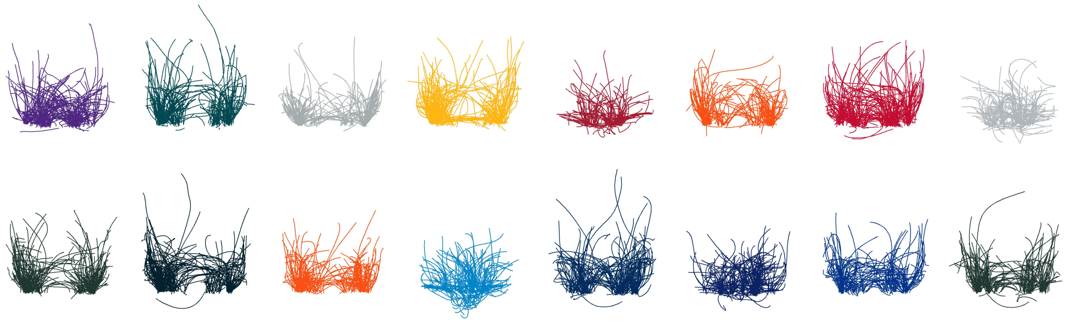
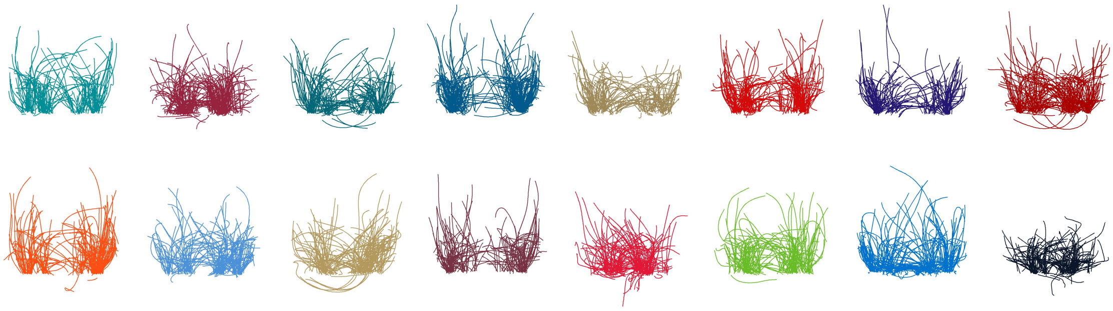


NFL Big Data Bowl

Dani Chu
Lucas Wu

Matthew Reyers
James Thomson



1 Introduction

The execution of pass plays in the National Football League (NFL) is crucial to offensive success, especially during a new era of football with an increased emphasis on passing. This year, LA Rams' coach Sean McVay was heralded for his offensive schemes (Simmons, 2018). Teams are continuously trying to improve their offense and due to McVay's success are trying to find a similar head coach (Willis, 2019). Improving their offense, however, may be easier than hiring anyone that has ever made eye contact with McVay as tracking data can be effectively leveraged to help design better offenses.

Next-Gen Stats' player tracking data from all NFL games is available to all 32 NFL teams for the first time this season. The NBA has had similar data available since 2013 (NBA, 2013). This has encouraged more public literature about the use of tracking data to improve strategy in sports. For example, Miller and Bornn (2017) used a probabilistic clustering algorithm for Functional Data to identify common actions in NBA possessions. This same strategy can be applied to the similar world of routes on passing plays in the NFL. It could also be expanded to recognize defensive formations, blocking schemes and run plays.

Though access to tracking data in football has been limited until the NFL Big Data Bowl, there has been publicly available play by play data via the nflscrapR package (Horowitz et al., 2018). Using this data Yurko, Ventura, and Horowitz (Yurko et al., 2018) proposed a Expected Points Added (EPA) model, Win Probability Added model and Wins Above Replacement model. Use of these metrics and data sets have breathed life into NFL analytics, building a firm foundation with which we can now expand.

