Use of NFL Tracking Data to Identify Defensive Coverages and Corresponding Offensive Play Outcomes

Taarak Shah

Abstract

Analyzing on-field performance is essential for any NFL team. There are many strategies to limit opposing performance from both an offensive and defensive perspective. We use NFL tracking data on all passing plays from the 2018 regular season, provided by the 2021 NFL Big Data Bowl, to answer questions about offensive and defensive performance. This paper looks into using an unsupervised classification model to identify defensive coverage to be either man coverage or zone coverage. We create variables relevant to this labeling problem then feed it into a Gaussian mixture model to classify the coverages. We then conduct an exploratory analysis of how receivers perform against these coverages. We found that our models led to over 75% of cornerbacks being classified as man coverage defenders, while free safeties and strong safeties were classified as zone defenders over 95% of the time. We also find that receiver play outcomes vastly differ depending on if they run their routes against man or zone coverage. We are able to determine that certain routes perform better or worse against man or zone coverage cornerbacks.

Contents

List of Figures

List of Tables

1 Introduction

The main objective of the game of American football is to score as many points as possible and limit the opposing team from scoring. At its highest level in the NFL, there have been countless years of study, analysis, and strategizing about methods to do this as effectively as possible. With the advent of the NFL's Next Gen Stats division, there are opportunities to use detailed player tracking data to rigorously examine strategies of the game from a statistical perspective. Previously, quantitative analysis of NFL data has been subject to heuristic, situational data about the passer, receiver, and outcome of the play, or statistics such as completions, tackles, sacks, and other basic counting measures.

There was not a way to determine what happened through the course of the play, much less any usable method to determine avenues for improvement until this data became available for use. This data was released to the public domain with the *NFL Big Data Bowl* [13]. The Big Data Bowl is a Kaggle competition held each year for students and professionals alike who are tasked with analyzing trends and performance based on the vast amount of data collected by the NFL Football Operations Division. The datasets consist of details about each play and tracking information of each player through the duration of the play. Now, with this available data, there are many opportunities to closely examine the tendencies of individual players and ways teams can exploit advantageous matchups.

In this essay, we aim to examine two key matchups that define a majority of play outcomes. The first focuses on analysis of coverage schemes for defensive players, specifically cornerbacks and safeties. This analysis was built on an extension of a paper by Dutta et al. [3]. In that paper, the authors examine using a Gaussian mixture model to evaluate classification of cornerbacks into man coverage or zone coverage based on traits such as position on the field, speed, direction, and distance from their opponent throughout the play. Their paper was constructed with data from the inaugural 2019 Big Data Bowl, which was limited and did not have information regarding player orientation, and only included the first six weeks of the 2017 NFL season. This forced them to only consider coverage assignments for cornerbacks, while forgoing analysis for other important positions like free safeties, strong safeties, and linebackers. We construct our own implementation of their methods later in this essay, but include additional engineered features and consider analysis for both free safeties and strong safeties. We choose to forgo analysis for linebackers, mainly due to computational constraints and the necessity to generate additional features specific to linebackers. Linebackers tend to assist in run support more frequently than defensive backs, so we would need different features to distinguish linebackers that are in run support on a given play instead of pass support. We would not be able to get a meaningful result with the same feature set as for defensive backs. This complicates the analysis greatly and is left to future work.

The second aim of this analysis focuses on how wide receivers perform against the two different types of coverage. We are provided with ground-truth labels of the routes that receivers are running, and we compare these with their individual performance against the labels assigned via the clusters in our reproduced Gaussian mixture model. The intent of this analysis was to provide an exploratory overview of how routes and individual receiver performance can vary largely with the type of coverage they face.

We will first detail the in-depth feature creation and consolidation for the coverage analysis, comparing our methodology to Dutta et al. [3], detailing key differences and improvements. We then consider performance of receivers against man and zone coverage, using the cluster labels as ground truth.

2 Methods

2.1 Data

This dataset was retrieved from the 2021 NFL Big Data Bowl [13]. As a whole, it contains all information about every passing play in the 2018-2019 NFL season. Additional information about accessing the data can be found in Section 7.2. There is information on

19,239 passing plays in the season. There are additional files separated by each of the 17 weeks that contain individual frame-by-frame data for each of these 19,239 plays, totalling to 18,309,388 frames of plays. A “frame” consists of a data capture that occurs every 100 milliseconds in time. At each frame, the tracking data contains information on player position, speed, direction, distance traveled since previous frame, and player orientation. This data was collected and verified by standardized procedures through NFL’s Next Gen Stats [16]. Player and football tracking data is monitored via RFID tags in each player’s shoulder pads and an RFID tag in each NFL football. Over 200 data points are created on every play of every game. Additional information about each player is collected and annotated by the NFL. This includes the event of the play, the player’s name, jersey number, position group, team, and if the player is a receiver, the route they ran on the play. A full list of variables provided in the tracking data from the original dataset can be found in Table 7 in Section 7.4. Details on how these features were adapted for our analysis are found later in Section 2.3.

2.1.1 Features in Dataset

A full list of features used in our clustering models (including generated features) is provided in Table 1 in Section 2.3.

We will primarily utilize information regarding their physical location on an xy-coordinate grid, then observe differences in their speed, acceleration, direction, and orientation. The x-coordinate spans from values 0-120, accounting for the 100 yards of the football field and the 10 yard width of both endzones. The y-coordinate spans from values 0-53.3, accounting for the 53.3 yards from sideline-to-sideline of an NFL regulation field. Player direction and orientation measures, in terms of degrees from 0 to 360, are relative to the line of scrimmage, the starting position of a given play.

Figure 1 provides a visualization of the football field and how we determine where the xy-coordinate grid begins and how other directional features are standardized among players.

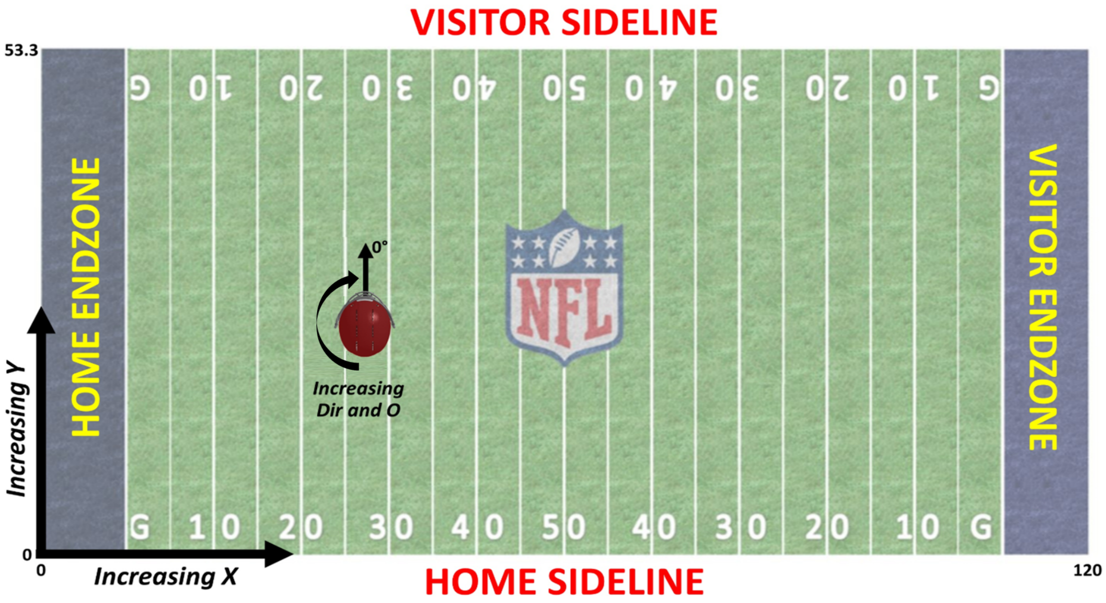

Figure 1: Example of tracking data coordinate information [13]

We can see that the x-coordinate increases from the home endzone to visitor endzone, and the y-coordinate increases from the home sideline to the visitor sideline. The "home sideline" refers to the sideline where the home team stands, likewise for the "visitor sideline". The "home endzone" is where the visiting team scores a touchdown (when the ball crosses x-coordinate 10) and the "visitor endzone" is where the home team scores a touchdown. The home and visitor sidelines are standardized across the unique arenas from when they were first installed [16].

However, there is an inconsistency that needs to be addressed. The rules of the game of football indicate that teams "switch sides" at the end of each quarter of play. This is done to ensure fairness of the game and even playing conditions for both teams throughout the game. This means that the home and visitor endzones are not consistent through the entire duration of the play. Specifically, switching sides means there are entire quarters where the home team will be scoring at x-coordinate 10, and the visiting team will be scoring at x-coordinate 110. In order to account for this, we do not look for the individual locations of the players on the field through the duration of the play. We instead look at how their motion changes relative to other players. Considering player location in this way is not dependent on the location of the field, and makes the inconsistencies in the way home and away teams progress towards the endzone irrelevant. This extends to information about player movement, player direction, and orientation. In Section 2.3, we further examine how we modify these variables as summaries of information for use in our models.

