Exploratory Data Analysis of Passing Plays using NFL Tracking Data

Adam Vonder Haar

Introduction

This report presents several methods of exploring NFL tracking data with the intent to uncover what makes passing plays work well. Passing plays are complicated, and a strong understanding of defensive alignment, coverage, receiver pass routes, and the combinations of all of the above is helpful for expanding this understanding. This report will be broken down in the following format:

- Cleaning and Restructuring Tracking Data
- Defensive Exploration
  - Role Classification
  - Starting Formation Classification
  - Convex Hull of Pass Coverage
- Offensive Exploration
  - Role Classification
  - Starting Formation Classification
  - Receiver Route Classification
- Combination Analysis

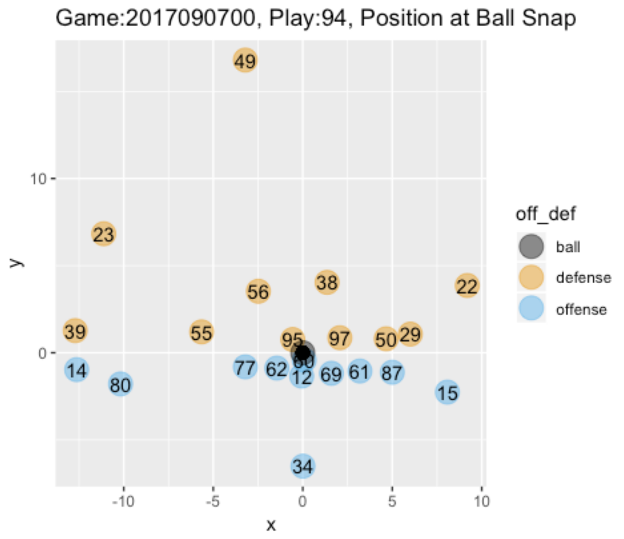
Cleaning and Restructuring Tracking Data

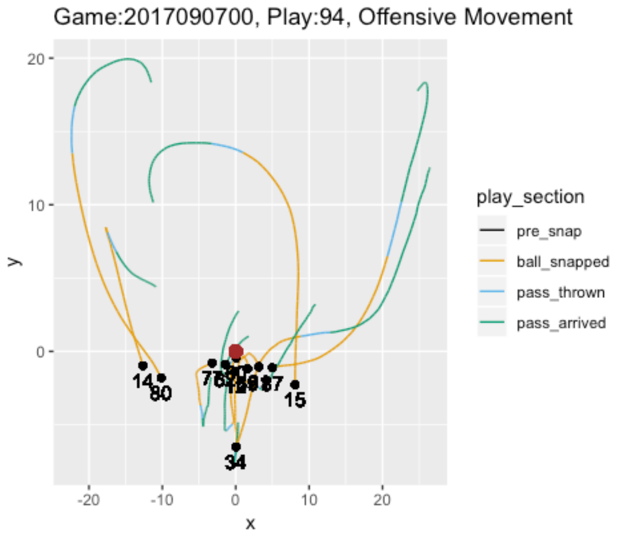
To begin, all data provided (tracking data for each game, game metadata, play metadata, player metadata), along with data pulled from nflscrapR, is loaded in and merged together. After all data is combined, some cleaning is done to standardize aspects of the data and also to normalize it with respect to the same directions. Additionally, "clean" pass plays are selected for this analysis, meaning passing plays that were relatively easy to identify and didn't require excessive manipulation to clean up discrepancies in the data.

An example of some cleaning that was done was with the event variable. There are cases in some plays where two different frames represent one event occurring, such as game 2017090700, play 4006 where the "ball_snap" event is shown to be occurring at both frame 13 and frame 15. In these situations, the event is naively assigned to the lowest numbered frame.

