Statistical Analysis of the Relationship Between NFL Quarterback Performance by Route and Team Winning Percentage

Rocco Trinci, 40129209
STAT 497, Sports Analytics
Dr. Joshua Wyatt Smith
Friday, October 28, 2022

Introduction

While American football[1] continues to grow, it is still not one the most globally known sports. That being said, one aspect that even the most novel sports fans likely know about football is the sport's violent and brutal nature. While this depiction is not without merit, often forgotten is the technical aspect also found within the game. One of the main areas in which this technique can be easily observed is in the paths executed by receivers[2] in order to create separation between themselves and the defenders attempting to cover them. The canonical term for these paths is routes and, while many variations have been developed throughout the sport's history, there exists a group of route types that form the foundation of all passing plays in football (Figure 1). Given their importance to the offensive aspect of the game, it is somewhat surprising that very little analytical research has been attributed to evaluating these different route types. As such, this paper aims to amend this research gap. Namely, this paper identifies which combination of routes best predicts a team's season winning percentage by analyzing the relationship between starting quarterback's total "Expected Points Added"[3] when targeting certain routes and team winning percentage through the use of multiple linear regressions models.

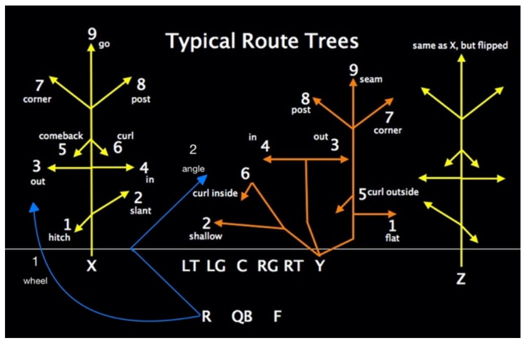

Figure 1: Typical Route Tree for all Eligible Receivers on a given play in American Football

Literature Review

When it comes to research regarding football routes, there have been numerous attempts to create route identification models using various techniques such as model-based curve clustering (Chu et al., 2019) and computer vision techniques (Ajmeri & Shah, 2017). However, these studies do not focus on analyzing the effectiveness of different routes. In terms of evaluating performance of route types using statistical analysis, there has not been a strong effort: Tarnowski & Saffran (2019) attempted to determine which routes "had the highest chance of success when provided the receiver, offensive formation, defensive formation, and the number of defenders in the box." However, the statistical method used to determine these probabilities and the overall results are not readily available. Alternatively, Mark Schofield et al.[4] used a Convolution Neural Network model to recognize which route was ran on any given play. However, the only effort to conduct any analysis was through simple comparisons of each route's EPA per play versus each type of defensive coverage. While insightful to a certain degree, the lack of effort to determine whether these differences are statistically significantly diminishes the overall impact of the study.

Although research into these different route combinations is clearly lacking, one area of football that has received in-depth analysis is Expected Points Added models, and specifically their usefulness in quarterback evaluation. First by Virgil Carter & Robert E. Machol (1971), this idea looks to quantify the value gained or lost by each play in a football game based on the current game context (i.e. location on field, amount of time left, etc.). While variations exist, the values are determined using the "historical mathematical chances a team will score based on the field position and down."[5] Through the use of this type of model, quarterback performance can be quantified by summing up total EPA of plays impacted by the quarterback. Additional research shows that the correlation

[1] American football and football will be used interchangeably
[2] The term receiver refers to any player who is eligible to catch the football being thrown by the quarterback
[3] Type of model with many variations which aims to quantify value of very play in football. Detailed information found in following section.
[4] Mark Schofield, "The Best Route Combination for Every Type of Coverage," USA Today (Gannett Satellite Information Network, August 9, 2021), https://touchdownwire.usatoday.com/lists/nfl-offense-best-route-combinations/.
[5] Joey DiCresce and Tej Seth, "Which QB Stats Are the Most Important?," M-Fans, April 14, 2020, https://mfootballanalytics.com/2020/04/06/which-qb-stats-are-the-most-important/.

between quarterback's performance in these metrics and overall team winning percentage is higher than other common statistics.[6] Specifically, looking through five seasons of quarterback data, it was determined that quarterback's total EPA[7] and quarterback's total EPA on passing plays[8] accounted for 43.7% and 30.7% of the variability in their team's win percentage on the season, respectively[9]. As such, it is through this lens of quarterback performance versus team performance that the gap in route performance analysis will be addressed.

