Modelling Run Outcomes by Field Control

What snapshots of player tracking data can tell us about play outcomes.

Caio Brighenti

Submission for the student subcompetition of the
NFL Big Data Bowl

Colgate University
Hamilton, NY
December, 22nd, 2019

1 Data Description and Motivation

In recent years, the football analytics community has all but discarded the value of running in producing successful NFL offenses, finding that runs are less effective in almost every situation outside of short yardage on late downs or in the Redzone. NFL coaches and players might still tout the value of "establishing the run" in faciliting the passing game, or in wearing out and fooling the defense, but there's strong empirical evidence against both these arguments.[1] But, if you're a fan of the Jaguars who watched Christian McCaffrey tear apart your run defense for 176 yards in Week 5, or a Cowboys fan who watched Aaron Jones put in 4 critical touchdowns that same week, it probably won't be easy to convince you the running game doesn't matter. Regardless of what the statistics tell us, anyone who watches football consistently knows how destructive the run *can* be, in the right conditions. So what exactly are these conditions? Instead of rehashing similar arguments as to why rushing is strategically inferior to passing, I explore what the conditions for positive, effective runs look like.

I leverage the availability of player-tracking data for 23,171 NFL rushing plays made available for this year's NFL Big Data Bowl. For each play, positional, directional, and movement data is available for each player on the field at the moment the ball is handed off to the running back. In other words, the data offer a snapshot of each play, alongside how many yards were gained on that play, as well as a host of game-status variables such as the down, quarter, and time on the clock. Many studies use game-status variables to predict the outcome of plays, answering, for instance, the likelihood of converting on 4th-and-1 at your own 30. My work, instead, aims to predict the outcome of plays using only what's happening on the field, agnostic to "strategic" game variables. This is because, intuitively, the quarter or down should have no impact on how far a running back can run, outside of the effect that game-status variables have in the strategic play-design choices of coaches, which will naturally manifest in how the field status takes shape.

The player tracking data was first prepared for analysis by applying a few standardization steps. Given that offenses will go towards either direction on the field during the course of an NFL game, plays where the offense was moving towards the right were flipped in order to ensure consistency across all plays in the dataset. This required adjusting not only the (x, y) coordinates of each player, but also their direction of motion.[2]

2 Clustering Run Types

I began by first identifying what different kinds of runs are observed in the dataset. To avoid biasing my grouping based on my prior assumptions, I applied a simple k-means clustering approach using the ball carrier's position on the field, distance from the line of scrimmage (LOS), speed, and direction of motion. After some experimentation, $k = 6$ appeared to best cluster these groups, as for values of k past 6, differences between some clusters appeared negligible. I visualized the resulting clusters by plotting the cluster center of each as a vector, alongside a sample of 20 plays from each group. It was immediately clear these clusters varied primarily in runner's direction of motion, and could be labeled by the different positions in the defensive line being attacked. This plot, along with the label I assigned to each cluster, is shown below in Figure 1, where the LOS is the highlighted line.

[1] For great work on both questions, see Josh Hermsmeyer's 538 article or Ben Baldwin's Football Outsiders article, both listed in the references section.

[2] Credit to the NFL's own Michael Lopez for standardization code used.