Pass plays were selected based on a set of criteria that aimed for passing plays in normal situations, and did not involve any runs. The potential outcomes for these plays included completed pass, incomplete pass, or interception; quarterback scrambles, sacks, and two-point conversions were not included at this time. Additionally, plays that had missing tracking data for any of the 22 players or the ball, along with plays that did not have the "ball_snap", "pass_forward", and "pass_arrived" events were excluded at this time. There was also some subjective exclusion of plays where the tracking data did not line up in a way that was expected, such as the ball at snap being positioned 15 yards behind the rest of the players. In the future it would be possible to clean the data to include more potential plays, but this process provided a large representative sample of plays where receivers ran routes and a pass was thrown. In the end, a total of 3487 passing plays were analyzed.

Finally, all plays were oriented to be "behind" the offense, that is with all movement up the field for the offense (toward the opposing team's end zone) correlating to an increase in the y axis, and movement from left to right across the field correlating to an increase in the x axis. The origin (0,0) for each play is equal to the position of the football at event "ball_snap". This orientation is different than that described in the data schema, but appeared more intuitive to me for plotting purposes. Additionally, plays were segmented based on the events that occurred during the play. An example of a cleaned, reoriented passing play is shown below, both with the position offense and defense at ball snap and also with the movement of the offensive players throughout the play.

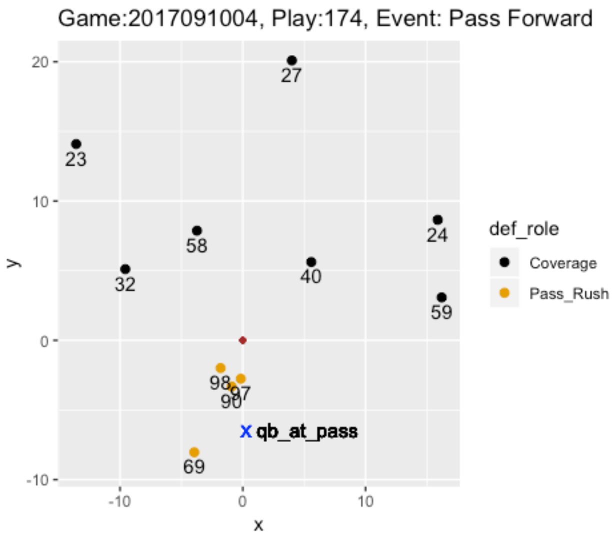




Defensive Analysis

Role Classification

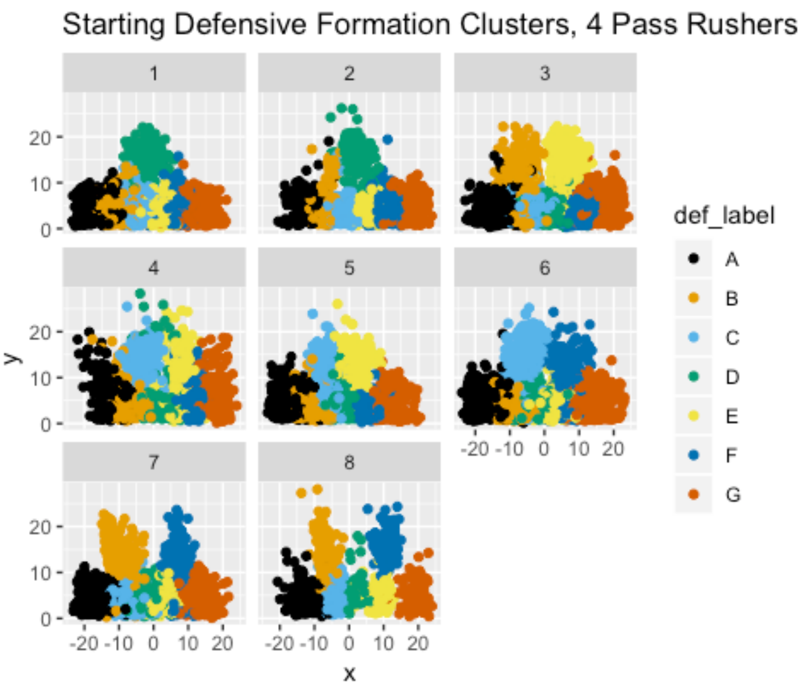
On the defensive side of the ball, this particular analysis is primarily concerned with what the pass coverage is doing during a play more so than the pass rush. Obviously, pass rush has an important role in the success of passing plays, but some work has already been done in that area such as with Brian Burke's Pass Block Win Rate stat. One way to view the split in roles is that pass rush is attempting to expedite the quarterback's decision with where to place the ball, while the pass coverage is attempting to limit his options for where to place the ball. This analysis will attempt to explore how well the offense is able to prevent the coverage from limiting the quarterback's options. In order to the analyze the pass coverage as a distinct unit, each defensive player must be assigned a role of either pass rush or pass coverage. To accomplish this, a variable provided in the play metadata, "numberOfPashRushers", is used in combination with each defensive player's distance from the quarterback at the time a pass is thrown. If a play is known to have 4 pass rushers, the role of pass rusher is assigned to the four players closest to the quarterback at the time the ball is thrown, and the role of coverage is assigned to the other players, as shown to the right.