We make use of other relevant information provided in the tracking data, and detail those nuances here. We have information about the event of each frame of the play, such as the moment the ball is snapped, when the ball is thrown, when the pass is caught, and more. We also have important information on each individual player's position. This allows us to separate our results based on the position of interest we want to discuss. For example, we expect that cornerbacks are likely to be more divided between man coverage and zone coverage responsibilities, but we expect that free safeties and strong safeties will likely be playing zone more frequently by the nature of their position.

2.2 Unsupervised Learning Models

Unsupervised learning is often used when we do not have a ground-truth label for the target variable we are interested in. There are many approaches we can take to an unsupervised learning problem, but we specifically focus on clustering in this essay. Clustering is when we use a feature space to group observations independent of a target variable. Clustering will help determine what groups exist in our data and what similarities or differences we can find between our groups.

Two clustering approaches were considered: a Gaussian mixture model (GMM) with 2 components and an unconstrained VVV covariance matrix, and a K-means clustering model with 2 clusters. This was done to compare what is traditionally considered a "soft" clustering method (GMM) versus a "hard" clustering method (K-means). The difference between a soft and hard clustering method is that in a hard clustering method, a single observation either belongs to a cluster, or it does not. In a soft clustering method, we are able to estimate probabilities of belonging to a certain cluster. We evaluate both methods to determine which output should be used for our analysis and results. Both models were fit with the `scikit-learn` package in Python [17]. Information and description of these methods in Section 2.2.1 and Section 2.2.2 were provided by Hastie et al. [6] and James et al. [8].

2.2.1 Gaussian Mixture Model

The Gaussian mixture model is a type of clustering algorithm that fits a mixture of probability density functions to a dataset, where each density is representative of a single group or cluster [11]. In a Gaussian mixture model, we carry out the following procedure. We assume X_i is from a mixture of Normal distributions with the probability density function:

$$f(x; \Phi_k) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, V_k)$$

where $\phi(x; \mu_k, V_k)$ is the probability density function of $N(\mu_k, V_k)$. Each component k is a cluster with prior probability π_k, constrained by $\sum_{k=1}^{K} \pi_k = 1$. For a fixed K, we use the expectation-maximization (EM) algorithm to estimate ϕ_K, to obtain the maximum likelihood estimate. This process is as follows. We specify k multivariate Gaussians; these are the k components. We initialize their mean and variance randomly. We then calculate the probability of each observation being produced by each of the k components, then assign the observation to the component with the highest probability. We then update the mean and variance of the component to reflect the mean and variance of all of the observations assigned to that component, then we iterate until convergence. Detailed information on this process can be found in McNicholas [11].

In our analysis, we chose to use 2 components since that was determined to provide the most clear and natural separation between man and zone coverage. We verify which cluster corresponds to the type of coverage in Section 2.3. We expect to have high collinearity in the feature space, so we use the VVV parameterization of the covariance matrix. A full table of possible parameterizations for the covariance matrix V_k is provided in Table 1 of Fraley et al. [4]. This is also identical to the number of components and the covariance structure used in the paper by Dutta et al. [3].

2.2.2 K-means Clustering

The K-means algorithm partitions our feature space into distinct, non-overlapping clusters. This method was originated by Lloyd [10]. The key difference between K-means and Gaussian mixture models is that the Gaussian mixture model assigns probablistic assignments to clusters, while K-means assigns deterministic assignments. This goes back to the difference between hard and soft clustering methods. In K-means, we carry out the following procedure, described by James et al. [8]. We specify k centroids and randomly decide their cluster centers M_k, initializing their coordinates at this random location. We randomly assign each observation in our data to a cluster from 1 to k, which will serve as our initial cluster assignments. We then iterate over the following two steps:

1. Compute the centroid for each of the k clusters, and the distance of each data point to the centroid. This centroid is the vector of p feature means for the kth cluster.

2. Assign each data point to the cluster with the nearest centroid, where closest is defined by Euclidean distance.

This process continues until cluster assignments cease to change. Further information about the K-means algorithm can be found in Jin and Han [9].

In our analysis, we utilized K-means clustering to provide an alternative to our coverage labels from the Gaussian mixture model. This was done to demonstrate performance with a hard clustering algorithm. We only have information on the final cluster labels, but nothing about the probability of assignment to a particular cluster. We do not use these coverage labels when examining our analysis for corresponding offensive players. This is solely done to compare our Gaussian mixture model results and observe if different clustering algorithms perform relatively similar in classifying man and zone coverage by position with the feature space we have defined.

2.3 Clustering for Coverage Assignments

We focus on the two models described and their respective results for clustering between man and zone assignments. We first standardize the data with the `StandardScaler` function from `scikit-learn` [17], so that each feature has mean zero and unit variance. This is general convention for working with data that exists in different units (for example, our xy-coordinate grid is presented in yards, while orientation information is presented in degrees). We fit a Gaussian mixture model with 2 components, following the methodology described in Section 2.2, and a K-means model with 2 clusters. This is done for three subsets of players: cornerbacks, free safeties, and strong safeties.

We are given information on many useful events that occur throughout the duration of the play. We are specifically interested in generating a label that examines how their coverage label may be considered man or zone coverage at each major event of the play, not necessarily at each frame. We have information about the event occurring during a play, mentioned in our overview in our tracking data in Section 2.1.1. A more detailed breakdown of these events is provided in the full table of features provided in our dataset in Section 7.4. We consolidate the levels of the event variable to simplify our events to three time frames: before the snap, after the snap but before the ball is thrown, and after the ball is thrown. This allows us to focus on relevant moments in the play and determine how coverages shift during the course of a play, and provide better clustering performance based on how players move before and after the ball is in play. The events over which we aggregate the play were similar to those in Dutta et al. [3]. Figure 2 below, also from that paper, shows the breakdown of the events over which we calculate our features.

For each of the three events for during each play, we calculate the groupings in Table 1 of our key tracking variables. This helps to provide a data reduction and a summary of how player motion changes through key events in each play, without losing much individual information. The calculations for mean and variance are aggregated over the frames in each event of the given play. Intuitively, we expect that each of the attributes provide useful

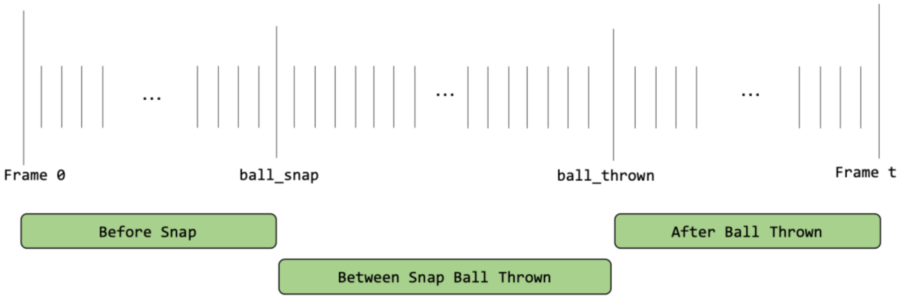

Figure 2: Events over which we generate different features [3]

information in determining coverages. The variance in both the X and Y coordinate give us information about how much the player is moving during different stages of the play. The actual location of the players is less relevant than how much they move over a given event in the game. Similar justification was used in only computing the variance of the speed of the players, since we typically expect that players in man coverage will be moving faster than those in zone coverage in order to keep up step-for-step with the opponent. We also compute features related to the orientation of the player, relative to both the line of scrimmage and the nearest opposition player. The remainder of the features were replicated based on Dutta et al. [3]. Their paper did not have access to information about player orientation. This improvement and access will allow us to examine more nuance in coverage labels between safeties and cornerbacks.

Predictor	Description
`var_X`	Variance in the x coordinate
`var_Y`	Variance in the y coordinate
`speed_var`	Variance in the speed
`opp_var`	Variance in the distance from the nearest opponent player
`opp_mean`	Mean distance from the nearest opponent player
`teammate_var`	Variance in the distance from the nearest teammate
`teammate_mean`	Mean distance from the nearest team mate

`opp_dir_var`	Variance in the difference in degrees of the direction of motion between the player and the nearest opponent player
`opp_dir_mean`	Mean difference in degrees of the direction of motion between the player and the nearest opponent player
`rat_var`	Variance of the ratio of the distance to the nearest opponent player and the distance from the nearest opponent player to the nearest team mate
`rat_mean`	Mean ratio of the distance to the nearest opponent player and the distance from the nearest opponent player to the nearest team mate
`orient_var`	Variance in the orientation of the individual player relative to the line of scrimmage
`orient_mean`	Mean orientation of the individual player relative to the line of scrimmage
`opp_orient_var`	Variance in the orientation of the player relative to the nearest opposing player
`opp_orient_mean`	Mean orientation of the player relative to the nearest opposing player

Table 1: Features used in the clustering models

One important realization is that we lose a lot of information about man-to-man coverage specifically when aggregating over so many frames for our designated events. Variables over this time frame do not truly account for the actual closeness of a defensive back against a receiver. Rather, we have information on how close or far they are *on average* from the opposing player. When looking only at averages, we could have a defensive back slip on the play, and as the receiver moves away from them through the course of the play, the average distance between the defender and receiver will grow to be much higher, even if the initial coverage was meant to be man-to-man. A potentially better way to assess this data would be to look at individual frames of the play, see how the created variables change relative to the previous frame, then collect those averages over the two frames and use that in the training set to build the models. This would take much longer computationally, but we would be able to gain crucial information about how much separation the receiver has from the defender, which would ultimately inform the difference between the coverage labels.