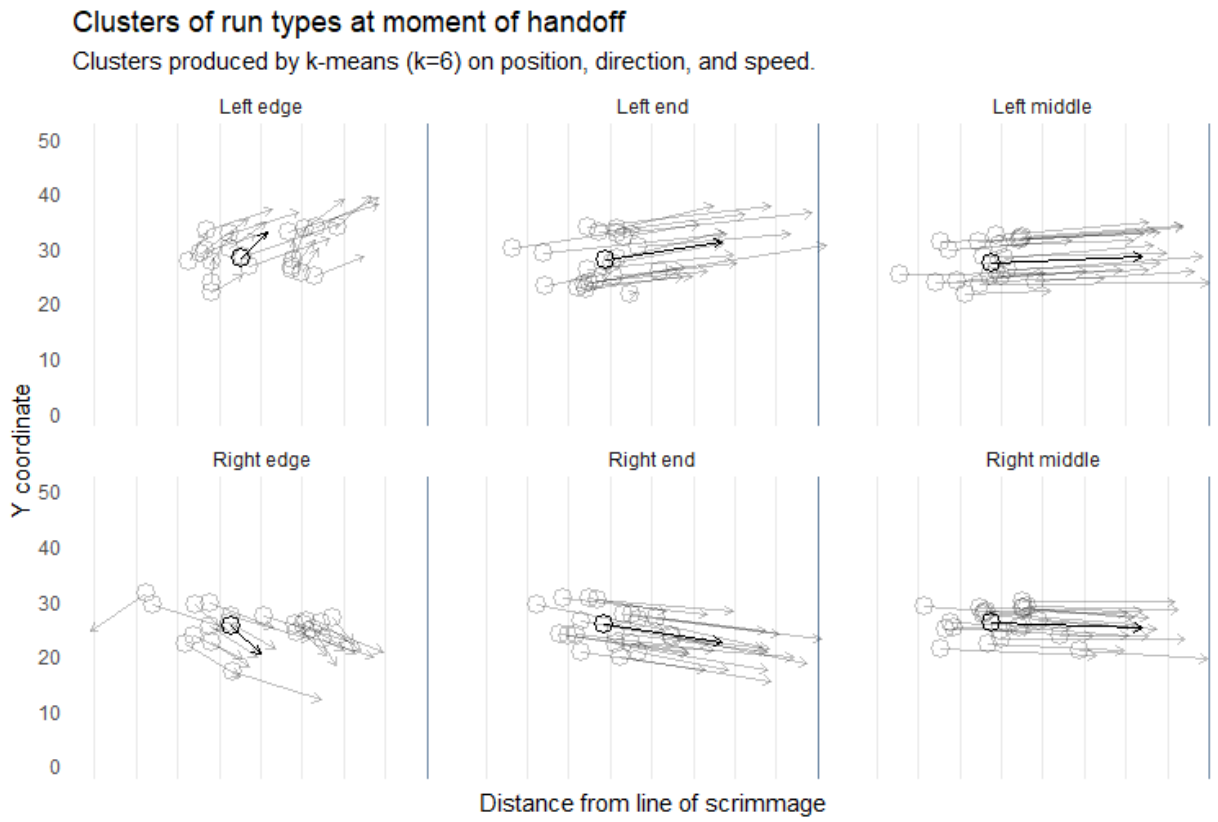

Figure 1: Resulting k-means clusters for different run types ($k = 6$).

Again, the above plot confirms that the clusters distinguish between themselves mostly by the ball carrier's direction of motion. To further analyze these clusters, the centers for each are shown in the following table.

Cluster	Distance from LOS (yards)	Y coordinate (yards)	Speed (yards/second)	Direction (°)
Left edge	5.54	28.01	5.09	7.21
Left end	4.90	27.93	4.34	41.21
Left middle	4.77	27.36	3.84	71.87
Right middle	4.77	26.16	3.83	106.24
Right end	4.86	25.64	4.27	137.47
Right edge	5.31	25.34	5.13	171.67

Table 1: Table of cluster centers from k-means clustering ($k = 6$)

The table of cluster centers also makes clear how little variation there is between pairs clusters of the same defensive position, but one towards the left and the other towards the right. The distance from the LOS and speed centers are nearly identical between each left-right cluster pair. This indicates that not only are clusters best grouped by direction of motion, but there is also no real difference between runs towards the left or right. In other words, runs towards the right can be considered simply mirrored versions of runs towards the left. With this in mind, I first reflected plays where the running back was moving towards the right ($\theta \in (90, 270)$) along the midpoint line ($y = 25$), then manually clustered the runs using the minimum and maximum angles for the above clusters, which had nearly no overlap. The resulting clusters are shown below in Figure 2.