Dataset

The original dataset used for this analysis was the NFL Big Data Bowl 2021 dataset[10], comprised of player tracking, play, game, and player information for all plays during the 2018 regular season. Most importantly, this dataset consists of EPA values for every passing play[11], as well as the route performed by the targeted receiver on almost all passing plays, the two key components needed to conduct this study's analysis.

Once the data was cleaned and manipulated, two tables of data remained: The first dataset contained a list of EPA values for every play that each of the key 11 routes[12] was the main route being targeted throughout the course of the season. Each route contained various amounts of EPA values, with **corner** routes having the lowest count at 421 and **hitch** routes having the highest at 2270. Next, a second table of data was created containing a total of 14 columns: The first column contains the name of each team's starting quarterback for the 2018 season. The next 11 are for each of the routes in question and contain each quarterback's total EPA on the season for each of these routes. Finally, the last two columns consists of the abbreviated team names of each quarterback and the team's season winning ratio (i.e. number of wins divided by total number of games played)[13]. Following this data cleaning process, the following methodologies were applied using these two constructed datasets.

Methodology

Before tackling the main question at hand, initial tests were conducted to validate the idea of differentiating routes based on their EPA production. Namely, a one-way ANOVA test was conducted to determine if a statistically significant difference existed between the sample mean EPA produced per attempt for each of the 11 routes. In addition, a multiple comparison post hoc test using Bonferroni Procedure was subsequently conducted to determine the specific pairwise significant differences. These tests were selected as they are the typical methods used to compare sample means from a number of different populations, which, in this study, are the route types. It must be noted that, by definition, one-way ANOVA is reliable if the residuals of the samples are normally distributed, which is technically not the case based on results from conducting a Shapiro-Wilks test for each of the sample sets[14]. However, given that the samples being used are large (all greater than 400 data points) and that Shapiro-Wilks test is sensitive to outliers in samples, normality was assumed. Once these tests were performed, a number of multiple linear regression models were constructed to address the thesis of this paper.

More specifically, using the second constructed dataset, several multiple linear regression were performed. While they all included the same dependant variable, starting quarterback's team win percentage for the season, the independent variables used varied between the models: A total of 11 independent variables were considered, each one corresponding to the quarterback's total EPA on one of the 11 route types being analyzed[15]. Accordingly, three different models were created using different combinations of these independent variables based on different statistical reasoning, with the goal of improving the variability in the dependant variable that is accounted for by the combination of independent variables, as well as the overall significance of the model and its independent variables.

Results

The one-way ANOVA provided promising results, revealing that there existed a statistically significant difference in the EPA produced per attempt between least two of the different routes being considered (p-value < 0.001). Moreover, Bonferroni's Procedure for multiple comparisons illustrated that, while not all routes had significantly different average EPA per attempt, roughly half of the possible route pairs exhibit a significant

[6] Joey DiCresce & Tej Seth
[7] Quarterback's total EPA is calculated by summing the EPA values of all plays in which they are involved in
[8] Quarterback's total EPA on passing plays is calculated by summing the EPA values of all plays where a pass was attempted or was planned to be attempted by said quarterback
[9] Joey DiCresce & Tej Seth
[10] "NFL Big Data Bowl 2021," Kaggle, accessed October 28, 2022, https://www.kaggle.com/competitions/nfl-big-data-bowl-2021/data?select=week1.csv.
[11] Model used to calculate these EPA values is not mentioned in dataset description
[12] The 11 routes that this study focuses on are: angle, corner, cross, flat, go, hitch, in, out, post, screen and slant
[13] Win ratio, win(nning) percentage and Win% will be used interchangeably
[14] Results of Shapiro-Wilks test found in provided Python code
[15] Names of variables are as follows: ANGLE, CORNER, CROSS, FLAT, GO, HITCH, IN, OUT, POST, SCREEN, SLANT where ANGLE = *Quarterback's total EPA on **angle** routes*, CROSS = *Quarterback's total EPA on **corner** routes,* and so on

difference in their sample means (Figure 2). These results validated the idea of comparing routes in terms of EPA, in turn allowing for the following tests to take place.

