NFL Big Data Bowl

Mitch Kinney

January 22, 2019

Introduction

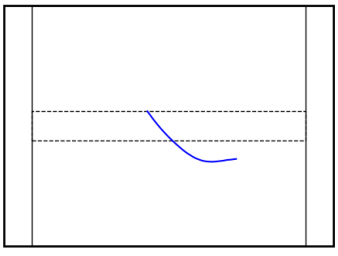
People love statistics in sports. Being able to quantify how great a team has been is alluring because of the power numbers and math have in our society. Summary statistics in football has become huge as analysts are able to define how well a player has done in a game or a season. Pro Football Focus and FiveThirtyEight has dominated this field and are giving the football community ways to relate how good a player has been compared to others in that position. The goal now is how to introduce statistics to the decision making process when actually preparing for and playing in a game. At the level of the pros it seems like many organizations still rely heavily and possibly exclusively on the decisions of humans and gut feelings. After the success of data analytics in baseball, it is clear that sports can be improved by utilizing a bit more math. In this report I will detail my efforts to answer Theme 3: Identify Best Receiver Route Combinations My overall goal is to see if there were routes that when run in combination forced certain spatial characteristics of the intended receiver and nearby defenders that corresponded with a successful play. The majority of this project was wrangling the data into a suitable form to analyze. Much of the data I wanted to extract involved the potential receivers and nearby defenders but there were other interesting things such as if the QB was in the pocket when he released the ball. My end analysis relied heavily on tabulating overall successful plays and using a logistic regression to find how routes ran on successful plays contributed to characteristics such as creating distance from defenders and turning defenders the wrong direction. As an avid watcher of football it was a lot of fun to explore the data and develop an original way to look at the game!

Wrangling the Data

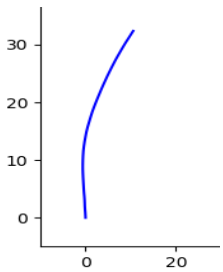
My theme was to diagnose route combinations that receivers ran successfully. Therefore the most important aspect would be to correctly label routes in passing plays. Another important feature was the defenders who were nearby the intended target. These two parts utilized the tracking data most extensively but I also gathered auxiliary information about the play such as if there was a blitz, the number of defenders pressuring the quarterback, the nearest sideline etc.

Labeling Routes

Only given tracking information about each player I decided to use a nearest neighbor type approach to label routes. Steps to do this included extracting the x, y coordinates for each receiver, transforming the coordinates to treat the left and right side of the field as symmetric, resizing the route and finding the closest match to a routes I pre-defined. To extract the x, y coordinates I looked for running backs, fullbacks, tight ends and wide receivers in all plays where a passing result existed. These are all incomplete passes, complete passes, interceptions, sacks and touchdowns. Then I found the coordinates that were in between the snap and approximately 4 seconds afterwards or when the pass reached its target. I shortened some route gathering to exclude excess running from broken plays. Based on the y coordinate of the receivers position at the snap and the direction the ball went after the snap I could determine where the receiver was on the field and which direction they would be running. That information allowed me to transform the receivers route coordinates to get similar starting points and directions. I decided my base would be routes running up and the middle of the field to the right. See figure below I wanted



(a) Route run from left side of the ball into the middle



(b) Transformed route

routes run from the left and right side of the ball to look the same so, for instance, a route running towards the middle of the field looked identical. After transforming I needed to compare the route to a