We will also build upon the foundation of adjusted plus minus models using regression techniques that have been used in basketball (Rosenbaum, 2004), hockey (Macdonald, 2011) and soccer (Matano et al., 2018). Specifically in (Macdonald, 2011) and (Sill, 2010), the authors used ridge regression to address the co-linearity in the data. The idea of adjusted plus minus models is to isolate the marginal effect of players on a given performance metric, accounting for their competition and teammates. We propose using a similar framework to evaluate the effectiveness of route combinations while accounting for defensive personnel and the other routes being run on the field.

We also explore the concept of "openness" of all receiving players at the time the ball is thrown and how that openness can be interpreted as success not just by the targeted player, but by the others who created the space for them by taking defenders away from the ball. We visualize the control on the field each team has and how that changes from the ball being thrown to the ball being caught.

In this report we propose the following contributions:

1. A model based clustering approach for functional data to recognize NFL routes
2. A selection of optimal route combinations for different situations
3. A method for measurement of openness to analyze the effectiveness of route combinations in creating control over the field

2 Pattern Recognition: Identifying Routes

The tracking data released for this competition covers all possessions from the 91 games that took place in the first 6 weeks of the 2017 NFL regular season. To identify routes run on passing plays we use the following steps to transformed the data.

First, we identified all the passing plays and kept only the tracking data for the Wide Receivers (WR), Running Backs (RB), Full Backs (FB), and Tight Ends (TE). All routes were analyzed with a common direction of play, line of scrimmage, and starting horizontal origin. We removed pre-snap motion, post-play movement (defined in apendix). We followed the work of (Miller and Bornn, 2017) and used Bezier curves (Olsen, 2018) to smooth route curves. The curves require double the observations to control points (frames) in a play, so all plays of 2 seconds or less were removed from the tracking data.

We then sample 200 evenly spaced x, y points from the bezier curves and clustered the points using multivariate model-based clustering of functional data (Bouveyron and Jacques, 2011) implemented in the funHDDC R package (Schmutz and Bouveyron, 2018) in conjunction with the fda R package (Ramsay et al., 2018).

We ran the clustering process for cluster sizes of 10, 15, 25 and 50 clusters. We chose to use the results from the clustering process with 50 clusters; this is because when the process was asked for more clusters it distinguished more horizontal variation in the route patterns instead of only clustering on depth of target.

While each cluster is distinct, we saw distinct routes in the data set and an opportunity to manually combine similar clusters into groups. In Figure 1 we plot the medians of each cluster for the left and right side of the field for 9 out of the 10 route groups.

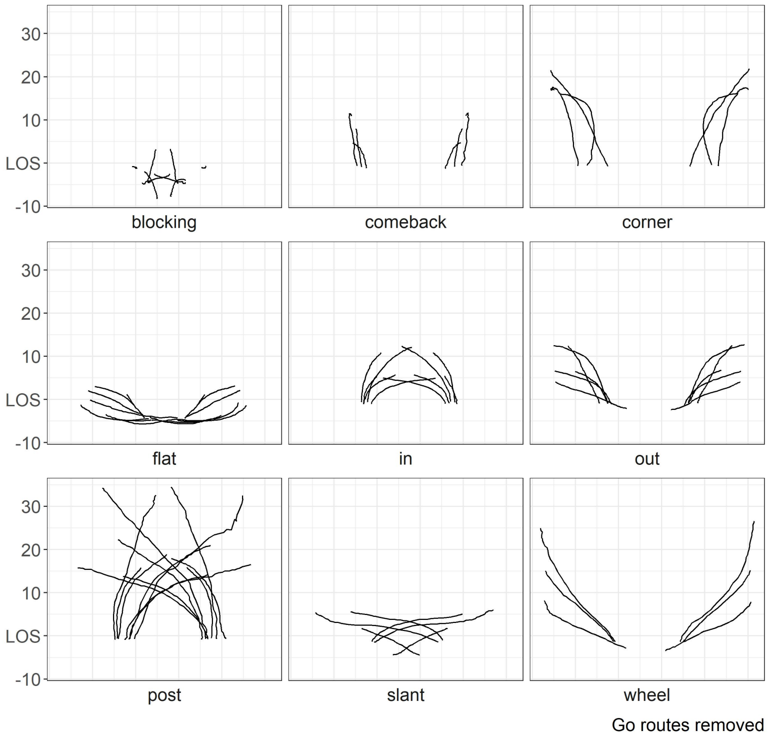


Figure 1: Median of clusters. Go routes not plotted for clarity.