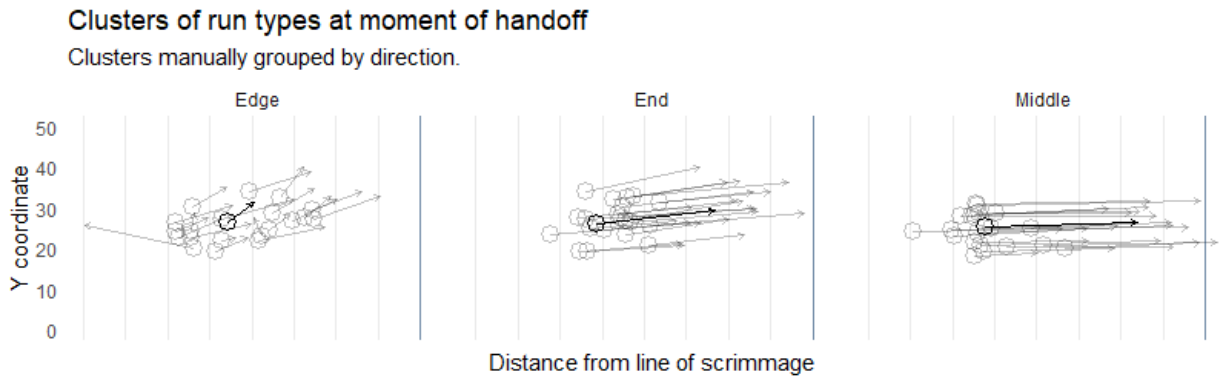

Figure 2: Clusters for plays manually grouped by direction of motion.

With these clusters in place, I now had groupings of runs with similar objectives. This is key to analyzing how varying field status might affect the outcome of a run, given that runs to different locations depend on different scenarios. In other words, a wide open gap towards the sideline is irrelevant if the runner is barreling down the middle of the field. The importance of run type and field status in predicting the outcome of a play can be graphically understood by comparing the following two plays with similar runner motion, but significantly different outcomes.

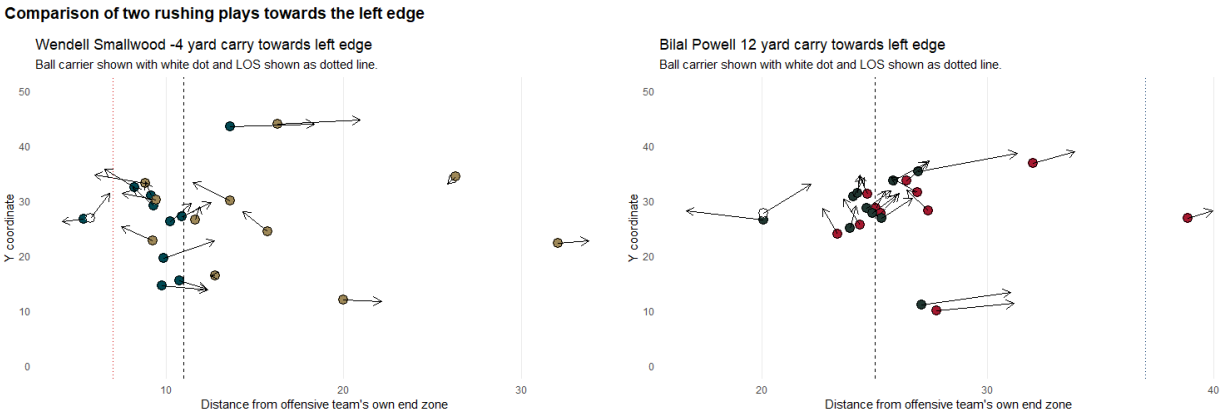

Figure 3: Comparison of two rushing plays towards the left edge.

It should be immediately obvious why the above plays in Figure 3[3] had such different outcomes. The question remains, however, is how this graphically visible field status can be quantified, so that it might be aggregated and analyzed for the entire dataset. To do this, I apply a method for quantifying and computing field control for any given play in the next section.

3 Computing Field Control

To compute the relative field control, I used the player influence technique presented by Javier Fernandez and Luke Bornn.[4] This approach treats a player's influence $I(p,t)$ over some point p at time t as a bivariate Gaussian distribution transformed by the player's speed (s) and angle of motion (θ). This transformation is done by adjusting the mean and covariance matrix.[5] Given the availability of positional and motion data in

[3] Again, credit to the Michael Lopez for the play plotting code I used as base.