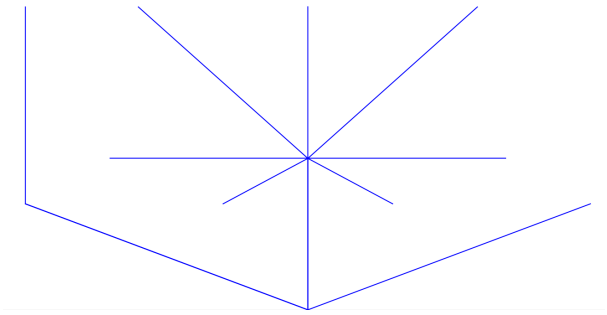
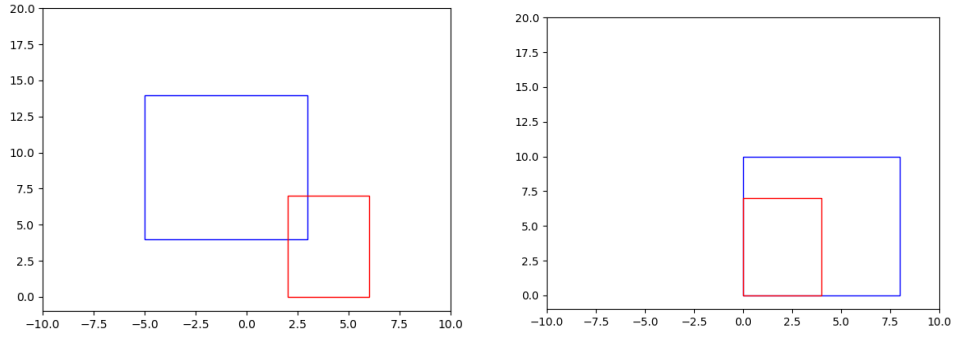
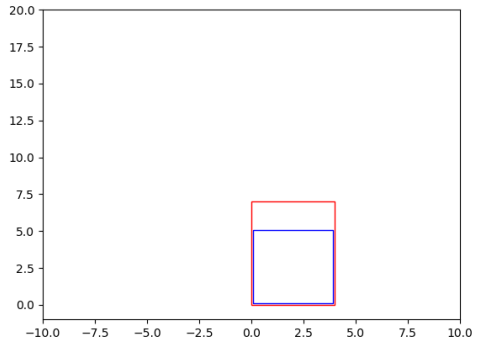


Figure 2: My route tree



(a) (b)



(c)

Figure 3: In (a) the receiver route is bounded in the red box and the pre-defined route in the blue box. In (b) the routes are moved to have (0, 0) bottom-left endpoint. In (c) the pre-defined route has the axis with the biggest discrepancy minimized to fit and still keep the same aspect ratio

few named routes that I defined. I chose to include pretty basic routes. They include a flat, a slant, a curl, a comeback, an out, a dig, a post, a corner, a streak, and a wheel. I have graphed them all in Figure 2. The pre-defined routes had a distinct shape but to use a nearest neighbors approach I had to reshape them to match as closely as possible to the extracted receiver routes in order to have a fair comparison. I made the pre-defined routes extremely large so they had to shrink down. Then to maintain the correct aspect I looked at the largest discrepancy between the max x and max y coordinate for the receiver route and the pre-defined route. Whichever discrepancy was larger I shrunk both x and y coordinates by that ratio. See figure 3 for details. After resizing the pre-defined routes I measured the distance between each point in the receiver route and the closest point in the pre-defined routes. I also did this with the pre-defined routes. Then I took averages of these two distances for each pre-defined route and summed them together. Whichever pre-defined route had the lowest sum was considered the nearest neighbor and the receiver route was labeled the same. Since running backs and tight ends were included I also needed a way to define if they were left back to block or were setting up for a screen. Any receiver who was behind the line of scrimmage when the pass was thrown or did not move more than six yards was labeled as blocking or waiting for a bubble screen. Unfortunately with this approach it is hard to say how good of a job it has done empirically since no true labels exist. The success was therefore measured using the eye test on a few plays in different games. I have included some in game graphics with the routes labeled

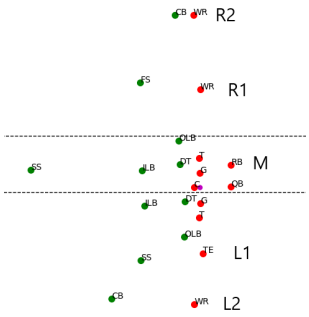


Figure 4: Route positions

and for completeness I will also include a few full games of routes that the judges can compare to true plays if they want. The position that the wide receivers started their route in was also part of the data collection. Originally I wanted to have the field broken into eight parts: 3 on the left, 2 in the middle and 3 on the right. Since there are usually only five positional players though this would leave a lot of holes in my data. I settled on making the positions only totaling five and have them relative to each other. Therefore route_1 would be the route run by the receiver to the left most of the ball, route_2 the next receiver in until route_5 is run by the receiver to the right most of the ball. If five total players of positions running back, fullback, tight end, or wide receiver did not exist on the play I assumed it was because extra lineman were brought in and so added a tight end to the center and gave him a blocking or bubble label. The five positions are from left most to right most are L2, L1, M, R1, R2. This may not be true though as three wide receivers could be on one side of the ball, but the positions are more to help assign names rather than be perfectively descriptive. See Figure 4