These route groups give us objects with which we can use to describe plays in football terms.

2.1 Player Route Tendencies

One example is that we can use the routes that we identified to see the usage of different teammates within the same offense. In Figure 2, we compare Vikings teammates Adam Thielen and Stefon Diggs. Diggs runs a greater proportion of go routes than Thielen. When Thielen does go long he prefers post routes.

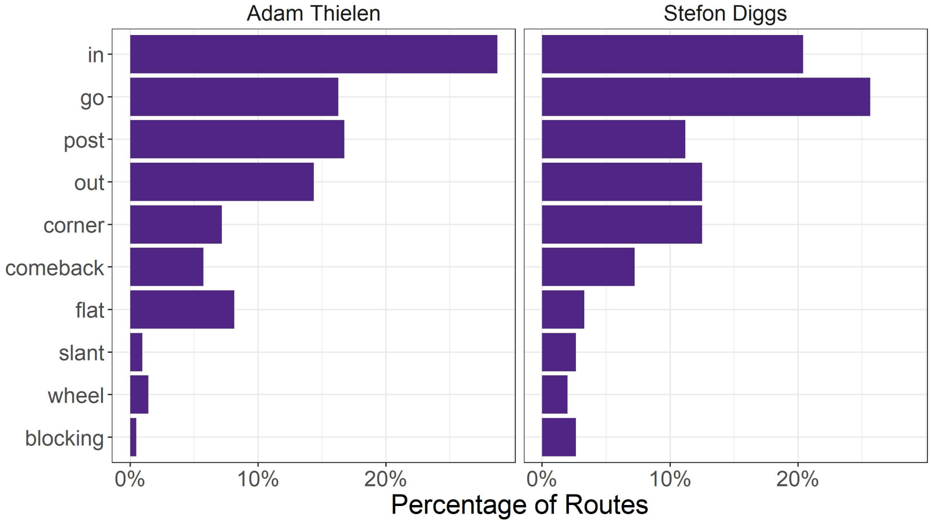


Figure 2: Comparison of Minnesota Vikings Wide Receivers

These type of route profiles can be used to cluster receivers based on their route habits. Without clustering, we can use the natural position groupings to see the distinctive route profiles of different players. In Figure 3, we can see that the route profiles for the three players lines up with our intuition of what their behaviour should be.

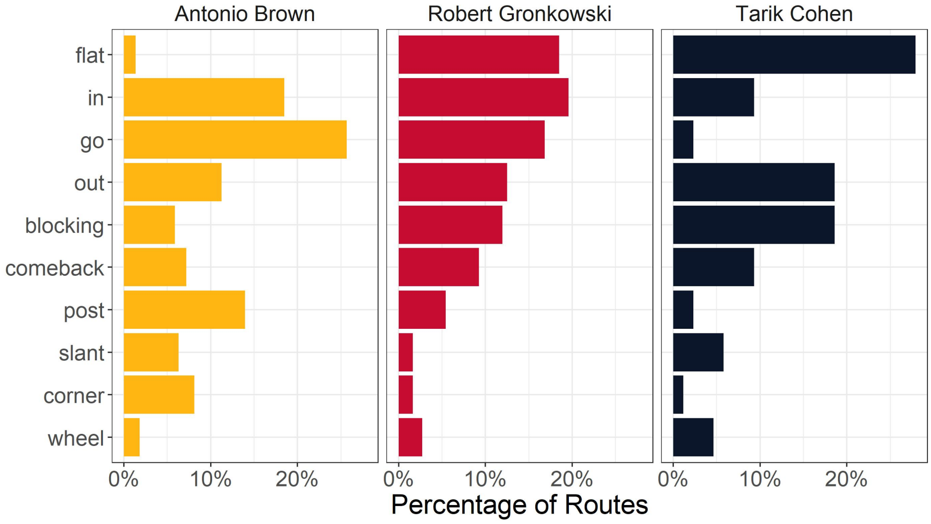


Figure 3: Comparison of routes by players of different positions

While these route profiles can be useful for scouting and describing offense tendencies, they also give us confidence that we have clustered and labelled routes in an appropriate way.