We also lose a fair amount of granularity of our data before the snap with the events we consider. We no longer have an idea about pre-snap motion or how this may influence the decisions a defense makes. It may be useful to assess how motion before the snap on offense influences how defenders decide to play in a man or zone coverage. Looking at what the model indicates at the first frame of the play for coverages and how that subsequently changes with an offensive player movement pre-snap could more appropriately inform the coverage labels.

For both the Gaussian mixture model and the K-means cluster labels, we visually examine a random set of plays and determine which cluster label corresponds to zone coverage and which cluster label corresponds to man coverage. The visualization in Figure 3 was used to examine the plays, without cluster labels pre-assigned. The label shown was either “1” or “2”. Approximately 100 plays for each model were examined to determine which cluster label corresponded to man coverage and zone coverage. This step is admittedly sub-optimal, and that having to manually determine which cluster labels correspond to which type of coverage may be an indication that cluster analysis is not the best method to approach this problem. The practicality and implementation of the algorithm is limited if subject to this manual review process.

It is relatively straightforward to distinguish these assignments, depending on position. If a cornerback is playing zone coverage, they often allow their nearest opposition player to pass by them and focus on another player during the play. If a cornerback is playing man coverage, they often follow their nearest opposition player through the duration of the play. If it was marked over the course of many plays that a free safety with no opposing player nearby was marked as cluster label 1, then that was a fair indication that the cluster label 1 corresponds to zone coverage. If a safety is playing zone coverage, they will often begin the play far from the line of scrimmage and hover in the deep portion of the field, behind the other defensive players. We do not expect large variation in movement from frame-to-frame, especially prior to the ball being thrown, as they tend to hover around the deep portion of the field. If a safety is playing man coverage, however, they may begin the play far from the line of scrimmage and accelerate towards an opposition player and follow them through the duration of the play. This is represented in our feature set as `opp_var` and `opp_mean`, the mean and variance of the distance from the opposing player at each event of the play.

The code for visualizing each play in Python is available publicly on Kaggle [14]. In Figure 3, there is an example of two frames throughout the course of one play visualized with correctly assigned labels and the corresponding routes ran by the receivers on the play. The correctly assigned labels are determined after our visual examination, while route labels are provided by the NFL in the tracking data. We see that intuitively, the free safety #20 is labeled as a zone defender, given there is no opposing player near him. The cornerback #28 is correctly labeled as a man-to-man defender, as his opposing receiver #19 is running a "go" route, meaning he runs in a straight line toward the endzone, and #28 is following him directly throughout the play. The other assignments for this play, man coverage for #25 and zone coverage for #22, reflect their total movement throughout the play as well and their position relative to the opposing players.

We examine the results of our clustering algorithms in Section 3 and discuss the differences between the Gaussian mixture model and K-means, along with useful insights about our coverage labels.

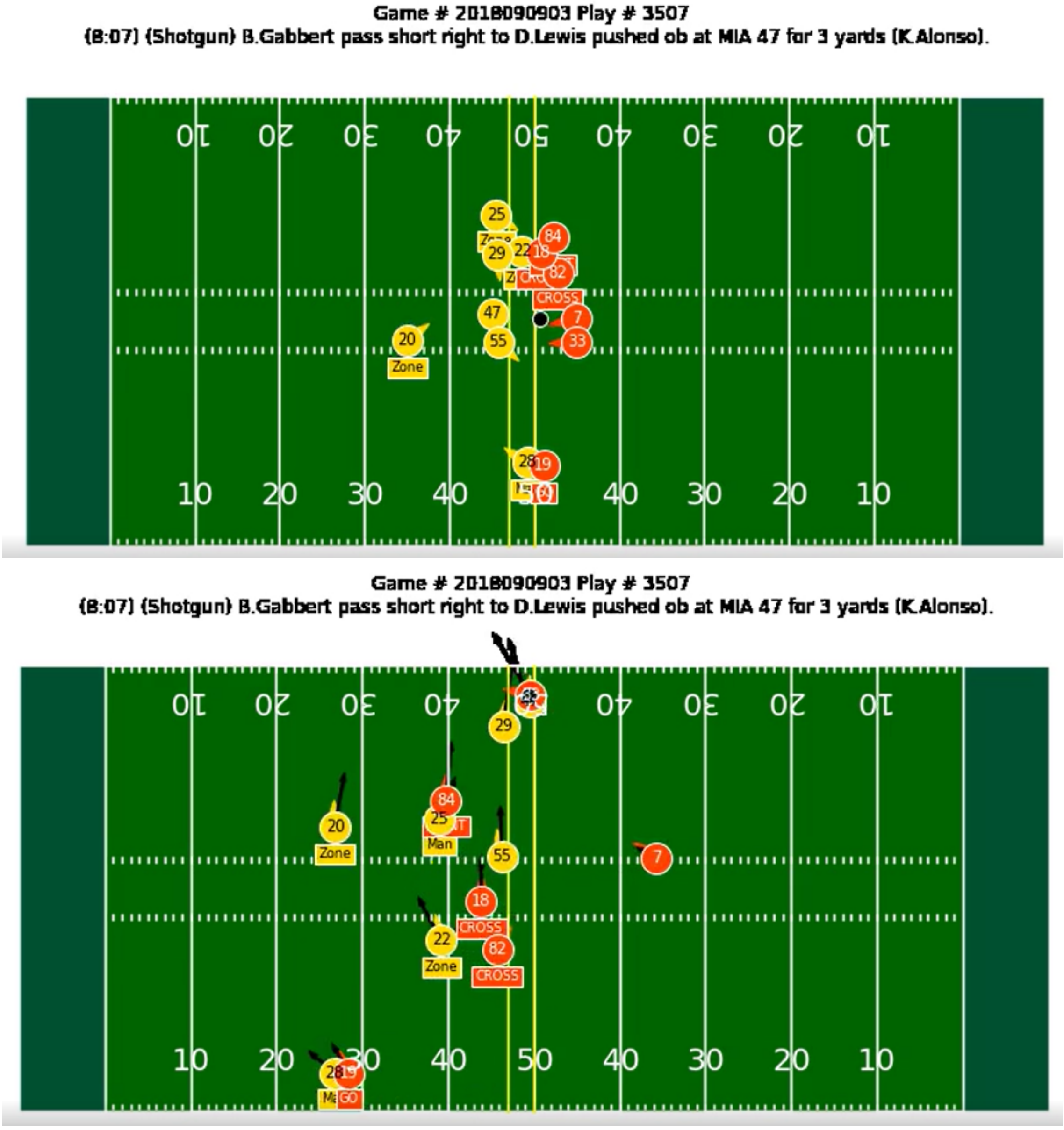
Game # 2018090903 Play # 3507
(8:07) (Shotgun) B.Gabbert pass short right to D.Lewis pushed ob at MIA 47 for 3 yards (K.Alonso).
10
20
30
40
50
40
30
20
10
25
29
22
84
18
82
CROSS
47
55
7
33
20
Zone
28
19
Game # 2018090903 Play # 3507
(8:07) (Shotgun) B.Gabbert pass short right to D.Lewis pushed ob at MIA 47 for 3 yards (K.Alonso).
29
84
25
Man
20
Zone
55
7
18
CROSS
22
Zone
82
CROSS
28
GO

Figure 3: Play visualization frames [14]

3 Clustering Results

Since the objective of this essay is to be considered to be an extension of the paper by Dutta et al. [3], we first look into a replication of their clustering results without the orientation features. These results are represented numerically in Tables 2 and 3, and visually in Figure 4. These tables show the breakdown of our results by each position of interest in our analysis. It is important to note the original analysis only considered cornerbacks, and only analyzed the first 6 weeks of the season, and did not use K-means clustering. Their Gaussian mixture model results showed an approximate 60/40 split between man and zone coverages respectively among cornerbacks.