Starting Formation Classification

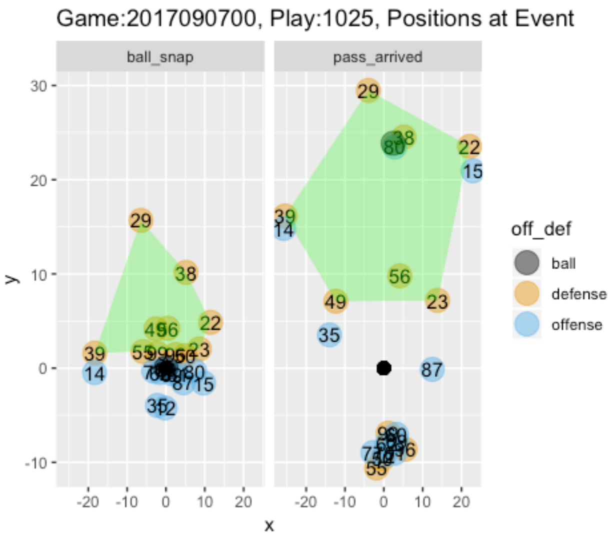
To explore potential starting formations of the defensive players in coverage, plays are first split by how many players were in coverage. 61% of the plays had 4 pass rushers and 7 players in coverage, so plays with this personnel grouping will be shown as an example. The roster position (CB, SS, MLB, etc.) is provided for each player in the player metadata, and while that is typically an adequate proxy for defensive role, this analysis ignored that variable in the case that a defensive player's roster position didn't correlate to their role they were serving on a particular play. Instead, coverage players were given a letter label relative to their starting position from left to right (with respect to the offense). For example, the player lined up farthest to the left was labeled the "A" player, and the player furthest to the right was given the label of "G" (in the 7-coverage player example). Hierarchal clustering was then performed in an attempt to group similar starting formations together, with the primary variables being the coordinate location of each coverage player at ball snap.

These clusters, along with their differences can be seen in the plot at right Some formations look different in structure, such as cluster 1 with most coverage near the line of scrimmage and one player playing safety contrasted with cluster 3 where there are two players fulfilling safety roles. Other clusters differ not in the structure as much, but which player is playing a role, contrasting cluster 3 with players B and E serving the safety roles and cluster 6 with players C and F serving the safety roles. This type of clustering is done for every personnel type (personnel being the split between number of pass rushers and players in coverage, not related to roster positions).