Defensive Parameters

A successful passing play depends a lot on how the other team chooses to defend. A wide open receiver will have a much higher chance of catching the ball than a covered receiver. When researching for this project I found that defining useful parameters about the defense can be difficult. For instance there are many types of schemes for how a defense will cover the pass. Many times these schemes cannot be determined by experts or the play is a hybrid of a couple of schemes. The expert eye can decide between man and zone coverage but the usual techniques involve seeing which direction a defender points their hips. This is impossible to discover with tracking data. Therefore instead of labeling a defense as in man or zone coverage I decided to look at the defenders closest to the outcome point. The outcome point is the x, y coordinate of the ball when the pass is complete, incomplete etc. I collected data on the two closest defenders to the outcome point at the frame when the pass was thrown and the frame when the pass reached the outcome point. This seemed to be a fair comprise to collect data on relevant defenders whether they were in zone or man coverage since defenders in zone coverage should show up when the ball is thrown and defenders in man and zone coverage will show up when the ball reaches the outcome point. The data I collected on each defender was their position, the distance to the outcome point, their speed, and the direction they were relative to the outcome point. For direction I wanted to know if the defender was running towards the outcome point, away from it or somewhere in between. Using the direction they were traveling in from the tracking data, the x, y coordinate of the defender and the x, y coordinate of the outcome point I was able to get an angle from 0 - 180 degrees with 0 degrees being staring directly at the outcome point. This should be able to tell me if some routes require misdirection at the point of throwing the ball in order to be successful such as a comeback or curl route. In this same vein I also collected the same information about the intended receiver at the time the pass was thrown. I wanted to know if the receiver should be accelerating toward the ball, not yet facing the outcome point or if a more successful pass happens when the receiver is already sitting near where he needs to catch the ball. Defense is hard to predict and relies heavily on post snap reads by the quarterback to identify which route will be most successful. Using the vast amounts of data available may help present an aggregate solution though on how what routes can help with separation.

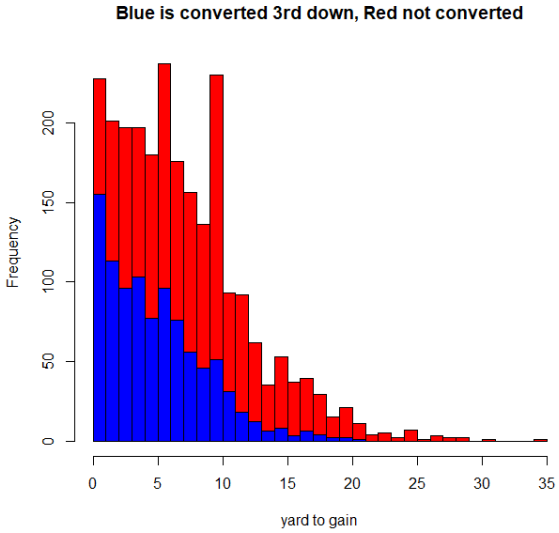


Figure 5: Histogram of frequency of success for third down

Defining Success

Not every pass play that is completed is a success which is why I defined a higher bar for a pass play to be considered a success. The idea of a successful play all hinges on getting a first down before fourth down. Whether a team is just moving the chains down the field or scoring a touchdown, getting the ball past the line to gain is a must. Third down is usually the most vital down to convert so I wanted to look at all the data and see if there is any major change point for success based on yardage. See table for odds of converting a third down at each yard to gain. Also note in figure 5 the histogram showing a visualization of the odds.