Position	Zone Percentage	Man Percentage
FS	0.8756	0.1244
SS	0.9385	0.0615
CB	0.5294	0.4706

Table 2: Gaussian Mixture Model Results by Position, without Orientation

Position	Zone Percentage	Man Percentage
FS	0.9908	0.0091
SS	0.9895	0.0105
CB	0.9909	0.0090

Table 3: K-means Results by Position, without Orientation

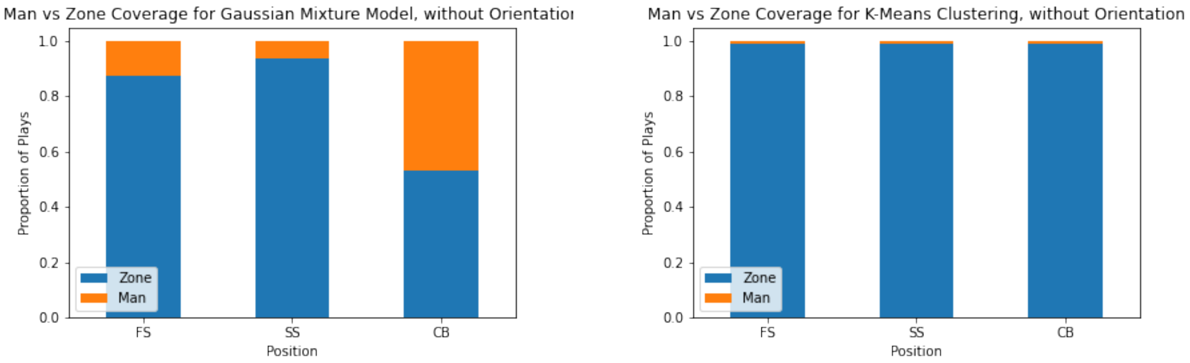

Figure 4: Man and Zone Coverage Percentage by Position for GMM and K-means, without Orientation

In our results, we can see that our Gaussian mixture model split for cornerbacks is roughly 50/50 man/zone, which is slightly different than 60/40 man/zone despite using the same feature space. Our model is classifying more cornerbacks in zone coverage than in man coverage. We could attribute this to many factors: exploring 17 weeks in our analysis instead of 6, the team-to-team and personnel variation over the course of a season, or simply that this data is for the 2018 season while theirs was the 2017 season. One driving factor we do acknowledge is that the percentage of man and zone coverage has been shown to vary wildly by team, as per *Pro Football Focus*' analysis of team usage of defensive coverage (Monson [12]). Thus, it is difficult to compare my implementation of the algorithm to previous work, but we can still determine how including orientation later in the paper is helpful. What is new to our analysis is that we examine these features for free safeties and strong safeties. We see this is largely dominated by zone coverage, as is to be expected by the nature of the position. It is slightly peculiar that strong safeties, who are generally larger and slower and used to support run defense, have a higher percentage of zone coverage, but this could also be attributed to different team tendencies and schemes.

The results for K-means are uninformative, as it appears the majority of points were assigned to the zone coverage cluster, which cannot be attributed to simple differences in team tendencies. Per James et al. [8], common disadvantages to K-means clustering include the idea that observations may belong to smaller subgroups in the data, and that we may simply be clustering noise with only $k = 2$ centers. We unfortunately do not have a good way of validating this. Additionally, it would be difficult to use $k > 2$ clusters in our analysis and manually confirm what each of the clusters would represent in our defensive coverage classification.

Next, we consider the results with orientation information included and compare that to the previous tables. These results are represented numerically in Tables 4 and 5, and visually in Figure 5.

In these results, we can see that our Gaussian mixture model split for our positional

Position	Zone Percentage	Man Percentage
FS	0.9893	0.0106
SS	0.9783	0.0216
CB	0.2220	0.7779

Table 4: Gaussian Mixture Model Results by Position, with Orientation

Position	Zone Percentage	Man Percentage
FS	0.9435	0.0564
SS	0.9692	0.0307
CB	0.1561	0.8438

Table 5: K-means Results by Position, with Orientation

coverage breakdown is slightly different with our orientation features. Our model classifies many more cornerbacks in man than in zone coverage, and more of the free safeties and strong safeties are classified as playing zone coverage than without the orientation features. Our feature creation of orientation relative to the opposing player is especially useful in this case, since we can associate the defender following the opposing player across the field and facing them as the play continues. The other relevant orientation feature is the orientation of the defender relative to the line of scrimmage. We would expect this to be more useful in identifying zone coverage, as the player tends to take more of a “wait and see” approach and covering a smaller area of the field, with their body facing the opposition as a whole.

Similar results are reflected in the K-means results as well. We are able to get somewhat informative labels with the added orientation features, though we may suffer from the same issues as we did without them. Compared to the Gaussian mixture model results, we are classifying both more cornerbacks and more safeties as playing man coverage. It appears we may be unable to distinguish observations classified as man coverage from noise or zone coverage despite our included features.

From this point onward, we only consider the results from our Gaussian mixture model with orientation features included. We produced other insights regarding the players who spend the most time in man coverage and zone coverage, which can be found in Section 7.6 in the appendix.

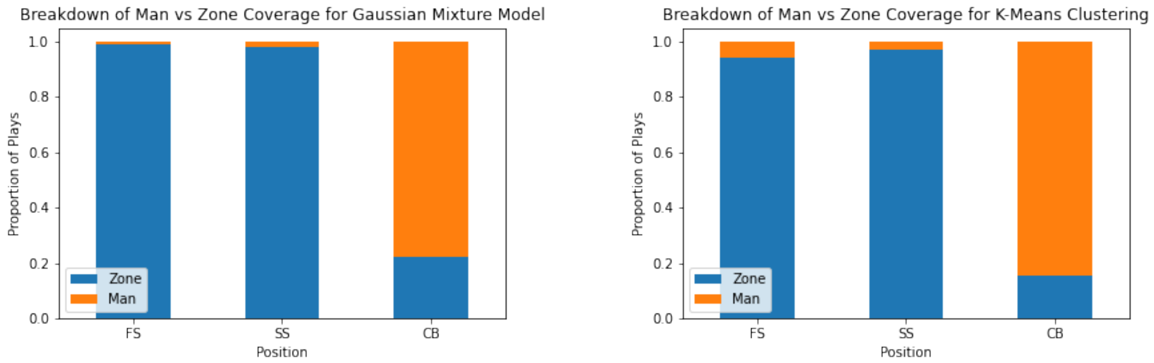

Figure 5: Man and Zone Coverage Percentage by Position for GMM and K-means, with Orientation

4 Quantifying Receiver Performance

In this exploratory analysis, we look to obtain some informative descriptive statistics using our labels created from the Gaussian mixture model. We depend on these estimates over the K-means estimates, since we have the probability of belonging to the assigned cluster which allows us to be more confident in the choice we made.

We only focus on the matchup between cornerback and wide receiver, since we determined that most corners are playing man-to-man coverage in Section 3. We first identify whether the cornerbacks we assigned a cluster label have their closest opponent as a wide receiver. This would indicate that they are covering a wide receiver on the play, and thus have some degree of influence on the play outcome. We only are interested in the corresponding offensive play outcome, so we narrow down our dataset to only receivers who are targeted on the play. This is determined by the location of the football through the play and with the play description labels provided. We then look at the specific route that the targeted receiver runs on the play and whether they catch the pass or not. Our goal is to identify if there is an association between two outcomes: players who maximize yardage gained against man coverage and zone coverage and which routes maximize yardage gained against man coverage and zone coverage.

We looked at 8,694 unique plays that consisted of cornerbacks directly matched up against wide receivers who were targeted on the play. Of these plays, 6,673 matchups were man coverage and 2,021 were zone coverage. In the figures below, we examine different measures of success to see how coverages may influence play outcomes.

We quantify play outcomes in two ways: with yards gained on the play and expected points added (EPA). EPA is a standardized descriptive statistic developed to measure the importance of yardage gained in context of the game situation. The argument for using EPA is that not all yards are created equal: an 11 yard pass on 4th and 20 is much less useful than 3 yards on 4th and 2. EPA attempts to adjust for play valuations. Further reading on EPA can be found in Yurko [19]. Measures of EPA on each pass play were in the dataset provided via the Big Data Bowl. We use yardage gained as our measure to provide a baseline comparison to using EPA. Inherently, this will miss out on information about game situation, such as down and distance and the intended goal of the offense. The idea is to provide an overview of which routes appear to be more effective and which route combinations can provide an advantage to the offense, under assumption that the play caller and quarterback can correctly identify the type of coverage.

A full glossary of receiver routes and their definitions can be found in Table 8 in Section 7.5. We first look at the breakdown of routes for man and zone coverage, quantified by both yardage and EPA. The results are provided in Figures 6 and 7.

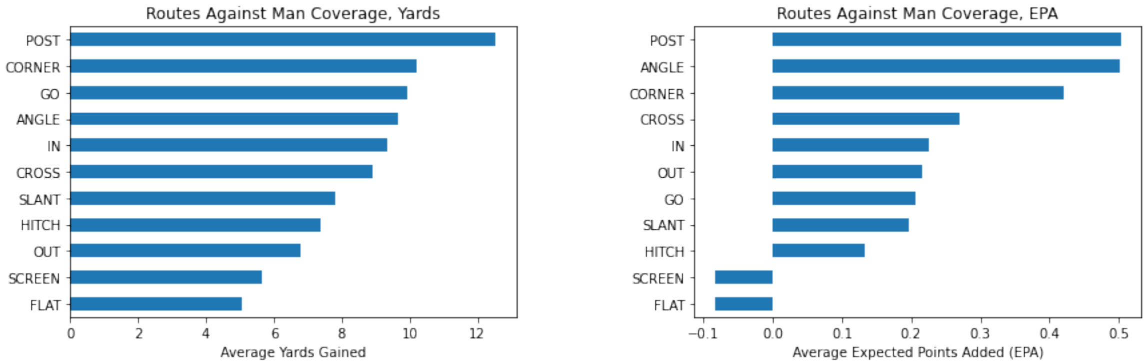

Figure 6: Play Outcomes by Routes vs Man Coverage

For routes against man coverage, both measures of yardage gained and EPA agree that post routes lead to the highest average gain per play, with corner routes and angle routes coming in close second. Each of these routes all consist of the player running in the same direction for a period of time, then sharply changing direction. For a defensive player in man coverage against this player, it becomes incredibly difficult to react quickly to these sharp movements, leading to larger expected gains when these routes are performed successfully. These perform less optimally versus a zone coverage defender, since there is often help from other defenders to account for the changes in direction by the receiver.