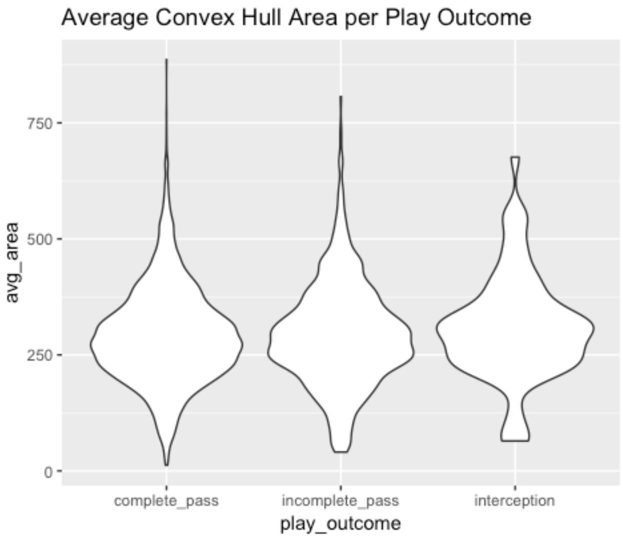


Convex Hull of Pass Coverage

One observation about pass coverage that would seem intuitive is that if the offensive routes are structured in a way that stretches out the defense, there is more available open space to which receivers can run and the quarterback can throw. One way to quantify and visualize this "stretching" of the pass coverage can be accomplished through the use of convex hulls, which can be thought of as a rubber band stretched around the outer edges of the defensive pass coverage. Convex hulls can be created for the pass coverage at different points throughout the play, and statistics related to these hulls can be created, such as the area of the hull at any particular point or the number of defenders found within the hull (more defenders inside the hull would mean less open space for two different hulls otherwise equal in area).



When looking at the average area of the coverage convex hull throughout a play compared to different play outcomes, there isn't a lot of signal initially. At the high end it seems possible that defenses that are especially stretched out throughout the play probably have a lower chance of preventing a completed pass. However more work on the relationship of the offensive players to the hull, along with some other descriptive metrics of the hull itself, is likely necessary for this methodology to add significant value.



Offensive Analysis

Role Classification

To assign roles for the offense is not quite as straightforward as it is for the defense, which is much easier thanks to the numberOfPassRushers variable. There are three general roles that are useful in evaluating pass plays: receivers (including running backs or tight ends that run pass routes and make themselves available for a catch), pass blockers, and the quarterback. The quarterback assignment is simple enough, and the criteria that was used in this case to identify receivers was an observed movement down the field of at least one yard from their position at ball snap to their position when the ball is thrown. There is a possibility this excludes some receivers who run routes parallel to the line of scrimmage, and in some instances might include pass blockers that end up further down the field than they started, but in general it should be a useful method for identifying players that are running pass routes.