[4] Javier Fernandez and Luke Bornn. "Wide Open Spaces: A statistical technique for measuring space creation in professional soccer". In: (2018)

[5] For more detail, see Appendix A of Fernandez and Bornn, or the Python implementation at this Kaggle notebook.

the NFL Big Data Bowl's dataset, this approach is a natural fit. By calculating the individual field control for each offensive player (i) and each defensive player (j) at any given point (p), the overall field control can is given by the following, where σ is the logistic function $\sigma(x) = \frac{1}{1+e^x}$.

$$FC(p,t) = \sigma(\sum_i I(p,t) - \sum_j I(p,t))$$

The logistic function is used to ensure the field control values remain in range [0,1], where a value of 1 represents complete offensive control, 0.5 represents equally shared control, and 0 represents complete defensive control. By calculating the above for each point on the field, a complete picture of the entire field's control at time t can be computed. To illustrate this, the two plays compared above are shown again, but this time showing just the ball carrier and the field control surface in Figure 4. Comparing these two plays, the influence of the defensive control—represented by the darker area—stretching around the area of offensive control and closing off the runner's path is clearly visible.

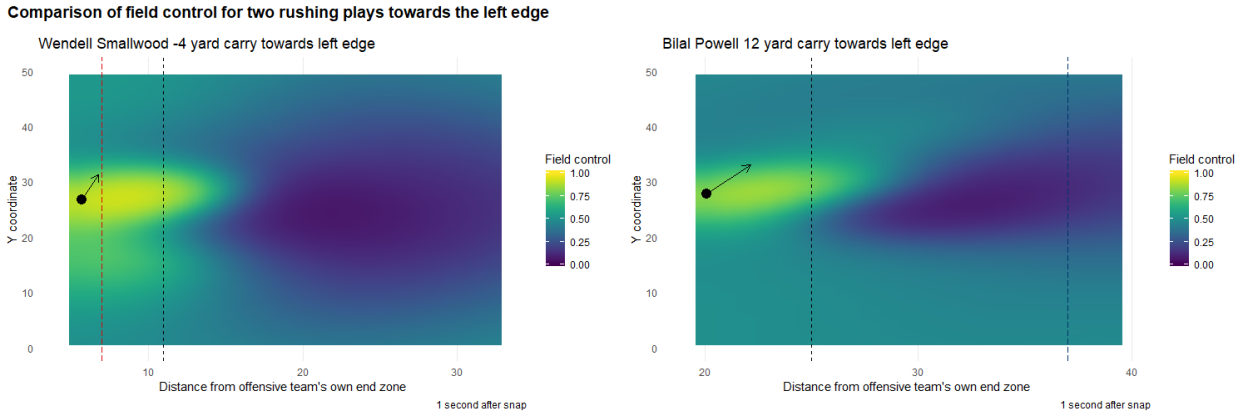

Figure 4: Comparison of field control for two rushing plays towards the left edge.

While looking at individual plays is helpful in illustrating potential patterns, I was more interested in identifying relationships across the entire dataset. For instance, I wanted to determine whether and how the field control changes for runs of different types, and for successful and unsuccessful plays of the same type. This can be done by calculating the field control for all plays in the dataset, then aggregating the results by the groups of interest. For this to be possible, I first had to standardize each play so that plays would be comparable with one another. I did this by centering all x coordinates for each play around the LOS, so that the x variable measures a player's distance from the LOS, and not their field position.

Next, I computed and stored the field control for each of the 23,171 plays in the dataset. For each play, I used a standard area of 37 yards horizontally (11 yards behind and 26 in front) and 50 yards vertically around the LOS, for a total of 1850 individual points per play.[6] The reasons for this were twofold: first, for all plays to be aggregatable they must cover the same area, and second, it would be computationally infeasible to compute the complete field control for all plays. Even using this limited area, calculating field control for all 23,171 plays was only possible by parallelizing the computations and running them on a 20-core compute cluster. With field control computed for each play, I averaged the control at each point p by run type, and plotted them in Figure 5. As before, plays towards the right were reflected around $y = 25$.

[6]It should be noted here some plays near the endzone had less than 26 possible yards towards the right of the LOS. For these plays the maximum available space was used, which can be computed by 110-LOS.