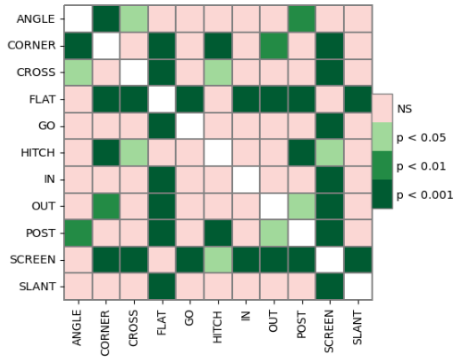

Figure 2: Heatmap demonstrating results of Bonferroni Procedure on samples of the 11 key routes

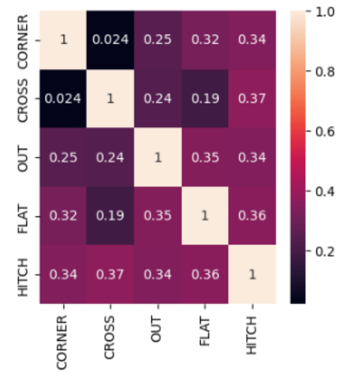

Figure 3: Correlation Coefficients between Total EPA on Corner, Cross, Out, Flat and Hitch routes

The first multiple regression model conducted considered each starting quarterback's total EPA on all 11 routes as independent variables. This model had an adjusted R^2 value of 0.382 and a p-value less than 0.05. However, none of the 11 independent variables had significant p-values in terms of their correlation to the dependant variable of team's win percentage. This lack of significance among the independent variables indicated that this model was not ideal.

To rectify this issue, a correlation matrix was produced to evaluate which of these 11 variables had the highest correlation to Win% on an individual level (Figure 4). The results of this matrix indicates that the total EPA produced on **corner, cross, flat, hitch** and **out routes** have the highest individual correlation. As such, using this information, a subsequent multiple linear regression model was created between these five independent variables and Win%.

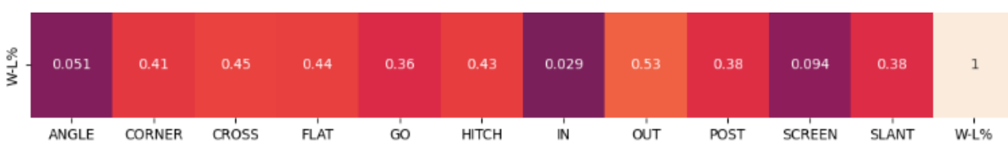

Figure 4: Correlation Coefficients between Win% and Total EPA on each different key route

This second model accounted for a larger portion of the variance in Win% with an adjusted R^2 value of 0.423 and was once again significant with a p-value less than 0.01. Additionally, while the previous model had no significant independent variables, both **CROSS** and **OUT** variables were determined to have statistically significant correlation to the dependant variables with p-values less than 0.05. However, the variables **CORNER**, and to a larger degree, **FLAT** and **HITCH** did not exhibit this correlation. As a result, while this model was an improvement over the initial model, issues still remained.

A final effort was conducted in order to determine if a better model could be attained: Another correlation matrix was created, this time comparing the correlation between the remaining independent variables of **CORNER**, **CROSS**, **FLAT**, **HITCH** and **OUT** (Figure 3). Through this matrix, the potential presence of multicollinearity became evident: The variables **FLAT** and **HITCH** exhibit a moderate correlation with the three other independent variables (**CORNER**, **CROSS**, and **OUT**), having a correlation coefficient greater than 0.32 with each of the three[16]. On the other hand, the correlation between **CORNER**, **CROSS**, and **OUT** variables is no greater than 0.25, with **CROSS** and **CORNER** exhibiting an extremely small correlation of 0.024. As such, given these values, it was determined that the variables of Flat and Hitch exhibit multicollinearity with the other variables, providing sufficient evidence to drop them in the final multiple regression model.

Therefore, the final model was incepted to compare the relationship between the three remaining independent variables, **CORNER**, **CROSS**, and **OUT**, and Win%. The outcome of this model was the most promising, having the highest adjusted r-squared value at 0.431 and being the most significant with a p-value less than 0.0005. Moreover, unlike the previous two models, all three of the independent variables had individual p-values less than 0.05, demonstrating significant correlation to Win%.

[16] Except between FLAT and CROSS. However, given correlation between HITCH and FLAT/CROSS are among the highest, can assume medium correlation exists between FLAT and CROSS.

Discussion

As mentioned above, the third model can be considered the most significant and effective. Given the results of these three models (Figure 5), it can be determined that, accounting for the total expected points a starting quarterback adds when targeting a given route, the routes that best describe the variability of their team's winning percentage in a significant manner are **corner**, **cross** and **out routes**. This implies that the expected points gained from these three routes predicts win percentage better than the expected points added on any other combination of route types, providing a solution to this paper's research question.