On the other hand, by both measures, screens and flat routes perform poorly against man coverage, even to the extent of having a negative average EPA. Both of these routes are short and originate behind the line of scrimmage. As a defender, if the assignment is to cover an offensive player one-to-one and they are standing behind the line of scrimmage, it is often straightforward to tackle them and minimize the offensive gain. The success of screen plays and flat routes are largely dependent on the execution of other blockers on the play. It follows that these are poor route choices in a man coverage situation.

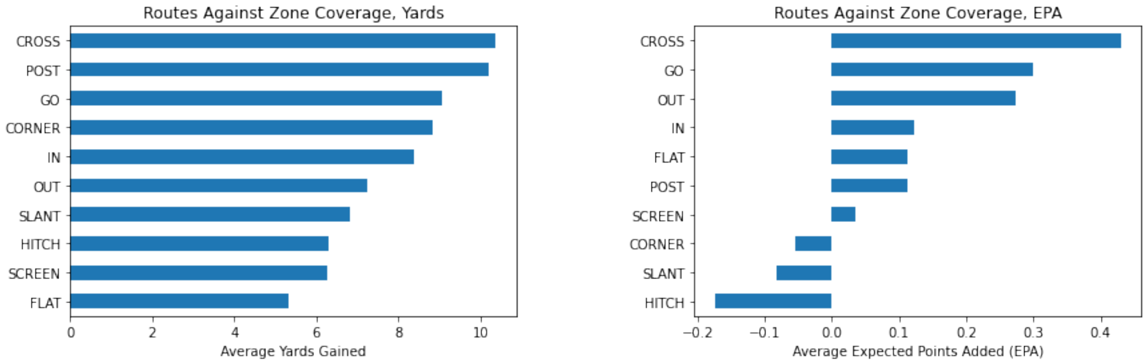

Figure 7: Play Outcomes by Routes vs Zone Coverage

When looking into routes against zone coverage, we see that crossing routes perform best by both variables. Crossing routes will break down the responsibilities of a zone coverage defender, since this type of route involves traveling from one side of the field to the other in its entirety, forcing the defense to communicate about who is responsible for covering the offensive player. In a man coverage scheme, it is easier to defend against crossing routes since it does not involve many sharp changes of direction and the assigned defender can follow the receiver across the field. This is reflected in the middling performance of crossing routes versus man coverage in Figure 6. Another common top route against zone coverage is the go route, which involves the receiver running in a straight line towards the end zone at top speed. This can be particularly effective since it can create mismatches between the defensive assignments. The cornerback facing the receiver at the line would be unable to react and keep up with the receiver, so the responsibility is passed on to the deep zone defender, often the safety, who may not be as fast or reactionary as a cornerback by nature of the position. This leads to go routes breaking down zone defenses more effectively than other routes. Again, it would come down to communication between the defense to effectively neutralize these routes.

Zone coverage seems to perform particularly poorly against hitch routes and slant routes, according to EPA. Hitch routes consist of a player running a certain distance forward from the line of scrimmage and abruptly stopping and waiting for the pass. Given how short the route is, there are often cornerbacks lurking underneath these receivers and waiting to intercept a pass. Since it is in a zone defense, the quarterback may not anticipate the defender waiting for the pass and be prone to making costly mistakes. Other less optimal routes against zone coverage include slant routes and corner routes, likely for similar reasons with defenders waiting underneath these routes in zone coverage.

There is an unseen component to using these measures to quantify receiver performance. Often in the name of strategizing, teams will use these less-than-optimal routes to set up misdirections that will pay off later in the game on a different play. Consider this hypothetical: a team runs a screen pass on a play for a minimal gain of 3 yards in the first quarter against man coverage. This is considered a low yardage gain and negative EPA by the chart in Figures 6. Later in the game, the team runs a play out of the same formation against man coverage, with the screen receiver in position ready to receive the pass. However, the quarterback fakes the pass to the screen receiver, the defender moves up in anticipation to stay on their man-to-man assignment, which would instead free up a receiver running a different route behind the defender. This type of game theory is not addressed directly in the model, and is an important consideration to make when assessing the results in terms of maximizing yardage and EPA. While it is true that EPA is built off of historical NFL play-by-play data, there still needs to be some type of adjustment to account for plays that set up larger gains on "highlight" plays. This is something that can be looked into by different situational measures like down and distance, distanced gained relative to distance needed, or distance gained in crucial game-time scenarios (i.e. in a two-minute drill at the end of a half, or when a team is on the verge of taking the lead, or when the field is compressed due to being within five yards of scoring a touchdown).

Generally, what we can take away from these results is that routes with sharp directional movements succeed against man coverage but fail against zone coverage, while routes that require timing and communication succeed against zone defenses but are more difficult against man-to-man defenses. Further analysis work was done with the receiver versus cornerback matchups, specifically regarding the best individual receivers against both types of coverage. Those results are provided in Section 7.6 to prevent clutter.

5 Discussion and Future Work

The main goal of this project was to expand upon the paper written by Dutta et al. [3] by including suggested new features and positional groups, examining an alternative clustering method, and undergoing exploratory analysis about offensive performance against these coverages.

We were able to see how our implementation of the Gaussian mixture model changed when we included new features compared to their analysis. It was ultimately found that including the features about orientation led us to classify more cornerbacks in man coverage than without it. We were also able to narrow down nuances between zone and man coverage for the safety position. While the majority of both free safeties and strong safeties play in zone coverage, there are a select few players who have a tendency to follow an offensive player through the duration of the play. These players are listed in Figures 11 and 13. With updated weekly data, we would be able to examine these trends through the course of a season and gain insight on individual player tendencies. There is potential to save time for coaches by automatically informing them of key matchups they can exploit on the offensive side, as well as identifying defensive weaknesses of their own team.

We also looked at one way we could use results about defense to inform choices that can be made on offense. By having a concept of which routes are effective against specific coverages, and which players tend to play in those coverages, coaches can effectively design plays that exploit these weaknesses. While not listed in this paper, it would be straightforward to break down offensive personnel by team and which routes they run the most, and which coverages they face the most. Having access to these insights can inform choices to make in player acquisition by identifying offensive weaknesses and seeking to upgrade. The applications of having a model to classify defensive coverage spans beyond only looking at offensive play outcomes or individual defensive player tendencies.

There are many opportunities for future work based on this framework. While it is a step forward to further refine classification for cornerbacks by including orientation and confirm what we knew about the safety position, there are many other positions of interest that could be examined. In this analysis, we chose to forgo analysis for the linebacker position, since their responsibilies include run defense and pass defense. Perhaps with this tracking data and our feature set, we could develop new features to summarize movement of linebackers and whether they are playing in run defense support or pass defense support. This would entail development of new features for this specific annotation problem, such as distance from the opposing offensive linemen, running back, and wide receiver at each frame. This would not be an easy or computationally efficient task, but it would prove useful for automatic annotation under this framework.

In a similar vein, in our offensive play outcome analysis we did not consider routes run by running backs or tight ends. We only looked at cornerbacks covering the opposing wide receiver. We know that running backs and tight ends tend to be covered by linebackers, since they are generally larger and a size mismatch for many cornerbacks. It would be interesting to see if there is a difference in route effectiveness against these coverages by position. It is not unreasonable to think that routes requiring quick changes of direction would be more difficult for larger, slower tight ends instead of quick-footed wide receivers.

One final improvement that could be looked into is how the coverage assignment for one player influences the coverage assignment for another player. For example, if a cornerback is playing man coverage and they know they have safety help over the top in a zone coverage scheme, they may be allowed to maintain some distance from their assigned offensive player and take more risks, perhaps gamble on an interception. We are unsure if the model would pick up on this association, and it could lead to an incorrect classification of zone coverage for the cornerback since they are not close to their assignment step-for-step through the duration of the play. Having information about another player's coverage label could provide better annotation in a case like this. We can also then determine if there are tendencies to exploit from a coaching perspective. With this information, we would be able to see combinations of coverage between different positions and determine team tendencies for playing in pure zone defense, pure man defense, or a hybrid between the two. Recognizing the situations in which teams will play certain coverages would provide an advantage in game planning and organization, leading to better team performance.

Building off of that, we can also associate how different personnel groupings give away tendencies about how individuals play man or zone coverage. To briefly introduce some terminology, teams have different packages of personnel groupings. Some common personnel groupings include "nickel", "dime", or "quarter". A nickel package means a team includes a fifth defensive back in lieu of a player at another position. A dime package means a team includes a fifth and sixth defensive back on the field, substituting out other positions. A quarter package means that teams will have seven defensive backs on a given play. Understanding how these different personnel groupings are utilized situationally can provide insight on team tendencies and reveal more areas where teams can exploit matchups.

Further research can improve on the methods and coverage labels assigned to these players. Another important note is that the dataset only consists of passing plays from the 2018 NFL season. If we had tracking information on run plays, we could attempt to label clusters of man and zone coverage based on movement before the snap and when the ball is in play before the handoff to the running back. We may see vast differences in coverage assignments for those events, since it is before the defense would converge on the running back, so the cornerbacks and safeties would still be focused on their assignments. This would provide more tendency information for coaches to build a game plan from, and provide more informative results of how coverages may shift through the duration of the play. We would be able to break down the coverage by the type of play (run or pass) that was run and determine play outcomes based on the pre-snap coverage. It would be even more helpful to have data over the course of multiple seasons, where we can pinpoint how personnel changes (from player additions, retirements, losses in free agency, etc.) would change a team's defensive strategy.