3 Evaluating Play Designs

3.1 Expected Points Added Metrics

One way plays can be evaluated is through performance based metrics. Yurko et al. (2018) implement an Expected Points Added (EPA) model and a Win Probability Added (WPA) model in their nflscrapR package (Horowitz et al., 2018) that can be easily joined with our data set. EPA is a good metric to evaluate plays, it's a measure of whether a given play increased or decreased a teams chances of scoring. The other metric we can use is a discretization of EPA to success rate. While there is debate amongst the NFL analytics community about the cut off to use for success rate, we define success rate to be any play in which a positive EPA is observed according to the Yurko et al. (2018) model.

The issue with using EPA based metrics to evaluate routes is that a play's design can be good when EPA is not and vice versa. For example, a receiver can be open due to play design and the receiver could drop the ball, not be targeted, fumble the ball, etc. These could all have negative EPA but does not mean that the play design was bad.

We can evaluate whether a player performed well by visualizing their zone of control. This looks at the space created on the field as a result of their actions, not the results of a complete or incomplete pass.

3.2 Zone of Control and Openness

Instead of looking at the outcome, we can use a process based method of evaluating plays. Evaluating a play in this data set comes with a few challenges. The first is understanding if what was done on the play was the best decision given the play design, coverage patterns, and openness of players. We first investigated the idea of Voronoi diagram which partitions the space by assigning every location to the nearest player, however, it does not capture the current velocities of players Bornn et al. (2018). We modelled team field control based on the work of Fernandez and Bornn (2018) quantified the control that the offense and defense held over contested areas of the field during passing plays. The output of the player ”openness” analysis allow for a visual understanding of field control and how that changes from when the ball is thrown to when it is caught.

The zone of control is measured using a bivariate normal distribution which takes into account the velocity and direction of all players on the field at an instance in time. Non-targeted receivers contributions to creating space for the target receiver are easily visualized. This could one day be used as a tool for analyzing an offense for better understanding of route combinations with mapped optimal catching zones. We can also use this to measure the result of a play based on individual performance, not only by whether yards/a touchdown were gained.

We can use EPA to measure the success of a play, though with only one realization of the play it becomes challenging to demonstrate another option would have been superior. The motivation for understanding the range of outcomes of a play stems from the desire to understand how well a combination of routes perform. Assuming what occurs on the field to be the optimal realization of a set of routes is faulty due to a collection of human errors.

4 Modelling Route Combinations

To determine complimentary route combinations we build two ridge regression models similar to (Macdonald, 2011) and (Sill, 2010).

We let

$$\begin{aligned}
X_{i,j} &= \begin{cases} 1, & \text{if a route j is observed play i;} \\ 0, & \text{if a route j is not observed on play i;} \end{cases} \\
R_i &= \text{\# of Pass Rushers on play i;} \\
B_i &= \text{\# Defenders in the Box on play i;} \\
DB_i &= \text{\# of Defensive Backs (DB) on the field on play i;} \\
LB_i &= \text{\# of Line Backers (LB) on the field on play i;} \\
DL_i &= \text{\# of Defensive Linemen (DL) on the field on play i;}
\end{aligned}$$