Moving along, inspecting the physical features of the three routes selected (Figure 1), it can be seen that this trio accounts for all varieties of routes. It can be deduced from this fact that the results of this study do not indicate to favour any type of route, for example favouring deeper routes over more shallow routes. While this may seem somewhat disappointing, this fact provides circumstantial evidence that the methodology used is not biased towards any specific type of route. Additionally, this variation among the selected routes allows for potential actionable decisions for similar type routes. For example, given their similarities in terms of design but not in impact on win percentage, a decision could be made to implement more crossing routes instead of slant routes or corner routes instead of go routes.

Model Number	Ind. Var.	Dep. Var.	Adj. R^2	P-value (F-stat)	Significant Ind. Var.	Beta values (in order listed)
1	ANGLE, CORNER, CROSS, FLAT, GO, HITCH, IN, OUT, POST, SCREEN, SLANT **(i.e. All Routes)**	W-L%	0.382	< 0.05	None	0.0046, 0.0050, 0.0034, 0.0024, 0.0004, 0.0003, -0.0022, 0.0045, 0.0063, 0.0041, -0.0028
2	CORNER, CROSS, FLAT, HITCH, OUT	W-L%	0.423	< 0.005	CROSS, OUT	0.0060, 0.0036, 0.0026, 0.0009, 0.0039
3	CORNER, CROSS, OUT	W-L%	0.431	< 0.0005	CORNER, CROSS, OUT	0.0074, 0.0041, 0.0045

Figure 5: Summary of Key Values from Three Multiple Regression Models

Limitations and Future Research

One of the main types of limitations of this study pertains to the data being used: While datasets containing event level EPA values for multiple seasons were available, no other readily available dataset contained route data for other seasons, limiting this study to the 2018 NFL season only. This leaves open the possibility that the information obtained is skewed given how that there tends to be statistical differences between seasons. Additionally, there is roughly 2000 data points missing in the dataset when comparing the 17671 passes attempted in the 2018 season versus the 15160 passing events for which route data is available in the dataset being used. This missing information could also be the cause of inexact results. Intuitively, the solution to these two limitations would be to obtain more various and accurate data, which is generally easier said than done. However, using available tracking data in junction with previously mentioned route recognition tools developed by other researchers provides a clear path into how this supplementary data could be obtained.

In terms future research, analyzing the relationship between EPA produced by route and a team's total points scored on a season could provide more context into the impact of certain routes given that EPA is a model that aims to predict the scoring of points. Alternatively, given that routes are connected to receivers, conducting analysis on how different receivers perform on different routes is another way in which the research into the differences between route types can be expanded.

Conclusion

The goal of this study was to determine which collection of routes best explained the variation in team Win% through the lens of the EPA produced by starting quarterbacks on each route type. Through the development of several multiple linear models, it was determined that the routes that fulfill this goal are **corner**, **crossing** and **out routes**. Specifically, these routes explain 43% of the variability in win percentage, which is on par with previous studies. Finally, the diversity of these routes demonstrates a lack of favouritism to any certain type of route, which in turn provides the possibility of altering route selection decisions. That being said, while this study provides solid foundation for the statistical analysis of football routes, limitations caused by the lack of data may have affected the final results obtained. These issues could be resolved through the use of route recognition models that continue to be incepted by other researchers. These tools would augment the currently available data, in turn opening the door for more expansive route analysis as it pertains to their relationship to other team performance statistics or to the performance of pass catchers.

Works Cited

Chu, Dani, Matthew Reyers, James Thomson, and Lucas Wu. "Route Identification in the National Football League." arXiv.org, August 7, 2019. https://arxiv.org/abs/1908.02423.

Ajmeri, Omar, and Ali Shah. "Using Computer Vision and Machine Learning to Automatically Classify NFL Game Film and Develop a Player Tracking System." MIT Sloan Sports Analytics Conference, March 3, 2017. https://www.sloansportsconference.com/research-papers/using-computer-vision-and-machine-learning-to-automatically-classify-nfl-game-film-and-develop-a-player-tracking-system.

Tarnowski, Jacob, and Kaili Saffran. "Optimizing NFL Receiver Route Combinations." Fisher Digital Publications, April 27, 2019. https://fisherpub.sjf.edu/undergraduate_ext_pub/28/.

Carter, Virgil, and Robert E. Machol. "Technical Note—Operations Research on Football." *Operations Research* 19, no. 2 (1971): 541–44. https://doi.org/10.1287/opre.19.2.541.