Estimating the Causal Effect of Defensive Formation on Yards Gained in Run Plays

2020 NFL Big Data Bowl

Aaron Kruchten

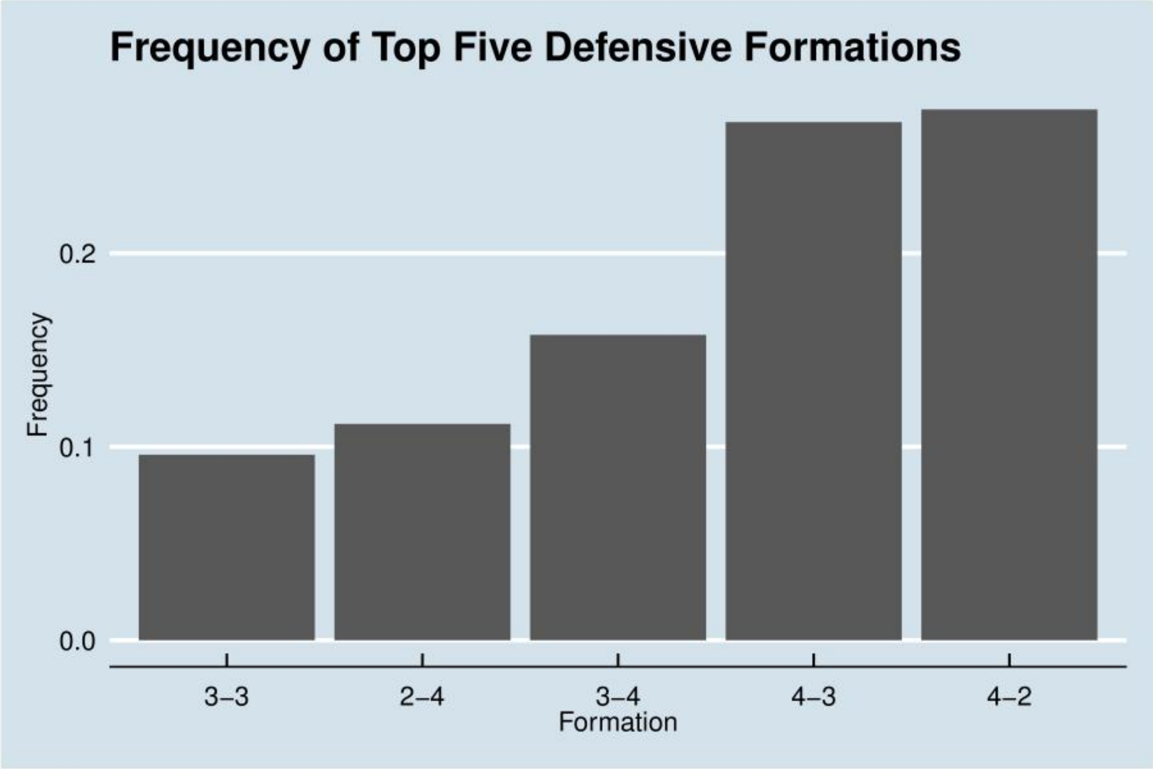
Introduction

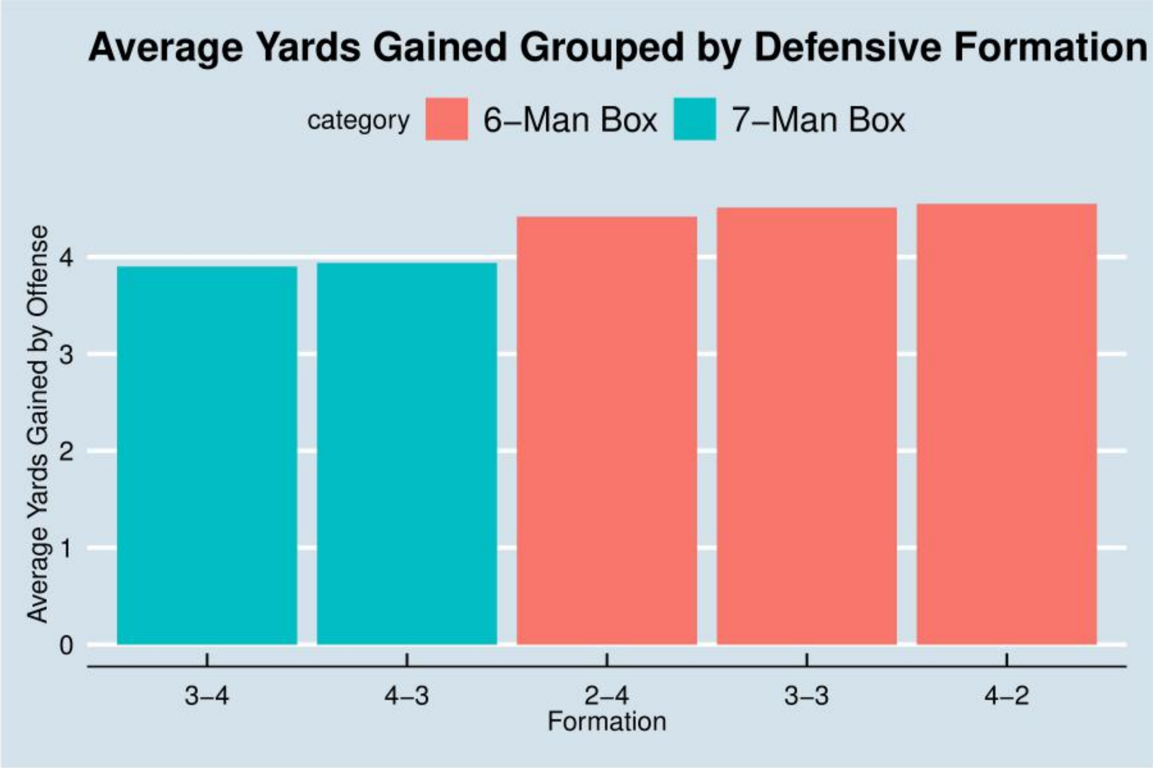
There is much debate about which defensive formations are the most effective. When choosing a formation for the defense, many factors must be taken into account. These factors include a formations's efficacy for pass protection, pass rush, and ability to stop the run. Understanding which formations are best for these three factors would be very valuable to NFL teams and defensive coordinators. Therefore, this analysis will focus on analyzing which formations are best at stopping the run game, and will estimate the causal effect of defensive formation on yards gained.

In order to estimate this effect, the best approach would likely be to conduct a randomized experiment with a within-subjects design by prescribing each formation to a set of NFL teams for a certain number of plays and recording yards gained. However, due to the competitive nature of football this approach is simply not feasible. Therefore, we must use approaches with observational data. In order to estimate the effect of defensive formation on yards gained, we will first conduct exploratory data analysis in order to identify the five most common defensive formations. Then, utilizing our knowledge of football, we will construct a directed acyclic graph (DAG) (sometimes also called a Bayesian Network), which expresses our beliefs of the factors which affect yards gained. Utilizing this graph, we will apply the *Back-Door Criterion* to identify a set of controls sufficient for controlling for confounding effects. Last, using this set of controls, we will build a predictive model to allow us to estimate the casual effect that defensive formation has on yards gained.

Exploratory Data Analysis

In this analysis, we will use the data set provided for the 2019 run prediction Kaggle contest. The data set has approximately 500k rows and contains information about all of the 22 players on the field at the time of hand off with a row for each player. After some data cleaning, and reducing the data to only one play per row, we are left with just under 23,000 rows of data. Our first step is to understand which defensive formations teams most often run. The graphs below display the five most common defensive formations and the mean yards gained by the offense for these formations.