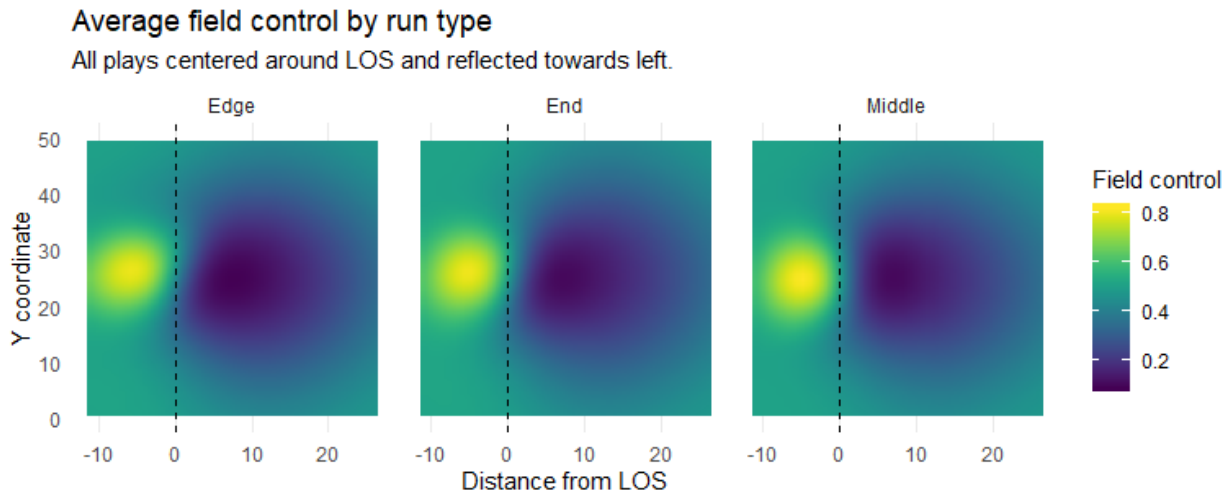

Figure 5: Average field control per run type.

Figure 5 indicates little instrinsic differences in the setup of the field by different run types, aside from field control for plays towards the middle being more densely concentrated at the LOS. Though perhaps uninteresting, this conclusion makes some sense. One might expect that offensive play design would closely relate the offensive line play—and consequently the field control situation—with the running back's motion. This, however, is likely to manifest only in *succesful* plays, given that the defensive line serves as an adversary with the goal of disrupting the intended field architecture of the play design. Given that these averages are comprised of over 23,000 plays in total, and that running plays more frequently fail than succeed, it is logical that this structure would dissapear over aggregation.

A more productive approach was to analyze how field control varied not just between run types, but between successful and unsuccessful runs of the same type. Typically, in the NFL analytics community, a successful play is defined as one that leads to an increase in the probalistic odds of scoring points, but this is a strategic measure. Instead, I define successful runs as ones that get past the line of scrimmage, gaining more than one yard. This approach clarifies what patterns in field control might lead to successful runs. To visualize this, I recreated the graph of average field control by run type, but instead plotting the average succesful plays minus the unsuccessful plays. This plot is shown in Figure 6.[7]

[7]This plot omits plays where the LOS is on or past the 95 yard line, as the lack of full range of points for these plays produced visual issues in the plot, and their exclusion did not change the visible patterns.