Yards to gain	1	2	3	4	5	6	7	8	9	10
Odds of success	2.1	1.28	0.95	1.09	0.75	0.68	0.76	0.56	0.51	0.28

We can see that as long as there is four or fewer yards to go, the offense has odds almost greater than 1 or probability greater than 50% of converting. Working backwards I decided to define success on second down as getting enough yards to get to at least 3rd and 4. Then on first down success is defined as getting at least half the yards required to get to 3rd and 4. For instance a three yard play on 1st and 10 would be considered a success. Then obviously on third down only converting is considered a success and similarly if a team went for it on fourth down only converting is considered a success.

Other Predictors

There was so much information it would have been impossible not to mine a little more out of the tracking information. These are mainly blocking variables that may not necessarily be able to be influenced by the offense in their route choices but affect the play none the less. The first is whether a blitz came after the quarterback defined as at least five defensive players crossing the line of scrimmage. Related to this is the number of defenders close to the quarterback when the ball is thrown where close is defined as within two yards. The time between the snap and the pass as well as the time between the pass and the outcome was of interest to see how long certain routes needed to develop. The distance between the outcome point and the closest sideline was included to see if the sideline helps or hurts the offense. Whether the pass came from the pocket or not was recorded to see how well bootlegs or broken plays compared to a quarterback hanging in the pocket for the duration. For offense and defense scheming I recorded whether the quarterback was in shotgun and the number of defensive backs were on the field. These are both pre-play considerations and I made sure to consider the positional match ups when looking at the intended receiver and the closest defenders.

Analysis

In this section I will use the data I have gathered to argue route combinations and spatial characteristics about the receivers and defenders have similar information and therefore can be thought of as interchangeable. This further implies that the route combinations can be assumed to be a sufficient condition for certain spatial characteristics including those that are conducive to having a successful play. My analysis is broken up into two parts: the first part is building a model on the spatial characteristics and other data to see what non-route information indicates a successful play; the second part is to find successful route combinations and see if they are also deemed successful using the model built in part one. Finding route combinations that were successful in a real game and predicted to be successful based on the spatial characteristics that arose out of them would mean the two data types have similar information and are connected. With this technique I was unable to get a straight forward ranking of the routes so instead I will go over a few route combinations that I found were successful and how they relate to the spatial characteristics. It is important to note that since both parts use route success it would be incorrect to overlap the data used so I used one third of the data for part 1 and two thirds of the data for part 2.

Summary of the Data

The full data I was successfully able to put together totaled 5,811 rows with a pass success percentage of 48%. This is handy for model building as the responses are nice and balanced. On 28% of the plays there was a blitz, 87% of throws came from inside the pocket and on average the quarterback was faced with 0.81 defenders in his face when throwing the ball. Surprisingly 78% of plays started in shotgun but it is important to note that I am not looking at any running plays. On average there were 4.94 defensive backs on the field.

Building a Model

My original thought was to build a model with all the route information, the spatial characteristics and the other data being treated as blocking variables. This proved to be an insurmountable task because to get an idea of effective route combinations I would have needed interactions and dummy variables. When I have five positions that receivers run from and each position has nine routes and each combination needs its own additive affect the extra variables needed added up quickly. Instead I only focused on the spatial characteristics and other data of each play which was much more manageable. I decided to stick with a pretty simple model since the goal is not prediction so much as explanation. I want to know how success fluctuates when spatial characteristics are changed. While adding in interaction and squared effects would make the model fit the training data better I believe it would not contribute much to explaining the effects of the model. The model I chose to use was logistic regression with a backwards step using the AIC criterion. I have found this technique has helps trim the fat from the models and bring out the important variables. Starting with the full model which contained parameters

> *blitz, closest defender direction, speed, distance, and position at outcome, second closest defender direction, speed, distance, and position at outcome, closest defender direction, speed, distance, and position at pass, second closest defender direction, speed, distance, and position at pass, closest sideline, in the pocket, intended receiver direction, distance, speed at pass, number of defensive backs, number of closest defenders to quarterback at pass, L2 route depth and position, L1 route depth and position, M route depth and position, R1 route depth and position, R2 route depth and position, time between snap and pass, and time between pass and outcome*