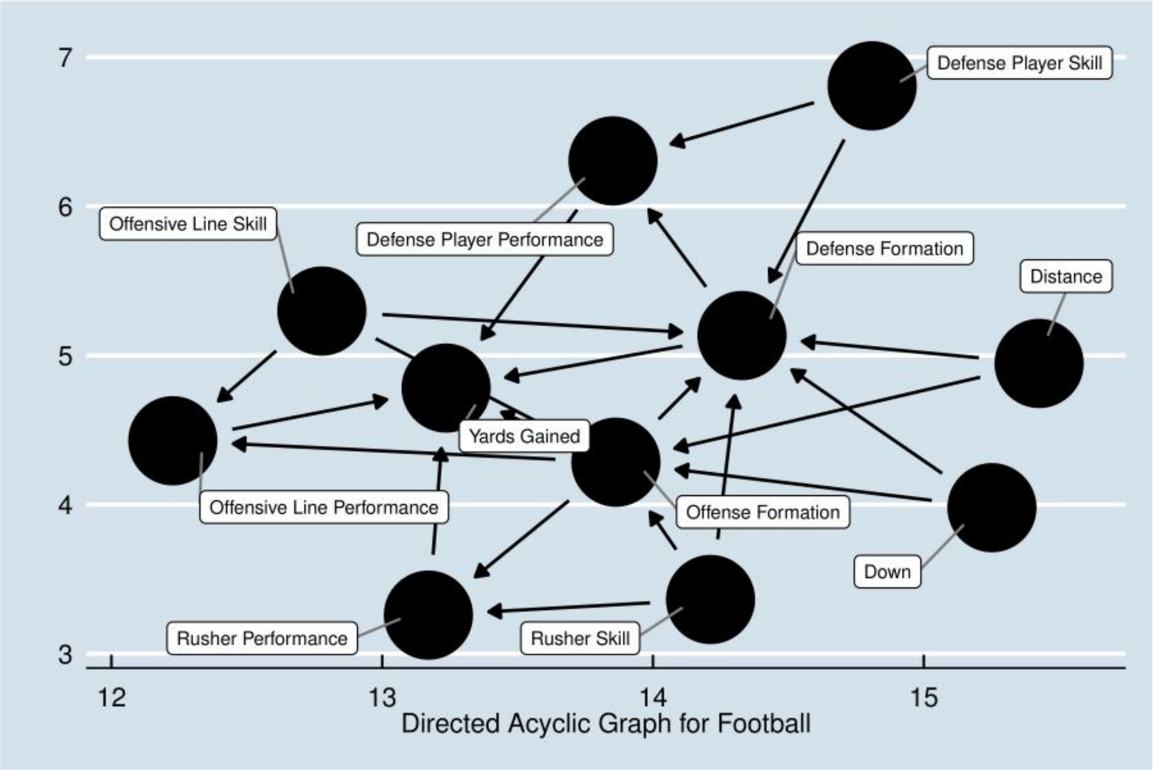


One will notice a few things about the above plots. First, the 4-2 defensive formation is both the most common defensive formation, and the formation that allows the most yards gained. Second, formations with a 7-man box appear to allow significantly fewer yards gained than formations with a 6-man box. Third, the 2-4 formation, although being the second least common formation among these five, appears to be the most effective at stopping the run among the formations with 6-men in the box.

Although this initial analysis appears to suggest some results, such as that the 2-4 formation is better at stopping the run than the 4-2 formation, we must be careful before reaching these conclusions yet. This result could be due confounding effects. For example, it might be that teams with better linebackers tend to run the 2-4 formation. This implies that the decrease in yardage is not due to the formation, but may be due to the skill of the players. The rest of this paper will be us developing a methodology to control for these possible confounding effects.

Graphical Causal Model

In this analysis, we would like to understand specifically how changing formation affects yards gained on a run play. That is, we would like to understand the causal effect that changing formation has on mean yards gained. In order to adequately measure this effect, we must control for all variables that affect both defense formation and yards gained. The following graphic is our estimate of the factors which affect yards gained. One can interpret the graphic as an arrow pointing directly from one variable to another implies that that variable causes the variable it is pointing to.



Although there exists methods to automatically discover causal structures such as the one displayed above (Glymour, C. et. al, 2019), we choose to form this model based solely on our knowledge of football. The first main assumption we make in forming this graph is that yards gained is affected by nothing except the players on the field. That is, yards gained is not affected by factors such as wind speed, stadium, or field type. Although this is not known to be true it seems reasonable. Additionally, at the time of writing this analysis, mid December 2019, the winning model for the 2019 NFL Big Data Bowl Kaggle Competition only used player location and speed features in order to predict yards gained (Gordeev, D., & Singer, P. 2019).

Besides this previous assumption, the graph above is mostly based on the following premises.

- Yards Gained is affected by the formation of the offense and defense and the performance of each player on the field.
- The formations coaches decide to line up in depend on the skills, strengths, and weaknesses of their players as well as the state of the game such as down and distance.
- The performance of each player depends on the formation of the offense and defense and each player's skill.

While we believe that most people knowledgeable of football will agree with the above premises, one part of our graph one may take issue with is the relationship we define between offensive and defensive formation. In our graph above, we have a directed edge from offensive formation to defensive formation. This implies that offensive formation causes defensive formation. In other words, the defense chooses their formation depending on what the offense does, however the offense does not change their formation depending on what the defense does. Although this is likely not completely true, we believe that this to be the case more often than not, and that the offense decides what formation they want to line up in and then the defense lines up their formation depending on what the offense does. While it is likely more true that both defensive formation and offensive formation affect each other, this would make estimating the causal effect of defensive formation difficult.

