, 5, 153-167.

Yurko, R., Matano, F., Richardson, L.F., Granered, N., Pospisil, T., Pelechrinis, K. and Ventura, S. (2020). "Going deep: models for continuous-time within-play valuation of game outcomes in American football with tracking data", *Journal of Quantitative Analysis in Sports*, 16, 163-182.

Yurko, R., Ventura, S. and Horowitz, M. (2019). "nflWAR: a reproducible method for offensive player evaluation in football", *Journal of Quantitative Analysis in Sports*, 15, 163-183.

Zilavy, G. (2018). "How to calculate NFL passer rating using a formula in Excel or Google Sheets", *Medium Data Science*, retrieved March 18/20 at https://medium.com/@gzil/how-to-calculate-nfl-passer-rating-using-a-formula-in-excel-or-google-sheets-54eb07246d1e

QB	Team	# Plays	Q_1	Q_2	NFL QB Rating
D Prescott	Dallas	140	44.5	43.9	86.6
K Cousins	Washington	154	42.8	43.4	93.8
J Winston	Tampa Bay	118	40.4	40.4	92.2
A Smith	Kansas City	196	39.9	39.3	104.7
M Ryan	Atlanta	164	39.5	39.7	91.4
D Carr	Oakland	117	39.4	39.3	86.4
C Wentz	Philadelphia	205	38.6	38.1	101.9
T Brady	New England	180	38.5	38.0	102.8
J McCown	NY Jets	179	37.8	38.0	94.5
P Rivers	San Diego	214	37.3	37.3	96.0
A Dalton	Cincinnati	135	37.1	36.2	86.6
D Brees	New Orleans	114	36.4	36.4	103.9
B Roethlisberger	Pittsburgh	193	35.7	35.1	93.4
C Keenum	Minnesota	136	35.0	36.0	98.3
E Manning	NY Giants	201	34.5	35.3	80.4
C Newton	Carolina	184	34.0	34.0	80.7
J Goff	LA Rams	170	33.4	33.3	100.5
T Siemian	Denver	162	32.0	31.5	73.3
A Rodgers	Green Bay	164	31.7	31.5	97.2
M Mariota	Tennessee	128	31.6	31.0	79.3
M Stafford	Detroit	182	31.2	31.3	99.3
J Brissett	Indianapolis	166	30.9	31.8	81.7
R Wilson	Seattle	179	28.9	31.0	95.4
T Taylor	Buffalo	163	28.8	28.3	89.2
C Palmer	Arizona	190	28.6	28.5	84.5
B Bortles	Jacksonville	129	28.1	28.8	84.7
J Flacco	Baltimore	146	26.0	26.0	80.4
J Cutler	Miami	126	25.6	26.4	80.8
D Kizer	Cleveland	128	24.5	24.8	60.5

Table 2: NFL Passer Ratings and rankings based on the Q metrics for the first six weeks of the 2017 NFL season.