In this essay, the main benefit was to look into expanded analysis for the cornerback and safety position groups, briefly examine the advantages and disadvantages of a soft and hard clustering algorithm, and examine how offensive play outcomes differ based on coverage. There exists many opportunities to expand on this dataset and look into more detailed analysis from this framework, where new features can be designed and implemented to answer many different annotation problems.

6 References

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, *49*(3), 803–821. Retrieved May 2, 2022, from http://www.jstor.org/stable/2532201

Bowen, M. (2017). NFL 101: Breaking down the basics of the route tree. https://bleacherreport.com/articles/2016841-nfl-101-breaking-down-the-basics-of-the-route-tree

Dutta, R., Yurko, R., & Ventura, S. (2019). Unsupervised Methods for Identifying Pass Coverage Among Defensive Backs with NFL Player Tracking Data. https://doi.org/10.48550/ARXIV.1906.11373

Fraley, C., Raftery, A., Murphy, T., & Scrucca, L. (2012). Mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report No. 597*.

Hartigan, J. A. (1975). *Clustering algorithms* (99th). John Wiley & Sons, Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning: Data mining, inference, and prediction*. Springer New York. https://books.google.com/books?id=yPfZBwAAQBAJ

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. Springer New York. https://books.google.com/books?id=qcI%5C_AAAAQBAJ

Jin, X., & Han, J. (2010). K-means clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 563–564). Springer US. https://doi.org/10.1007/978-0-387-30164-8_425

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

McNicholas, P. D. (2016). Model-based clustering - journal of classification. https://link.springer.com/article/10.1007/s00357-016-9211-9

Monson, S. (2017). Taking a closer look: Examining the NFL's coverage scheme tendencies. https://www.pff.com/news/pro-taking-a-closer-look-examining-the-nfls-coverage-scheme-tendencies

NFL Big Data Bowl. (2021). https://www.kaggle.com/c/nfl-big-data-bowl-2021/

NFL Big Data Bowl - Plotting Player Position. (2020). https://www.kaggle.com/code/robikscube/nfl-big-data-bowl-plotting-player-position/notebook

NFL Glossary of Terms. (2022). https://operations.nfl.com/learn-the-game/nfl-basics/terms-glossary/

NFL Next Gen Stats. (2022). https://operations.nfl.com/gameday/technology/nfl-next-gen-stats/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Reynolds, D. (2009). Gaussian mixture models. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of biometrics* (pp. 659–663). Springer US. https://doi.org/10.1007/978-0-387-73003-5_196

Yurko, R. (2017). NFL Player Evaluation Using Expected Points Added - CMU Statistics. https://stat.cmu.edu/~ryurko/files/greatlakes_2017.pdf

7 Appendix

7.1 Reproduction

The data and code used for the analysis, as well as the resources for the essay, can be found at: https://github.com/taarakshah/ms-thesis.

7.2 Downloading NFL Tracking Data

The data used in this analysis be downloaded from the Kaggle website here: https://www.kaggle.com/c/nfl-big-data-bowl-2021/data. The data will be downloaded in a zipped file of roughly 2.33 GB. There are 20 .csv files, 17 of which represent tracking data for the 17 weeks of the NFL season, and 3 of which provide supplementary information and keys to join on about individual games, players, and plays.

7.3 Glossary of NFL Terms

Some definitions were generously provided by the NFL Football Operations [15] division.

Term	Definition
Cornerback (CB)	A defensive player who focuses on guarding the opposing offense's wide receiver. Primarily serves to prevent pass completions.
Defensive Lineman (DL)	A defensive player who focuses on disrupting the quarterback's processing and attempts to tackle him for a loss of yardage.
Extra Point	A scoring play that results in 1 point for the offensive team. Can only occur after a successful touchdown.

Field Goal	A scoring play that results in 3 points for the offensive team. The ball must be kicked through the field goalposts for a successful try.
Line of scrimmage	The imaginary line marking the beginning yard line of a specific play. The ball is positioned here upon the start of each play.
Linebacker (LB)	A defensive player who is responsible for run defense and pass defense.
Offensive Lineman (OL)	An offensive player who focuses on protecting the quarterback from oncoming defenders.
Pass play	A type of play initiated when the quarterback receives the ball and throws forward to an eligible receiver, typically a tight end, running back, or wide receiver.
Play	A moment in the game where the ball is in action. Can be either a run play or pass play.
Quarterback (QB)	An offensive player responsible for initiating the play, throwing to a receiver or handing the ball off to a running back, and completing passes.
Run play	A type of play initiated when the quarterback receives the ball and hands it to a player in the backfield so that player can advance the ball, typically a running back.
Running back (RB)	An offensive player responsible for running the ball forward. Typically lines up behind or next to the quarterback.

Safety (S)	A defensive player who is responsible for covering zones to prevent passes being completed down the field. There are two versions of this position, known as “strong safety” (SS) or “free safety” (FS). Typically, strong safeties assist linebackers in run support, while free safeties are assisting cornerbacks in pass coverage.
Tight End (TE)	An offensive player that functions similar to a wide receiver, but is generally larger and assists in
Touchdown	A scoring play that results in 6 points for the offensive team. Can come in the form of a passing play, rushing play, or special teams play.
Wide Receiver (WR)	An offensive player responsible for running in predetermined patterns (routes) and catch the ball on passing plays from the quarterback.

Table 6: Glossary

7.4 Features Provided in Tracking Data

Predictor	Description
x	Player position along the long axis of the field, 0 - 120 yards (numeric)
y	Player position along the short axis of the field, 0 - 53.3 yards. (numeric)
s	Speed in yards/second (numeric)
a	Acceleration in yards/second^2 (numeric)
dis	Distance traveled from prior time point, in yards (numeric)

o	Player orientation (deg), 0 - 360 degrees (numeric)
dir	Angle of player motion (deg), 0 - 360 degrees (numeric)
event	Tagged play details, including moment of ball snap, pass release, pass catch, tackle, etc (text)
nflId	Player identification number, unique across players (numeric)
displayName	Player name (text)
jerseyNumber	Jersey number of player (numeric)
position	Player position group (text)
team	Team (away or home) of corresponding player (text)
frameId	Frame identifier for each play, starting at 1 (numeric)
gameId	Game identifier, unique (numeric)
playId	Play identifier, not unique across games (numeric)
playDirection	Direction that the offense is moving (text, left or right)
route	Route ran by offensive player. Unique values: hitch, out, flat, cross, go, slant, screen, corner, in, angle, post, wheel

Table 7: Features available in NFL tracking data

7.5 Glossary of Receiver Routes

Below is a table summarizing the names of routes run most frequently by wide receivers. The ”route tree” in Figure 8 provides a visualization of common routes run by wide receivers. This was provided courtesy of Bowen [2]. The table below does not include routes commonly run by running backs, since the analysis did not focus on them, nor are they represented in the route tree visualization.

Route Name	Description
Angle	A route that initially begins with the receiver moving toward the sideline then cutting sharply back toward the middle of the field.
Corner	A deeper route where the receiver travels in a straight line toward the end zone, then cuts sharply outward toward the sideline.
Cross	A route that begins on one side of the field and moves across the field. Less sharp cut than a slant route.
Flat	A route typically ran by running backs or tight ends where they cut from outside of the backfield and hover close to the sideline by the line of scrimmage.
Go	A straight route where the receiver sprints straight toward the end zone.
Hitch	A short curl route with an short depth of target, approximately 2-3 yards from the line of scrimmage.
In	A general term for a route that cuts away from the sideline toward the middle of the field at a 90 degree angle after a specific distance from the line of scrimmage is reached.
Out	A general term for a route that sharply cuts toward the sideline at a 90 degree angle after a specific distance from the line of scrimmage is reached.
Post	A deeper route where the receiver travels in a straight line toward the end zone, then cuts sharply inward away from the sideline.

Screen	A route where the receiver stays behind the line of scrimmage with intent to receive the ball and travel up the field with it. Often includes other offensive players as blockers.
Slant	A route that cuts sharp and diagonally to the middle of the field after a specific distance from the line of scrimmage is reached.