The models for EPA and Success are,

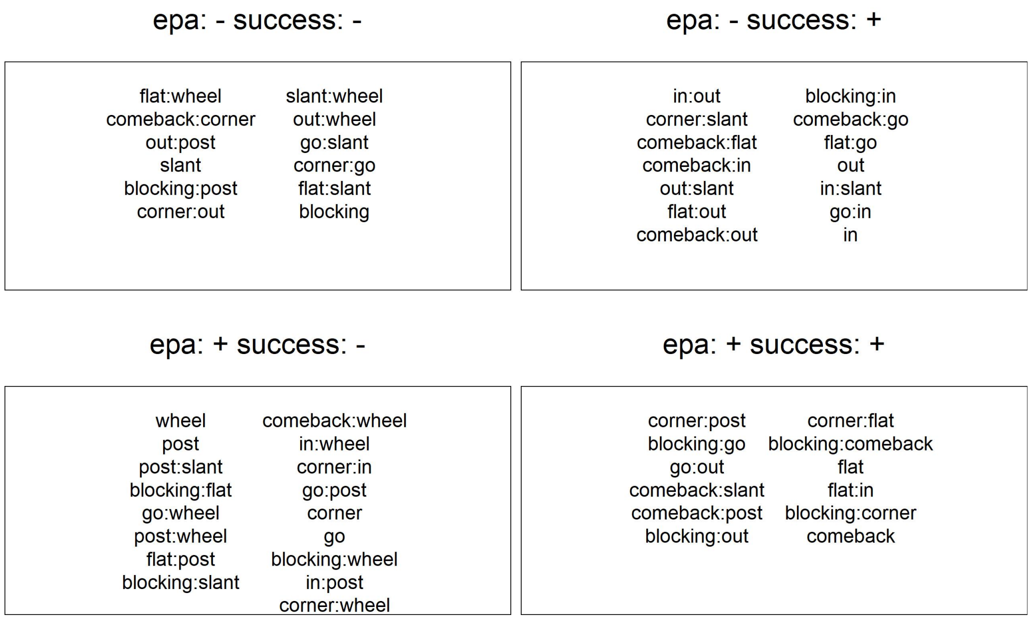
$$\begin{aligned}
\text{EPA} =& \beta_0 + \beta_1 X_1 + \dots \beta_J X_J + \beta_{1,2} X_1 X_2 + \dots \beta_{1,J} X_1 X_J + \dots \beta_{J-1,J} X_{J-1} X_J + \\
& \beta_R R + \beta_B B + \beta_{DN} DB + \beta_{LB} LB + \beta_{DL} DL \\
\text{Success} =& \beta_0 + \beta_1 X_1 + \dots \beta_J X_J + \beta_{1,2} X_1 X_2 + \dots \beta_{1,J} X_1 X_J + \dots \beta_{J-1,J} X_{J-1} X_J + \\
& \beta_R R + \beta_B B + \beta_{DN} DB + \beta_{LB} LB + \beta_{DL} DL
\end{aligned}$$

To fit both models we use ridge regression to account for the multicolinearity between the regressors. For the Success model we use a logistic ridge regression model. Both these models were fit using the glmnet package (Friedman et al., 2018).

In the player model case (Macdonald, 2011) we can interpret the player coefficients as the marginal change in performance for a given player accounting for their teammates and level of competition compared to an average player. In our case we can interpret the fitted coefficients in the same way, the coefficients for route and each possible route interaction, accounting for the other routes that were run, and defensive packages. We do not adjust for quarterback or quality of receivers.

Before interpreting the route coefficients, we will interpret the defensive results of the model. More defenders in the box, more DLs, less DBs, less LBs and less pass rushers result in higher EPA plays and higher success rates.

The two coefficients of the two models are summarized based on the coefficient value in 4.


epa: - success: -

flat:wheel
comeback:corner
out:post
slant
blocking:post
corner:out
slant:wheel
out:wheel
go:slant
corner:go
flat:slant
blocking

epa: - success: +

in:out
corner:slant
comeback:flat
comeback:in
out:slant
flat:out
comeback:out
blocking:in
comeback:go
flat:go
out
in:slant
go:in
in

epa: + success: -

wheel
post
post:slant
blocking:flat
go:wheel
post:wheel
flat:post
blocking:slant
comeback:wheel
in:wheel
corner:in
go:post
corner
go
blocking:wheel
in:post
corner:wheel

epa: + success: +

corner:post
blocking:go
go:out
comeback:slant
comeback:post
blocking:out
corner:flat
blocking:comeback
flat
flat:in
blocking:corner
comeback