Additionally, it is important to note that we are defining a difference between player skill and player performance. We define player performance as how well a player performed his duties in a given play. We define player skill has a player's inherent athleticism and ability to perform his duties.

As stated before, we are primarily concerned with the relationship between yards gained and defense formation. The node which represents yards gained can be found at approximately the coordinate (13.25,4.7) and defense formation can be found at approximately (14.25,5.2).[1]

In order to identify a set of controls for estimating the causal effect as we wish to, we will use the *Back-Door Criterion*. The *Back-Door Criterion* gives us a methodology to identify the variables which will be sufficient to estimate the casual effect of a predictor on a result variable. A formal definition of the *Back-Door Criterion* can be found in the appendix of this paper.

Applying the *Back-Door Criterion* to the above graph, gives us two possible sets of controls listed below.

- {Offense Formation, Offensive Line Skill, Rusher Skill, Defense Player Skill}
- {Offense Formation, Offensive Line Performance, Rusher Performance, Defense Player Skill }

Because of the difficulty in adequately measuring the skill of an offensive lineman (Benjamin Alamar & Keith Goldner 2011), and the player tracking data we are using in this analysis which we believe will allow us to measure the performance of an offensive lineman on any given play, we choose to include the second set of controls in our analysis.

Therefore, in simple terms, assuming the above plot is the true causal plot and all relationships in the plot are in reality how they are displayed and there are no other variables which may affect the

[1] Note that the coordinates do not have any meaning and are simply an artifact of the software used to generate this plot. They are only left in in order to better locate relevant nodes.

variables in our plot, then we can estimate the causal effect of formation on yards gained.

Now that we have a set of controls we must decide how we will model these variables. The first variable we will discuss is offensive line performance.

Modeling Offensive Line Performance

In order to effectively model offensive line performance we choose to model the two main goals we believe any offensive lineman will have on any given run play. We believe that by modeling an offensive lineman's ability to accomplish the following goals we can effectively model their performance. The list below details what we believe to be these two goals, and how we chose to model them.

- The offensive line wants to win the line of scrimmage by driving the defensive line backwards off the ball, and get up to the second level and cover up the linebackers. In order to model this, we simply compute the distance between the offensive lineman's x position and the line of scrimmage at the time of hand off.
- They want to open a horizontal gap in the line of scrimmage for the runner to travel through. In order to represent this, we first go through the data and select which offensive linemen are likely to be playing each position, center, left or right guard, and left or right tackle depending on their y positions. We then compute the distance between each of the adjacent offensive lineman, and the distance from the out of bounds line and each tackle.

Modeling Rusher Performance

A rusher's performance in a given play is the sum of many variables. Some of these include a rusher's performance in breaking tackles, evading defenders, and finding open holes and room to run. Because we only have data about the players at the time of hand off, we can only model a rusher's ability to find and hit a gap in the offensive line. We do this by utilizing the rusher's speed, acceleration, direction, and the size of the gap between the two offensive lineman in the direction the rusher is heading. If a rusher is performing well at time of hand off they should have high acceleration, high speed, and should be running towards an open gap in the offensive line.

Modeling Defensive Player Skill

Now, we must consider how to model defensive player skill. Because corners and safeties are primarily concerned with guarding against the pass, and usually only stop run plays if the rusher has already passed the defensive line and linebackers, we will only model the skill of defensive lineman and linebackers. There are several possible approaches to how we could model defensive player skill.

- We could model a defensive player's skill against the run by considering statistics such as the ratio of tackles they have made over how many plays they have been in. However, this approach has several issues. First, is this approach is difficult with current data. Although data is accessible that would allow us to compute this ratio for every player (Maksim Horowitz et. al, 2019), there is no easy way to match players in these data sets with players in the data set we are using in this analysis. Second, by using tackles as a metric we may also be modeling *player performance* instead of *player skill.* Doing this would violate the conditions for the *back-door criterion* and we would no longer be estimating the effect of defensive formation on yards gained.