Table 8: Description of routes run by receivers

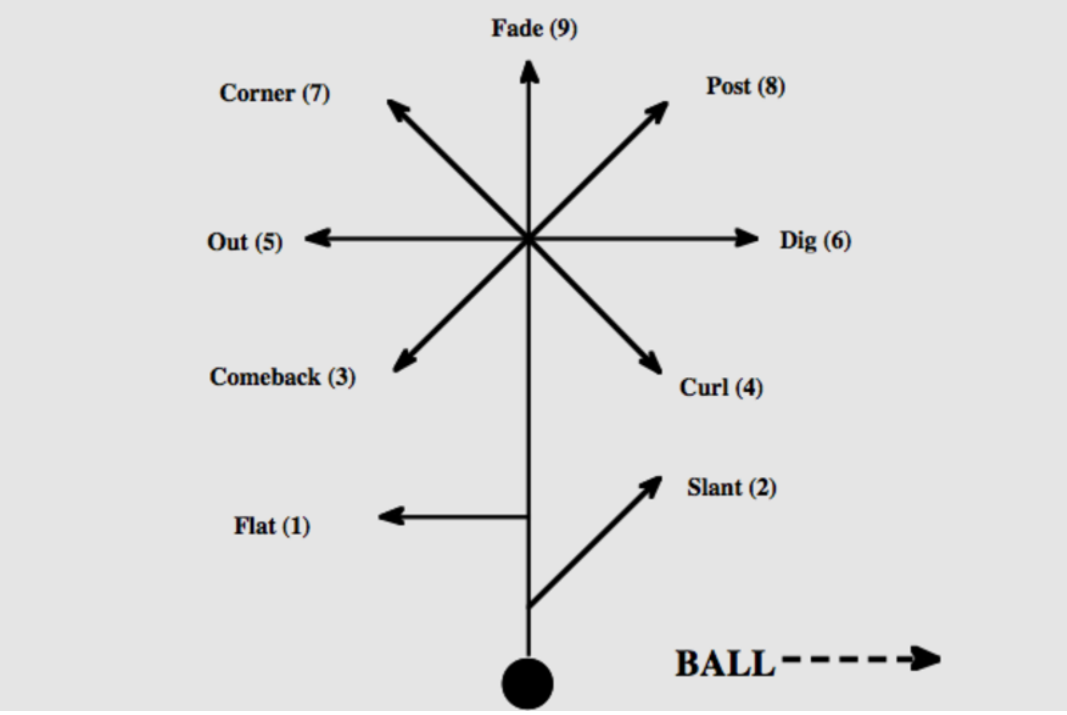

Figure 8: Visual of NFL route tree [2]

7.6 Additional Figures

The figures in this section represent additional interesting slices of data from our results. Figures 9 through 14 show the players with the highest percentage of man coverage and zone coverage, sorted by position (minimum 100 plays required). The margins are noticeably thin for the safety position, as the majority of players in this position tend to play zone coverage.