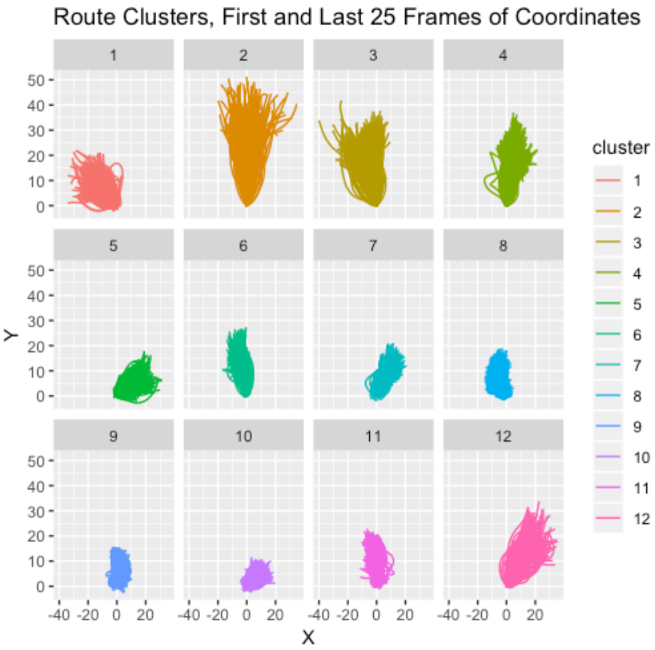
Starting Formation Classification

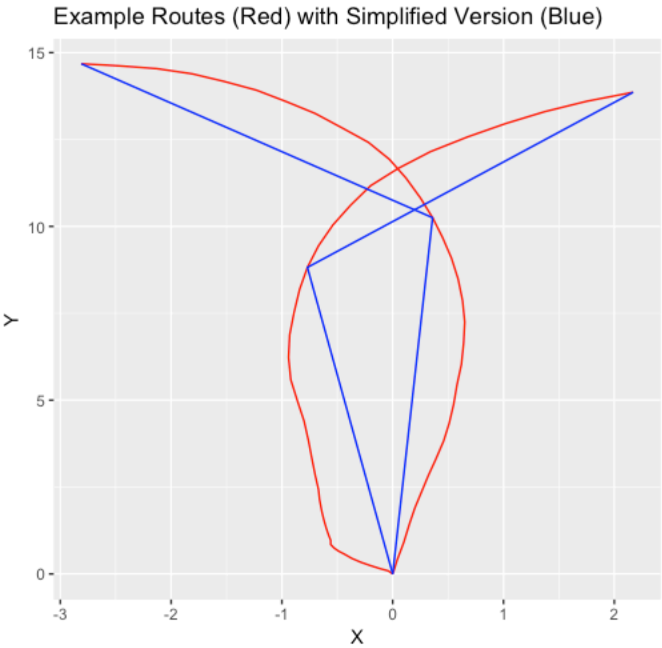
In general, starting formations for the offense are identified very similarly to how they are for defense, using hierarchal clustering to identify similar formations that receivers start in.

Receiver Route Classification

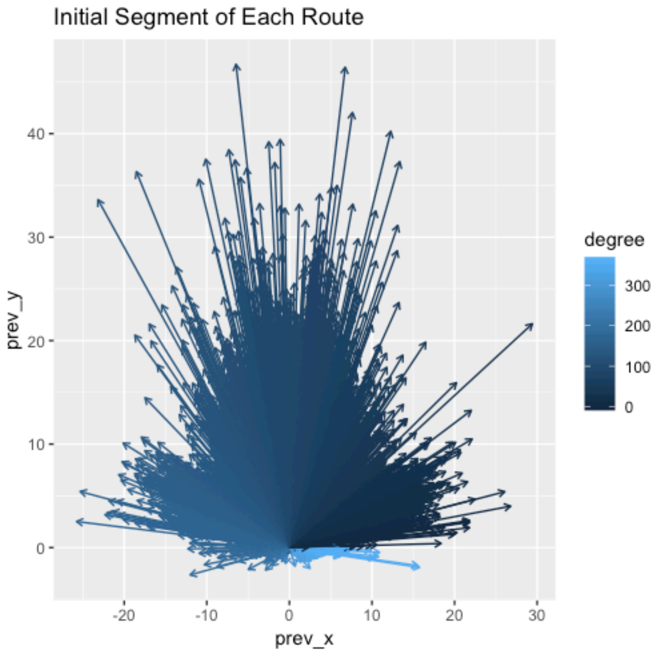
The are several methods that are utilized to classify receiver routes. Potentially, some combination of subjective human identification could help here, but a fully unsupervised approach was used in this case so that minimal assumptions were built into the classification.