After backwards elimination the model was reduced to

Predictor	Coefficient	Probability Value
Closest defender direction at outcome	1.39	0.003
Closest defender distance at outcome	-0.004	∼0
Closest defender speed at outcome	0.23	∼ 0
Closest defender direction at pass	-0.004	0.002
Closest defender position at pass DL	-0.699	0.006
Closest defender position at pass LB	-0.133	0.43
Second closest defender speed at outcome	0.07	0.08
Second closest defender position at pass DL	-0.79	∼0
Second closest defender position at pass LB	-0.16	0.29
Closest sideline	0.03	0.00003
In the pocket	-0.24	0.233
Intended receiver direction at pass	0.006	∼0
Intended receiver distance at pass	-0.82	∼0
Number of defensive backs	-0.19	0.07
Shotgun	-0.24	0.17

Assumptions necessary for a logistic regression model (of which there are very few) show nothing too extraneous and the residual deviance is low enough to not consider rejecting the model. When looking at these summaries I am most interested in the probability value and the sign of the coefficient attached to each predictor. A low probability value implies that the coefficient is not likely to be equal to zero and therefore the sign can be interpreted to mean how adding a unit of the predictor affects the overall success positively or negatively. With this in mind the variables with low probabilities that have signs that make sense to me are the closest defender-at-the-outcome's distance and speed and the closest sideline, the intended receiver distance, the second closest defender at the pass' speed, and the number of defensive backs on the field. All of these coefficients make sense in context of running a successful play such as increasing defenders defense increases probability of success, passes toward the middle of field have a higher probability of success, and slower defenders make it easier to catch a ball up field. The coefficients of the directions of the intended receiver and defenders do not make much sense to me however. I would expect the opposite since lower directions indicate a player is looking at the outcome point. Yet defenders have a negative coefficient and intended receivers have a positive coefficient. The intended receiver is measured at the frame of the pass so this might imply some mis direction is needed when the pass is thrown. Overall though even though these coefficients have low probability values they also have extremely low effect sizes so only a massive swing in one direction or another will change the success of the play drastically. The other strange coefficient is the one for the position of the two closest defenders at the pass. Intuitively a defensive lineman covering a route would indicate that a catch and run for modest gain is likely. What I hypothesize though is that when defensive lineman are close then that means the play is akin to a short dump to a running back. These plays can be successful but are also utilized when all other option have run out and a quarterback is only looking to avoid a negative play so not high probability of success plays on average. Distance of the closest defender at the pass and the outcome have opposite signs. The outcome sign makes sense but at the pass could mean a couple of things: man coverage is better for success or the covering defender should be running past the outcome point at the time of the pass. More exploration on the relationship between man coverage and passing success is needed. Another interesting thing to note is the complete drop off of the receiver information such as the position and route depth. This is most likely due to the shifting positions and lack of more crucial route information. While I believe route depth is an important factor to consider, I did not have the infrastructure in my model required to support it.
As far as routes go what this says to me is that successful routes will attempt to out pace defenders and use misdirection to make a defender slow down when the pass is thrown and need to catch up as the pass is arriving. The successful route combinations I found in the next part take steps toward achieving these goals.

Route Combinations

In this section meticulous tabulation was used to find successful route combinations. I used the other two thirds of the data since I feared that finding such specific combinations would subset the data quickly. My approach was to look at each of the five positions that a receiver could run out of and look at a table of success against the routes when the receiver was the intended target. At this stage I was only

interested in what routes were being targeted by the quarterback. From this I found the most successful routes at each position and then proceeded to look at the routes ran in the adjacent positions to see if any of them stuck out. As an example looking at the R2 position which is the position furthest to the right of the ball I see that post has done the best and has a sufficient number of observations. Next I took the rows where a post route was the intended target in the R2 position and looked at the routes run in the R1 position which is one towards the middle. I found that a slant was highly successful when run in conjunction. The next step was to look at the predictions of success of these plays when plugged into the model from the previous part. A high prediction of success would indicate that this route combination has the spatial characteristics deemed to be sufficient to run a successful play. I will also provide game and play Ids showing the route combination in action and how the model coefficients line up with the players actions.