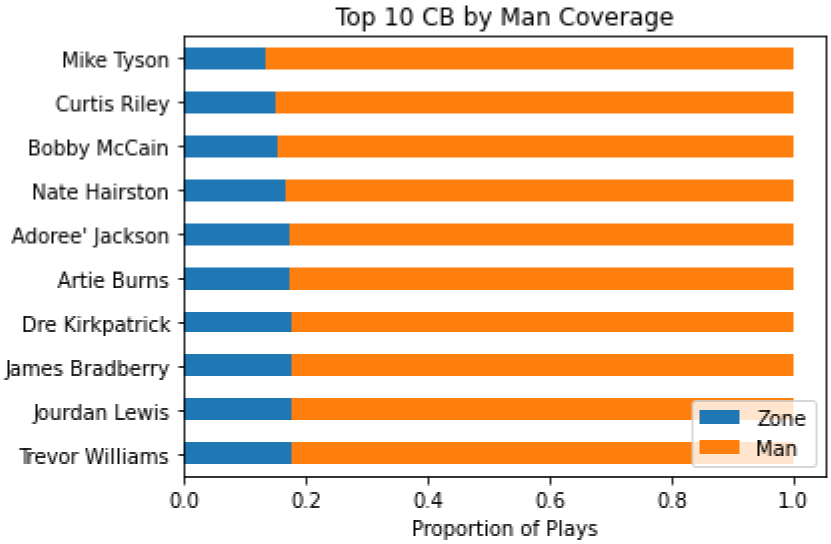

Figure 9: Top Man Coverage Cornerbacks, by Percentage

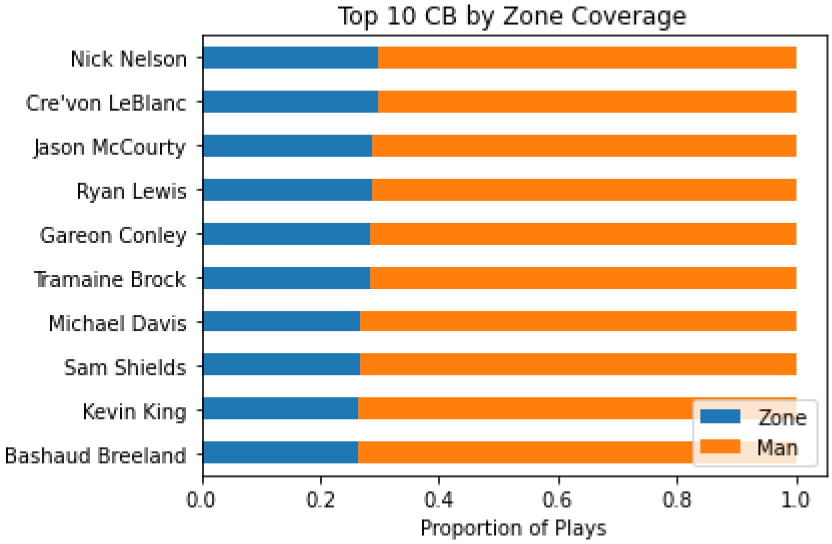

Figure 10: Top Zone Coverage Cornerbacks, by Percentage

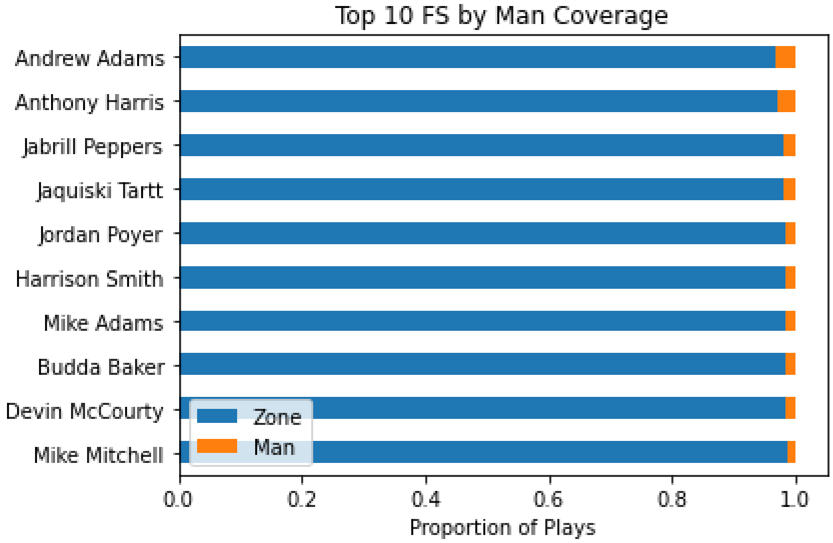

Figure 11: Top Man Coverage Free Safeties, by Percentage

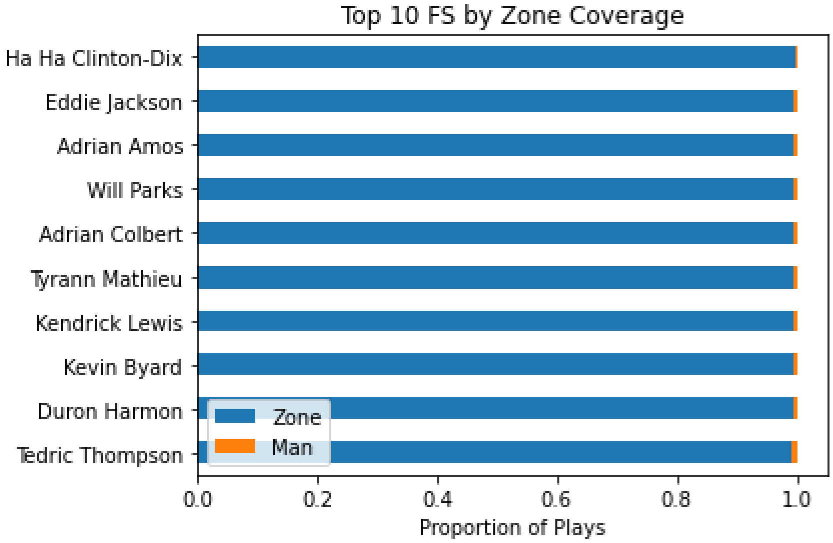

Figure 12: Top Zone Coverage Free Safeties, by Percentage

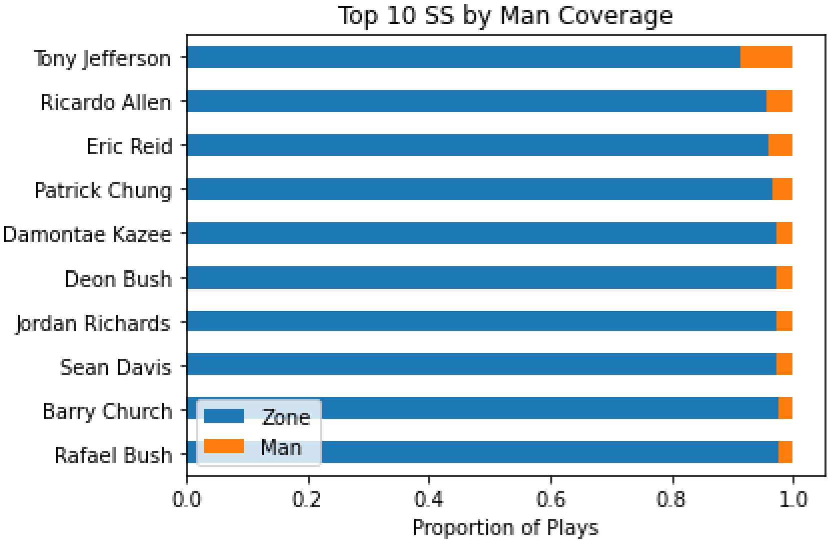

Figure 13: Top Man Coverage Strong Safeties, by Percentage

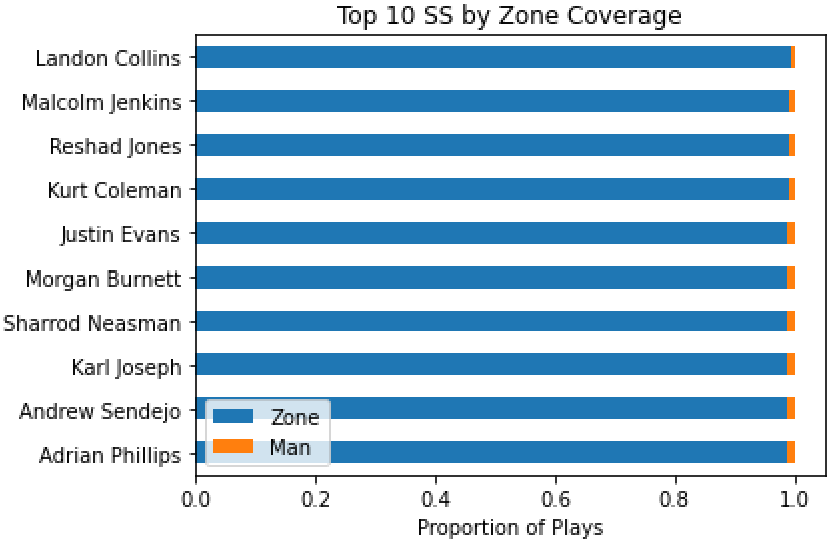

Figure 14: Top Zone Coverage Strong Safeties, by Percentage

Figures 15 through 18 show the top 20 receivers against man coverage and zone coverage, quantified by both yardage gained and EPA.

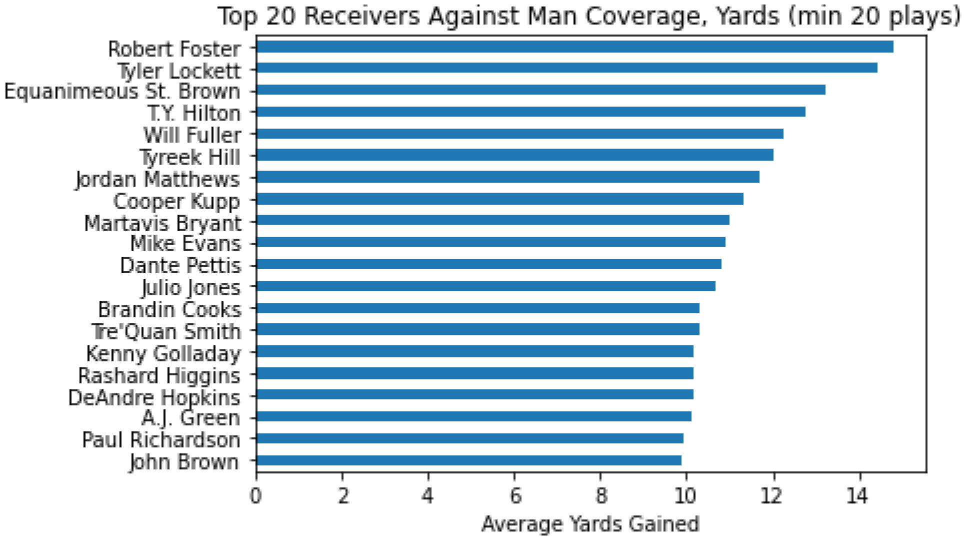

Figure 15: Top WRs vs Man Coverage, by Yards Gained

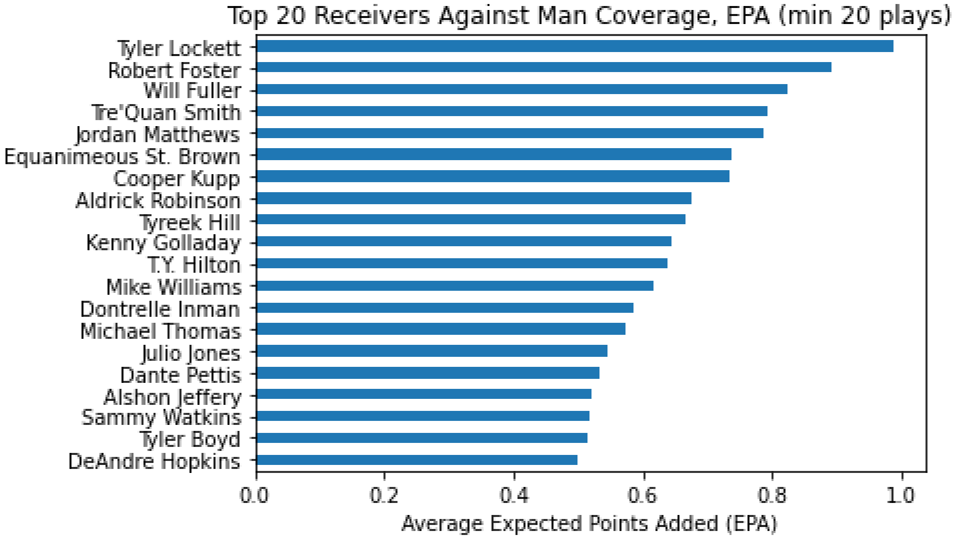

Figure 16: Top WRs vs Man Coverage, by EPA

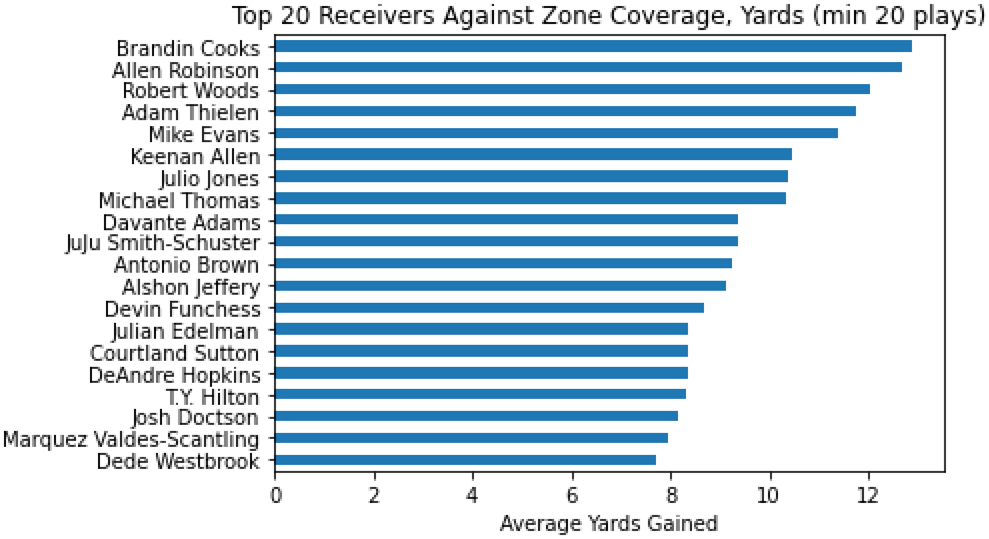

Figure 17: Top WRs vs Zone Coverage, by Yards Gained

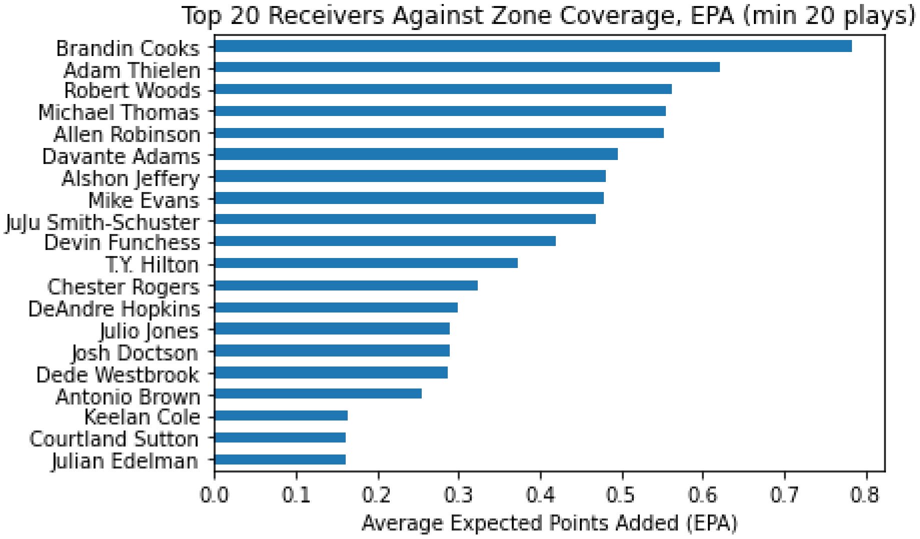

Figure 18: Top WRs vs Zone Coverage, by EPA