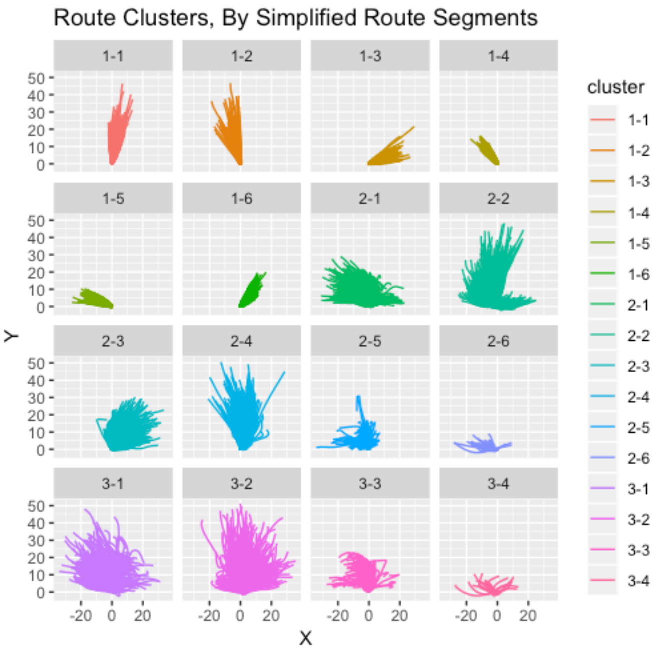
The first approach that was used was to look at the first and last 23 frames of each route. Out of the 15,106 pass routes pull from the selected plays, 14,044 (or 93%) have a least 23 frames of coordinate data. Of those, 12,830 contain 51 frames or less of data, which means that at least 90% of the route would be represented in these variables. For routes with more frames, some of the middle of the route would be missing with this methodology, but it should be sufficient enough for these purposes. Additionally, because the shape of the route is the concern here more than the location of the route, all routes are normalized to have a start point at the origin (0,0). With these frames of coordinate data, each route was clustered with similar routes using hierarchal clustering. Further work could be done to optimize for the best number of route clusters, but for now 12 different clusters were created, and the algorithm appeared to do a pretty solid job of separating routes. These clusters will be examined further later.





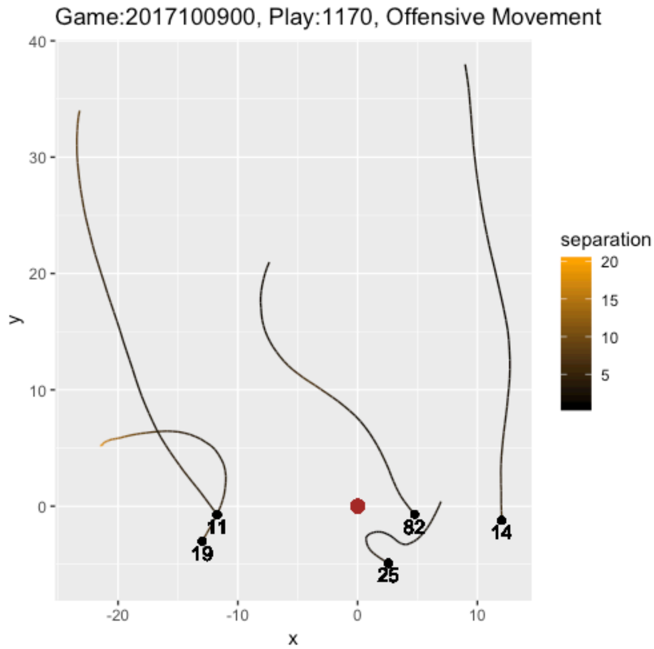
Another way of evaluating routes is to attempt to simplify them before clustering. Often the granularity in the tracking data can add more noise than necessary to the clustering, and if some of that can be stripped out, it may be easier to group similar routes. An example of that can be seen to the left: a random post route and corner route with their simplified versions made up of just two-line segments. This dramatically simplifies the number of points in each route; of the 15,106 routes in the dataset, 14,510 (96%) are composed of three different line segments or less. With these line segments, it is trivial to calculate the length of each segment along with the direction the segment is angled. These degree and distance measurements will be used to cluster these route segments. The initial segment for each route is plotted below, with the length of each segment corresponding to the segment distance in yards, and the direction relative to moving down the field. After grouping similar routes, 16 new clusters are created and are shown below.