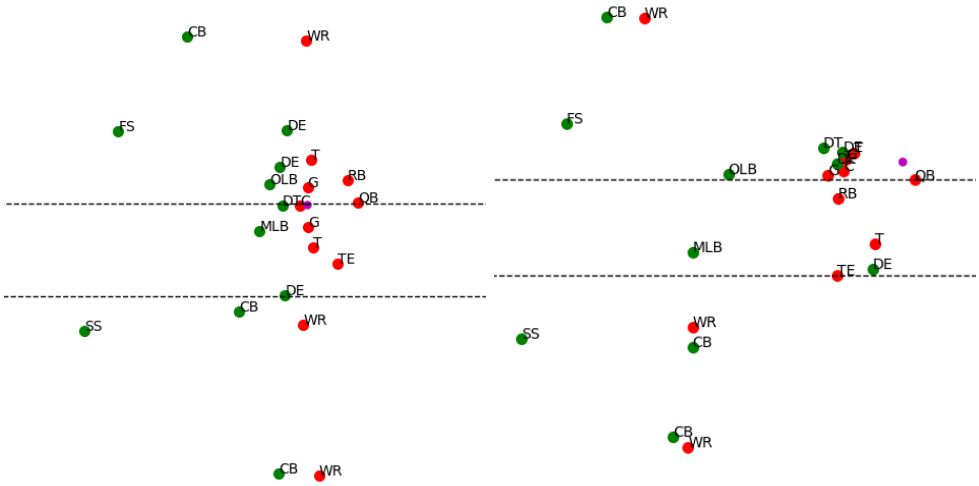
A Post and a Flat

This route combination will be when the intended target is a receiver in the L1 position running a post route while the receiver in the M position will be running a flat route. As can be seen in gameId 2017092406 and playId 1802. When looking at the play some things that are apparent are that the defense is in a zone coverage and the middle line backer is responsible for players in the middle of the field. At the time of the pass and the outcome the closest defender is the strong safety. At the time of the pass the strong safety is backing away and changing direction to attack the ball. The receiver seems to slow down before catching the ball to sit in the hole so is extremely close to the outcome point when the ball is thrown. At the time the ball gets there the strong safety is running quickly towards the outcome point trying to catch up. All these are predictors that the model in part 1 says contributes to a successful pass. The other consideration is the flat route being run underneath to pull away the middle linebacker. This is a common theme in many successful plays that I have seen: misdirection for the deep players and overloading the short players. In the end this play nets a 17 yard gain. See figure 6. There are 13 plays in the segmented dataset that matches all these route constraints. The model from part 1 when applied to these rows was able to predict 10 of them correctly for an accuracy of 76%. This matches close to what the model was able to achieve in testing. This implies that knowing only the routes run from these two positions matched providing all the information from the spatial characteristics. This is an extremely small group though so more examples are needed to verify.

Two Digs

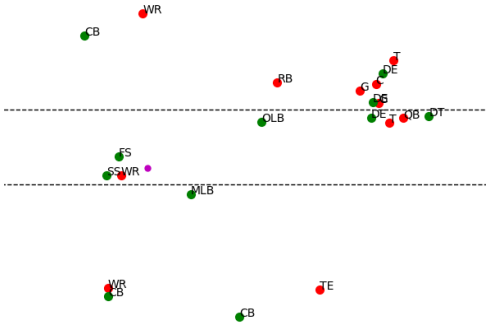
In this route combination the wide receiver in the R1 position will be the intended target running a dig and the receiver in the R2 position will also be running a dig. An example of this play is gameId 2017091004 and playId 2293. The defense in this play is playing a combination of man and zone. The free safety has outside leverage while the linebackers need to provide short and inside support. The closest defender to the outcome point is the right inside linebacker. At the pass he is just starting to accelerate towards the outcome point. Then at the outcome he speeds up to the point too late to block the pass and runs through the ball while the covering free safety is trying to play catch up from behind as well running towards the wide receiver at the pass and the outcome. From the model's perspective the good things about this route are the direction the linebacker faces at the pass, the speed and direction of the linebacker at the outcome, the speed of the safety at the pass and the distance from the sideline. The second dig route is crucial as once again the routes overload the short players and make them decide who to leave open while the deep players are misdirected by the in cuts. See figure 7. There are 24 plays which match this type of combination and of them 18 are considered a success. When run through the model from part 1 there is a 75% accuracy or 18 correct predictions. Note that 18 true successes and 18 correct predictions are not directly related. Again this shows that successful routes also exhibit the spatial characteristics believed to lead to success from part 1.