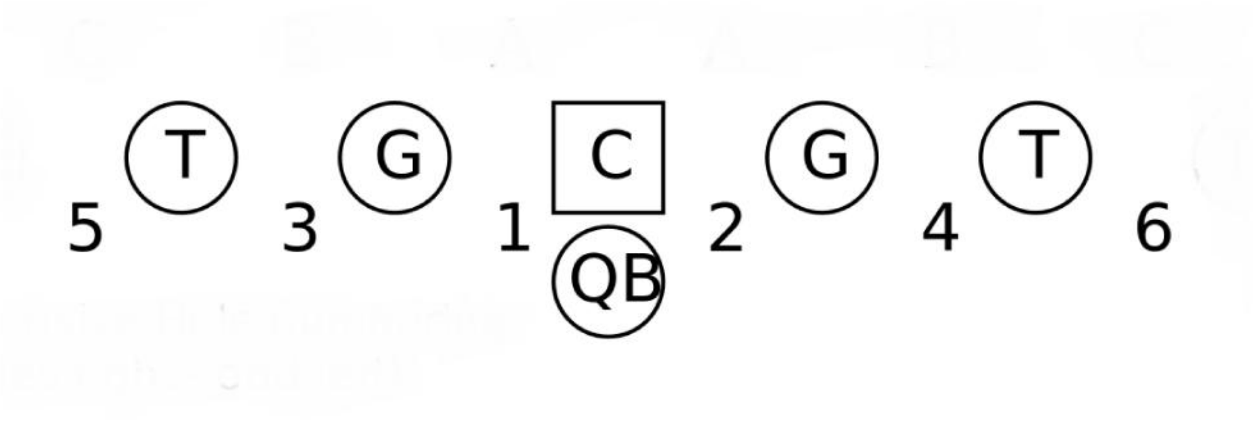


Figure 1: Naming convention for gaps in this analysis

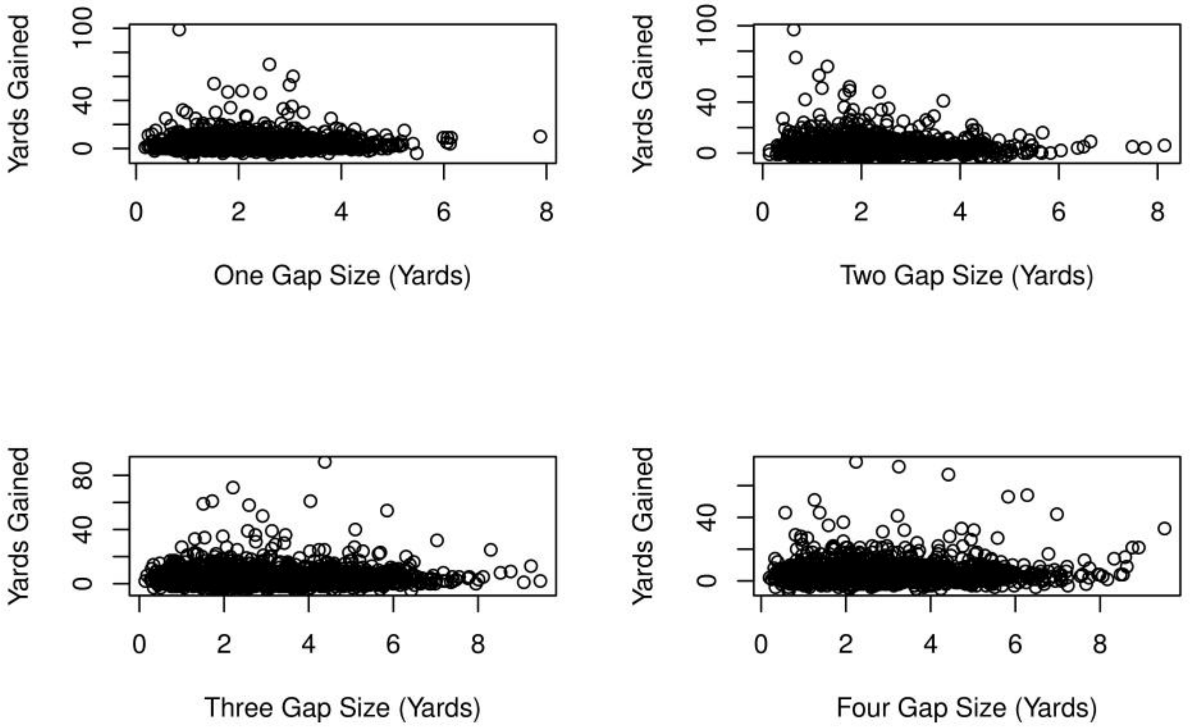
- We could model a defensive player's skill with their draft pick. When NFL recruiters and general managers choose which players to draft, they analyze countless statistics and footage in order to find the best player they can. By using draft pick to measure defensive player skill, we will be summarizing those statistics as well as recruiters' choices on who to draft in a single statistic. This approach is not without drawbacks, however. Some of these issues include length of time from draft. If a player has been in the NFL for several years, his draft pick may no longer be a good measure of skill.