Figure 4: Route Combination Groupings by Coefficient for EPA and Success Rate

Negative EPA values represent a value less than the average EPA gain in a given play. These plays which have a negative value but were successful, did not do much for the team under the given circumstances. Routes that have negative coefficients in the EPA regression and a positive coefficient for the success regression go in the

"EPA: - success: +" bucket. We see those route combinations as more consistent options that have less big play potential. The "EPA: + success: -" group is seen as the big plays that are less consistent. Finally, the "EPA: + success: +" group are the optimal combinations and the "EPA: - success: -" are not. While we have grouped these routes into buckets most of the coefficients are quite small.

We summarize the plays that had routes from each group in Table 1. We look at the Average EPA, Success Rate Percentage and Completion Percentage for each group. We can see while the uncertainty overlaps the averages to line up with our understanding of the clusters.

Coefficient Group	Mean EPA	SD of EPA	Success %	Completion %
epa: - success: -	-0.06	1.71	39.71%	51.1%
epa: - success: +	-0.03	1.63	45.42%	62.1%
epa: + success: -	0.05	1.73	40.25%	50.5%
epa: + success: +	0.01	1.61	45.19%	61.8%

Table 1: Summary of Play Performance

4.1 Complimentary Route Combinations

The out and go combination is one that is identified as better than average in both models. We pulled a play from the Oakland Raider at Denver Broncos game on October 1st 2017. The out route (#88) complements the go route (#16) by pulling away the defensive back (#21) from the go route, giving the go route outside leverage on his safety (#26).

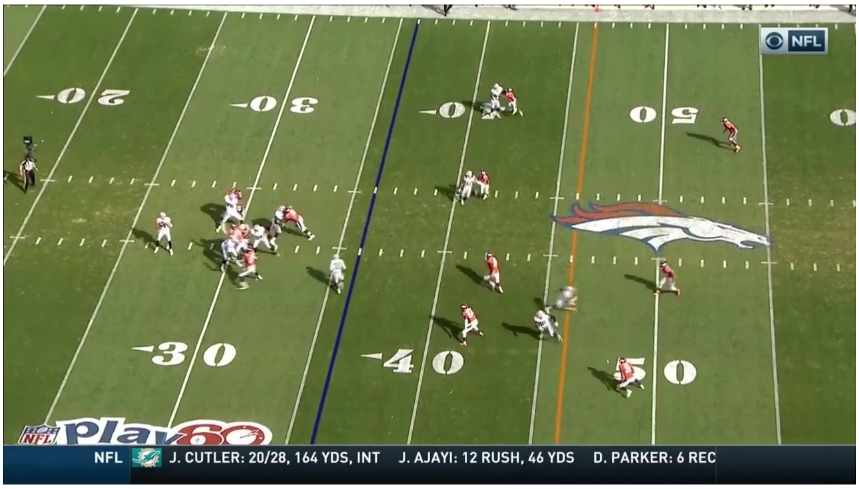


Figure 5: Example of an Out and Go. Video will play in Adobe Acrobat or on YouTube at 2:49.