While both of these examples are extremely small sample sizes, it is a good start to matching route combinations and player actions to see patterns of success. The two parts are independent given the segmenting of the data at the beginning so success in one part should not directly impact success in the other part. Finding success using both route combinations and actions between the parts is the key.

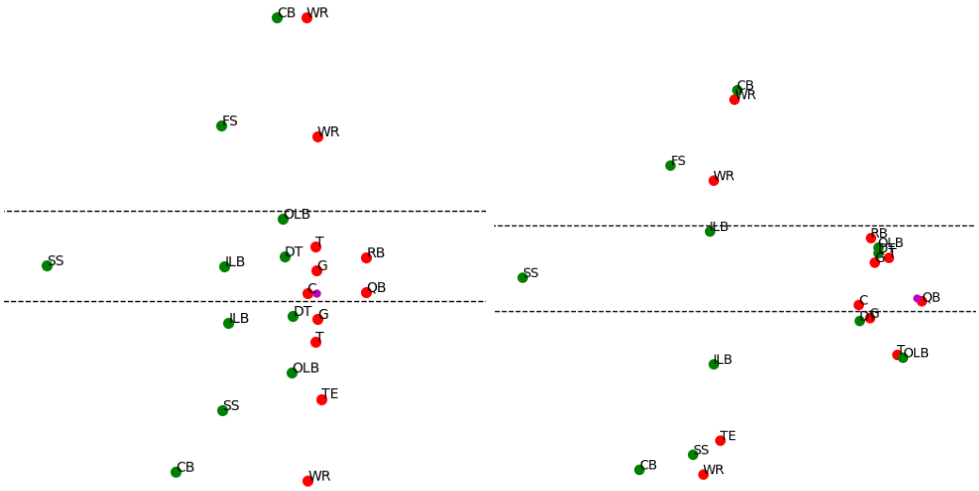


(a) (b)

Figure 6: Still shots of a Post and a Flat

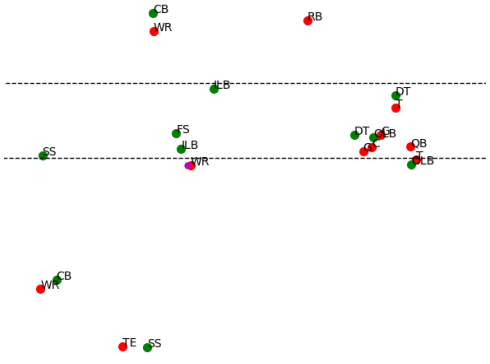


(c)



(a) (b)

Figure 7: Still shots of a Dig and a Dig



(c)

Conclusion

Discovering the optimal combinations of wide receivers is a step towards introducing data science into the football decision making process. In this project I built an unsupervised method to label simple wide receiver routes as well as define a method to identify what routes were considered good combinations. I used a model that was able to accurately identify what spatial characteristics lead to a successful pass and find route combinations that were empirically successful. These route combinations caused the spatial characteristics of a successful pass as defined by my model. While this is not a definite ranking, evidence of successful routes causing players to react in a way that predicts a successful play is a good start. My code will be available on my github at https://github.com/kinne174.

If I had more time...

I would like to explore more with the response of yards after the catch and attempting to classify defensive schemes. Working with a binary response of pass success gives limited information whereas looking at not only if a pass was caught but also how to design a play so there is running room after is an interesting challenge. A big part of the success or failure of a route is dependent on the defensive coverage. Looking at plays it was sometimes clear to me if a defensive back was shadowing a receiver or if they dropped off into a zone coverage after the pass. Knowing the type of coverage ahead of time can influence the play call or switching the play at the line of scrimmage. This was a lot of fun and I'm glad I was able to delve into the data and see football, a game I love, from a statistics perspective. Thank you and go Broncos!