In order to model defensive player skill we will use the second approach and include player draft pick.

Predictive Model

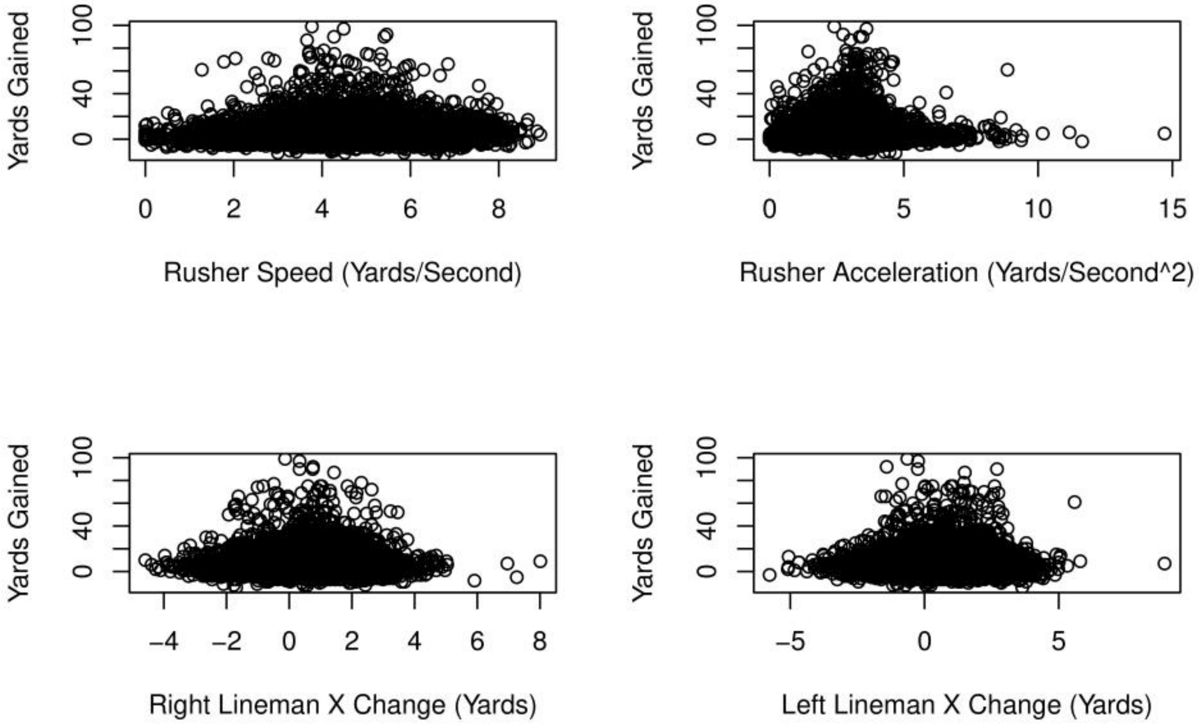
In order to estimate the causal effect of defensive formation on yards gained, we decide to use a general additive model. A general additive model is an extension of linear regression that allows our predicted variable (yards gained in this case), to depend on an arbitrary smooth function of our predictors instead of strictly linear relationships.

Below, we plot the yards gained by intended gap size. Intended gap is calculated as the gap in which the runner is moving towards at the time of hand off. The gap size is the horizontal distance between the two offensive lineman that the gap lies between.



One will notice that we don't seem to see much of a linear relationship between the gap size and yards gained. However, looking at the plot of one gap size and yards gained (top left in the plots above) we appear to see more runs with exceptionally large yards gained when the gap size is approximately two yards wide. Additionally, looking at the plot for two gap size, we appear to see that yards gained increases as we get close to 2 yard wide gaps, then decreases as gaps get larger. This implies that we should model the relationship between gap size and yards gained in our model not with as a linear relationship, but as an arbitrary smooth relationship.