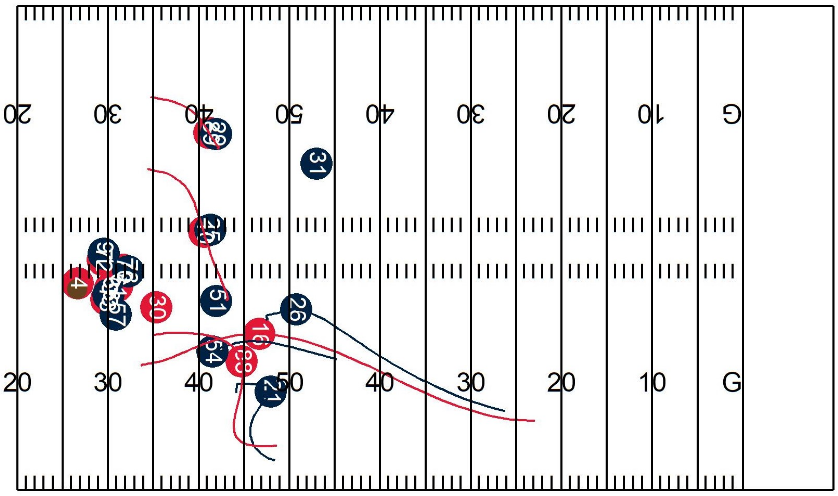


Figure 6: Example of an Out and Go with data. Made with gganimate (Pedersen and Robinson, 2019). Video will play in Adobe Acrobat.

4.2 Mean Control of Field for Route Combinations

Visualizing the control on the field at the time the ball is released by the QB, we can see that the offense (red) has created space for the received to move up the field with the ball. Through proper route combination design, the control of the field is 69% in favour of the offense. In the future we would be interested in whether there is a relationship between zone control and EPA/Success. The target receiver is located around (

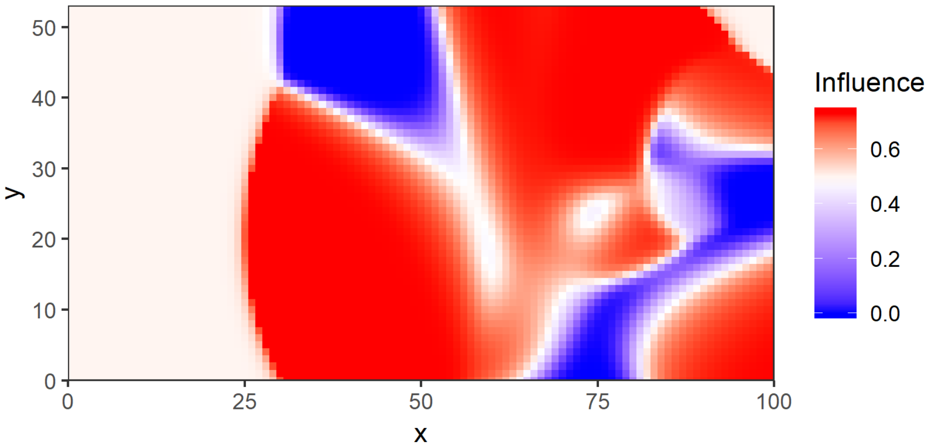


Figure 7: Offensive influence at the moment the ball is released by the QB to the receiver

5 Future Work

In the future, we hope to use this for better evaluation of receivers who run their route, but are not targeted by the QB. A better understanding of the route combinations which create the space necessary for routes to have both a positive success and EPA rate could uncover new combinations for offenses to exploit. The side of the field effect is an area of research we are interested in exploring further, as well as identifying play action plays, and defensive coverage (zone vs man). Finally, we could evaluate which receivers are the best at creating space and may be undervalued or perhaps quarter backs who make great decisions.

We also recognize that the tracking data only covered the first 6 weeks, and that with more data over long periods of time we may have found different conclusions on the effectiveness of the specific route combinations we identified in this paper.

6 Conclusion

Based on the available data we clustered routes and analyzed the combinations based on their success rates and whether they improved the EPA of a team. We identified a number of route combinations that we consider consistently strong in both of categories. We have the ability to visual a play based on the direction and velocity of players to determine which zones of the field are under their control, we extended that to team control to visualize the effect of route combinations on opening up the field for the target receiver.

7 Acknowledgements

We would like to thank Barinder Thind and Cherlene Lin for their helpful discussions around functional data analysis, Michael Couture for his help with defensive formations and Tim Swartz for his guidance and supervision. To the developers of the tidyverse family of packages (Wickham, 2017) without which we would not have been able to perform this research. Finally, to Michael Lopez and the NFL for hosting this competition and givins us this opportunity.