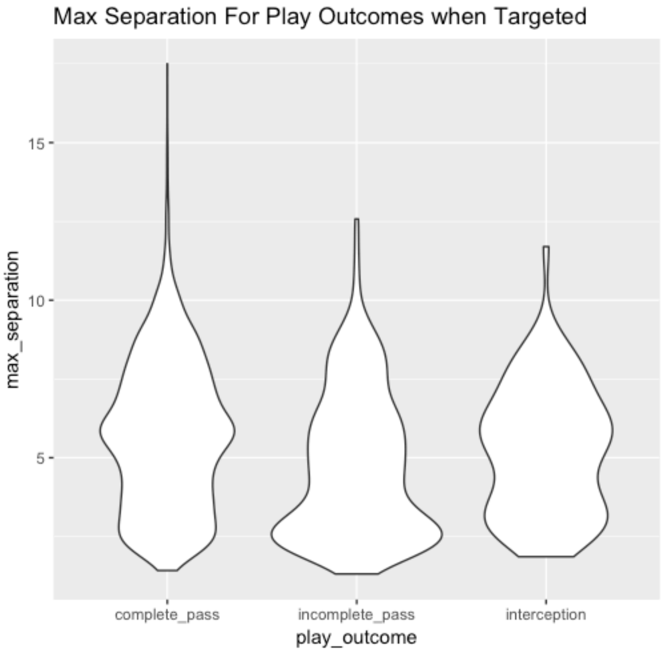


Combination Analysis

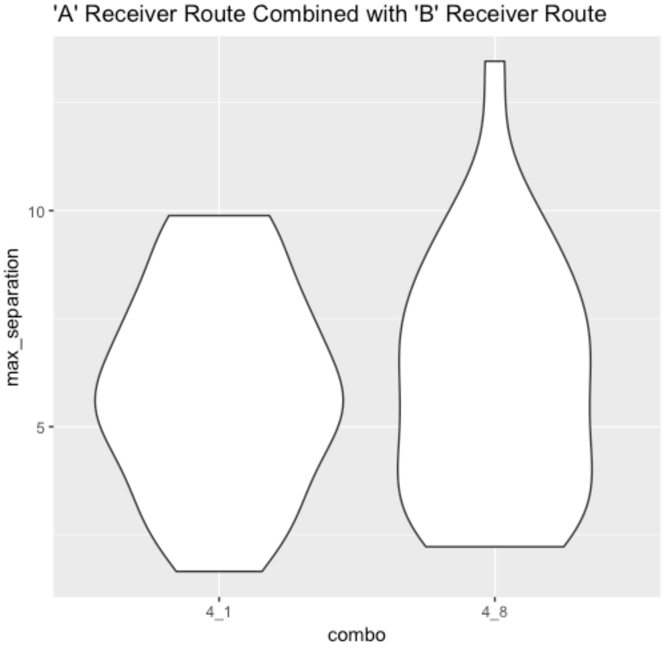
In order to identify which combinations of pass routes are valuable for an offense, it is important to define what might be considered successful outcomes for a particular pass route or pass play. One variable that will be used here is max separation, which is essentially the most “open” a receiver is at any point in his route, defined as the distance from the nearest defender. It is easy to determine each receiver’s distance in yards from his nearest defender at each point during the play, so this variable will just be whenever that distance is the greatest. Separation an important variable to consider because it can potentially reveal situations where quarterbacks are not targeting the most open receiver, or they are not targeting a receiver at his most open point in the route. For instance, the receiver with the highest “max separation” at any point in the route is only targeted on 21% of the plays in the dataset. The outcome of the play can be studied, but it isn’t possible to know how the play might have turned out differently had a more open receiver been targeted at the optimal time.



For instance, on the play to the right, an incomplete pass was thrown deep to player number 14, who had a max separation of about 6 yards from his closest defender. Number 19, however, was 20 yards away from his nearest defender at one point, and likely could have caught an easy pass and gained many yards after the catch had the quarterback not had his eyes locked downfield. This is just the beginning of the type of analysis that could come with separation data.



When looking at players who were actually targeted, it becomes apparent that the ability to achieve maximum separation plays are clear role in play outcome.



Additionally, separation can be used to identify route combinations that work well together. In the plotted example below, the leftmost receiver (the "A" receiver") is runner a cluster 4 route. When combining that with the route of the receiver immediately to his right (the "B" receiver"), an 8-cluster route run by the B is able to achieve better separation for the A receiver than a 1 cluster route.

Tracking data brings so many possibilities for analysis like this, and hopefully this report has been successful in highlighting some of these opportunities.