Additionally, below we plot rusher speed, rusher acceleration, left lineman x change, and right lineman x change going from left to right in the plots below. Left lineman x change is the distance that the offensive lineman left of the intended hole is from the line of scrimmage. A positive value implies he is winning and driving defenders back, while a negative value implies he is being driven back off the line of scrimmage. Right lineman x change is identical but for the offensive lineman right of the intended gap. We see similar nonlinear trends, and therefore we will use an arbitrary instead of linear relationship for these variables in our model.



Final Model

Now, we can describe our final model[2] that we will use to estimate the effect defensive formation has on yards gained. All player specific variables are recorded at the time of hand off.

- GS_g = Gap Size for gap $g \in \{1, 2, 3, 4, 5, 6\}$
- $\mathbb{I}_g = 1$ if g is the intended gap and 0 otherwise with $g \in \{1, 2, 3, 4, 5, 6\}$. Intended gap is the gap the rusher is moving towards at the time of hand off.
- RA = Rusher Acceleration
- RS = Rusher Speed
- LL = Distance of the offensive lineman left of the intended gap from the line of scrimmage.
- RL = Distance of the offensive lineman right of the intended gap from the line of scrimmage.
- $ADLP$ = Average draft pick of the linebackers on the field.
- $ADLBP$ = Average draft pick of the defensive lineman on the field.
- $\mathbb{I}_{OF} = 1$ if $OF \in \{Empty, I{-}Form, Jumbo, Pistol, Shotgun, Singleback, Wildcat, No\ Formation\ Listed\}$ is the formation of the offense and 0 otherwise.

[2] Our model was fit with the mgcv package in R

- $\mathbb{I}_{DF} = 1$ if $DF \in \{3-4, 4-3, 2-4, 3-3, 4-2\}$ is the formation of the defense and zero otherwise.
- f_i is an arbitrary smooth function

$\mathbb{E}[Yards\ Gained] = \beta_0 + f_g(GS_g)\mathbb{I}_g + f_7(RA, RS) + f_8(LL) + f_9(RL) + \beta_{D-Line}ADLP + \beta_{LB}ADLBP + \beta_{OF}\mathbb{I}_{OF} + \beta_{DF}\mathbb{I}_{DF}$

- $\forall g \in \{1, 2, 3, 4, 5, 6\}$
- $\forall OF \in \{Empty, I-Form, Jumbo, Pistol, Shotgun, Singleback, Wildcat\}$
- $\forall DF \in \{3-4, 4-3, 3-3, 4-2\}$

Note, this implies that we are using $No\ Formation\ Listed$ as our reference level for offense formation, and we are using $2-4$ as our reference level for defense formation.

Second, we include the function $f_7(RA, RS)$. The function of two variables in this case allows interaction between rusher acceleration and rusher speed. We decide to include interaction between these terms because, for example, a rusher with low speed and high acceleration is likely not performing nearly as well as rusher with high speed and high acceleration at hand off.

Additionally, note that we choose to represent defensive player skill as average defensive lineman draft pick and average linebacker draft pick. We choose to compute the average in order to avoid multicollinearity in our model. If we did not average these and included the draft pick for each defensive lineman and each offensive lineman we would introduce multicollinearity into our model between defensive formation and these draft picks.

Estimated Causal Effect

Now that we have our predictive model with controls chosen by the back-door criterion, we can estimate the causal effect of defensive formation on yards gained by considering the coefficients for each defensive formation in our model. In the table below, we report the estimates of the coefficients as well as a bootstrapped[3] 95 percent confidence interval.

Table 1: Estimated effect of defensive formations with 2-4 formation as reference level

Formation	Effect Estimate (Yards)	Lower Bound 95% CI	Upper Bound 95% CI
4-2	0.087	-0.201	0.399
3-3	-0.013	-0.389	0.344
3-4	-0.413	-0.744	-0.053
4-3	-0.490	-0.799	-0.153

As one can see from the table above, the confidence intervals for our estimate of the causal effect of the 4-2 and 3-3 formation with the 2-4 as reference are centered around zero. This implies that the 4-2 and 3-3 formation likely give no significant advantage over the 2-4 defense formation. Additionally, we see that the estimated effect of the 3-4 formation is -0.413 and the estimated effect of the 4-3 is -0.490. However, due to the large overlap between these confidence intervals, this

[3] We chose to bootstrap by resampling residuals. This is because we trust the shape of our regression function, however the residuals are likely non-normal and skewed right because of some running plays with very large yard gains.

difference does not appear to be significant. All of these results together imply that the different defense formations are equally as good at stopping the run up to the number of players in the box. Adding an additional player to the box decreases the expected yard gain by approximately 0.4 to 0.5 yards.