8 Appendix

Set of Route Ending Events = {pass outcome caught, pass outcome incomplete, qb sack, run, touchdown, pass outcome interception, pass outcome touchdown, fumble, qb strip sack, pass shovel, handoff, qb spike}

References

Luke Bornn, Dan Cervone, and Javier Fernandez. Soccer analytics: Unravelling the complexity of the beautiful game. *Significance*, 15(3):26–29, 2018. doi: 10.1111/j.1740-9713.2018.01146.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2018.01146.x.

Charles Bouveyron and Julien Jacques. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300, Dec 2011. ISSN 1862-5355. doi: 10.1007/s11634-011-0095-6. URL https://doi.org/10.1007/s11634-011-0095-6.

Javier Fernandez and Luke Bornn. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *Proceedings of the 2018 MIT Sloan Sports Analytics Conference*, 2018.

Jerome Friedman, Trevor Hastie, Rob Tibshirani, Noah Simon, Balasubramanian Narasimhan, and Junyang Qian. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, 2018. URL https://CRAN.R-project.org/package=glmnet. R package version 2.0-16.

Maksim Horowitz, Ron Yurko, and Samuel Ventura. *nflscrapR: Compiling the NFL Play-by-Play API for easy use in R*, 2018. URL https://github.com/maksimhorowitz/nflscrapR. R package version 1.8.1.

Brian Macdonald. Adjusted Plus-Minus for NHL Players using Ridge Regression with Goals, Shots, Fenwick, and Corsi. *arXiv e-prints*, art. arXiv:1201.0317, December 2011.

Francesca Matano, Lee F. Richardson, Taylor Pospisil, Collin Eubanks, and Jining Qin. Augmenting Adjusted Plus-Minus in Soccer with FIFA Ratings. *arXiv e-prints*, art. arXiv:1810.08032, October 2018.

Andrew C. Miller and Luke Bornn. Possession sketches : Mapping nba strategies. In *Proceedings of the 2017 MIT Sloan Sports Analytics Conference*, 2017.

NBA. Nba partners with stats llc for tracking technology, September 2013. URL http://www.nba.com/2013/news/09/05/nba-stats-llc-player-tracking-technology/. [Online; posted Sep 5, 2013].

Aaron Olsen. *bezier: Toolkit for Bezier Curves and Splines*, 2018. URL https://CRAN.R-project.org/package=bezier. R package version 1.1.2.

Thomas Lin Pedersen and David Robinson. *gganimate: A Grammar of Animated Graphics*, 2019. URL https://CRAN.R-project.org/package=gganimate. R package version 1.0.0.

J. O. Ramsay, Hadley Wickham, Spencer Graves, and Giles Hooker. *fda: Functional Data Analysis*, 2018. URL https://CRAN.R-project.org/package=fda. R package version 2.4.8.

Dan T. Rosenbaum. Measuring how nba players help their teams win, April 2004. URL http://www.82games.com/comm30.htm. [Online; posted April 30, 2004].

A Schmutz and J. Jacques & C. Bouveyron. *funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces*, 2018. URL https://CRAN.R-project.org/package=funHDDC. R package version 2.2.0.

Joseph Sill. Improved nba adjusted +/- using regularization and out-ofsample testing. In *Proceedings of the 2010 MIT Sloan Sports Analytics Conference*, 2010.

Myles Simmons. Mcvay, offensive staff excelling in play design to get best of matchups, October 2018. URL https://www.therams.com/news/mcvay-offensive-staff-excelling-in-play-design-to-get-best-of-matchups. [Online; posted Sep 5, 2013].

Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. URL https://CRAN.R-project.org/package=tidyverse. R package version 1.2.1.

George Willis. The sean mcvay coaching tree is already starting to grow, January 2019. URL https://nypost.com/2019/01/12/the-sean-mcvay-coaching-tree-is-already-starting-to-grow/. [Online; posted Jan 12, 2019].

Ronald Yurko, Samuel Ventura, and Maksim Horowitz. nflWAR: A Reproducible Method for Offensive Player Evaluation in Football. *arXiv e-prints*, art. arXiv:1802.00998, February 2018.