Limitations and Future Work

Because this analysis has been done with observational data there are of course limitations and possible issues. The first and most important possible issue is with our directed acyclic graph. Our estimate of the causal effect relies on this being the true causal graph. Although we believe that our graph is well justified and mostly correct, it is impossible to know this for sure.

Second, our estimates of defensive line skill and linebacker skill could likely be improved. To model these players' skill we decided to use their draft picks; we argued that the draft pick was representative of players' skill level because they are chosen in a competitive draft and coaches and general managers want to get the best players. While we believe this is true, and we believe that draft pick is an adequate measure of player's skill more sophisticated measures of player skill would have been interesting and may give us a slightly more accurate estimate of the causal effect.

Conclusion

In this analysis, we have attempted to estimate the causal effect that defensive formation has on the run game. In order to do this, we used a predictive model to estimate the number of yards gained from several defensive formations while controlling for a carefully selected set of variables. We found that there was no significant difference between the formations with 6 players in the box, and there was no significant difference between formations with 7 players in the box. However, by introducing a 7th player into the box the yards gained in a run play is expected to decrease by about 0.4 or 0.5 yards.

This analysis has shown that the stopping the run game should not be a consideration for defensive coordinators when choosing the defensive formation up to the number of players in the box. In other words, because all of the common formations with 6 players in the box stop the run game all equally as well and all common formations with 7 players in the box stop the run equally as well, defensive formations should be chosen only considering the pass game. Defensive coordinators should first decide how well they want to protect against the run, and with that the decision they should choose if they want a formation with 7 players in the box or 6. Then, they should choose which formation to line up in after that only considering the pass game, and only considering factors such as how much of a pass rush they want or how well they want to defend against the pass in the open field.

Appendix

The definition of the *Back-Door Criterion* and supporting definitions are mostly paraphrased from *Advanced Data Analysis from an Elementary Point of View* (Cosma Rohilla Shalizi).

- *Path*: A *Path* is a sequence of edges that begins at a vertex and travels through the vertices of a graph. *Paths* can be both directed and undirected. A directed path must follow the direction of the arrows while an undirected path does not have to.

- *Back Door Path*: A *Back-Door Path* is an undirected path between our dependent variable, Y, and our independent variable, X, with an an arrow going into X. In our analysis, Y = Yards Gained and X = Defense Formation.
- *Collider*: Consider a path with at least three nodes. Consider any arbitrary set of three nodes in sequence along that path. Label these nodes A, B, and C with A coming before B and B coming before C. The node B is a collider along that path if and only if edges coming from both A and C are pointing into B.
- *Non-Collider*: A node which is not a collider.
- *Blocked Path*: A path is a *Blocked Path* if by conditioning on some set of nodes S if one of the two following statements hold for some node, Z, along the path.

1. Z is a non-collider and in S.
2. Z is a collider and neither Z nor any of its descendants along the path are in S.

Finally, the *Back-Door Criterion* states that if a set of variables or controls, S, blocks every *Back-Door Path* between X and Y and no variables in S is a descendant of X then S satisfies the *Back-Door Criterion.* Therefore, the set of controls S is sufficient to estimate the effect X has on Y.

References

Benjamin Alamar & Keith Goldner (2011) The Blindside Project: Measuring the Impact of Individual Offensive Linemen, CHANCE, 24:4, 25-29, DOI: 10.1080/09332480.2011.10739883

Cosma Rohilla Shalizi, Advanced Data Analysis from an Elementary Point of View (Cambridge University Press), http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV

Maksim Horowitz, Ron Yurko and Samuel Ventura (2019). nflscrapR: Compiling the NFL Play-by-Play API for easy use in R. R package version 1.8.1. https://github.com/maksimhorowitz/nflscrapR

Malumphy, C. (2000-2019). Drafts by Year. Retrieved from http://www.drafthistory.com/index.php/.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. Frontiers in Genetics, 10. doi: 10.3389/fgene.2019.00524

Gordeev, D., & Singer, P. (2019, November 28). Public LB 1st place The Zoo. Retrieved December 17, 2019, from https://www.kaggle.com/c/nfl-big-data-bowl-2020/discussion/119400.