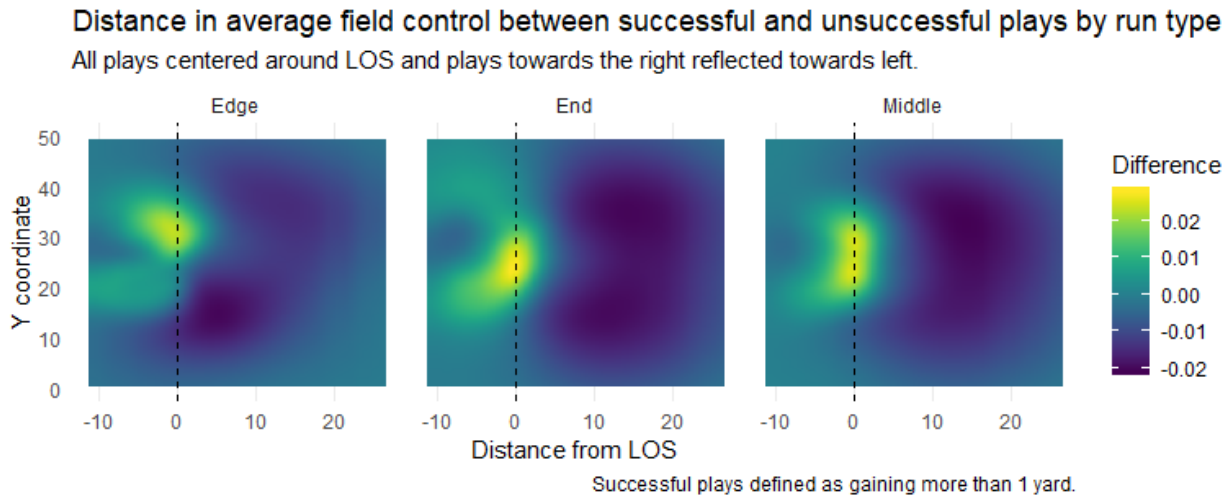

Figure 6: Distance between successful and unsuccessful plays in average field control by run type.

The interpretation for Figure 6 is slightly different than Figure 5. The bright and dark spots do not necessarily indicate a predominance of offensive or defensive control, respectively. Instead, bright spots are areas where the offense has, on average, greater control on successful plays than unsuccessful plays, and the opposite for darker spots. Strikingly, these plots indicates that on plays where the running back is able to get past the LOS, the offense has *less* control over the area a few yards past the LOS . On the other hand, in all three cases the offense has stronger control over sections of the LOS, and the location of this shifts to the left as the run angle increases. In other words, the "gap" the running back is running towards is directly visible. This makes sense, and indicates that what is key to distinguishing between runs that will pan out and those that will get stuffed is the offensive control over the running back's target gap.

To test this finding, I recreated Figure 6 using a minimum of 5 and later 10 yards gained as the boundary for successful plays. In these plots, the darker areas immediately past the line of scrimmage receded farther back towards the endzone, resulting in nearly no difference between successful and unsuccessful plays shortly past the line of scrimmage. At the LOS, however, the bright spots remained, indicating that even for medium and long yardage runs, the most important factor is control of the target gap.

If a gap in the runner's path is indeed the most salient indicator of successful runs, then quantifying whether the runner is heading towards a gap could be a strong approach for predicting the outcome of runs. Aditionally, identifying whether a gap exists—even if not on the runner's path—could be used to determine the "ideal" path for the runner, which in turn could be used to quantify how easily the runner could adjust to the ideal path, another likely strong predictor of run outcomes. With these objectives, I proceed in the next section to determine direct measures of these ideas, and fit a linear regression model to predict yards gained.

4 Modeling Play Outcomes

Before implementing a regression model, I first established how to determine what, at the moment of the handoff, the ideal running angle would be. Given the conclusions from the previous section, an ideal running angle for rush success should point directly to the largest gap on the LOS, where the defensive team has the weakest presence. Given the standardization approach centered coordinates on the LOS, and the vertical range is $[0, 50]$, the set of points on the LOS is expressed by

$$P_{los} := \{p \in P_{los} \mid p = (0, y),\ \forall\, y \in [0, 50]\}$$

The ideal point on the LOS (p_i) is defined as

$$p_i := arg\ \underset{p \in P_{los}}{max}\ I(p, t)$$

In other words, p_i is the point on the LOS with maximal offensive control. In the case of a tie, one is chosen at random.[8] Similarly, the *expected* point where the runner will arrive at the LOS (p_e) can be easily obtained by calculating where the runner's motion vector will intersect with the LOS. Formally, given the player's position ($p_{player} = (x, y)$), horizontal distance from the LOS (d_h), and angle of motion (θ), the expected intersection point is[9]

$$p_e := (x + d_h,\ min(y + \tan(max(\theta, 0))d_h,\ 50))$$

To illustrate how the expected and ideal intersection points on and paths towards the LOS are calculated, I revisited the unsuccessful example play. Figure 7 again plots the field control and running back for this play, but this time also plots the expected and ideal paths, where the expected path moves towards the sideline, and the ideal towards the middle of the field. This makes clear how different the two points can be.

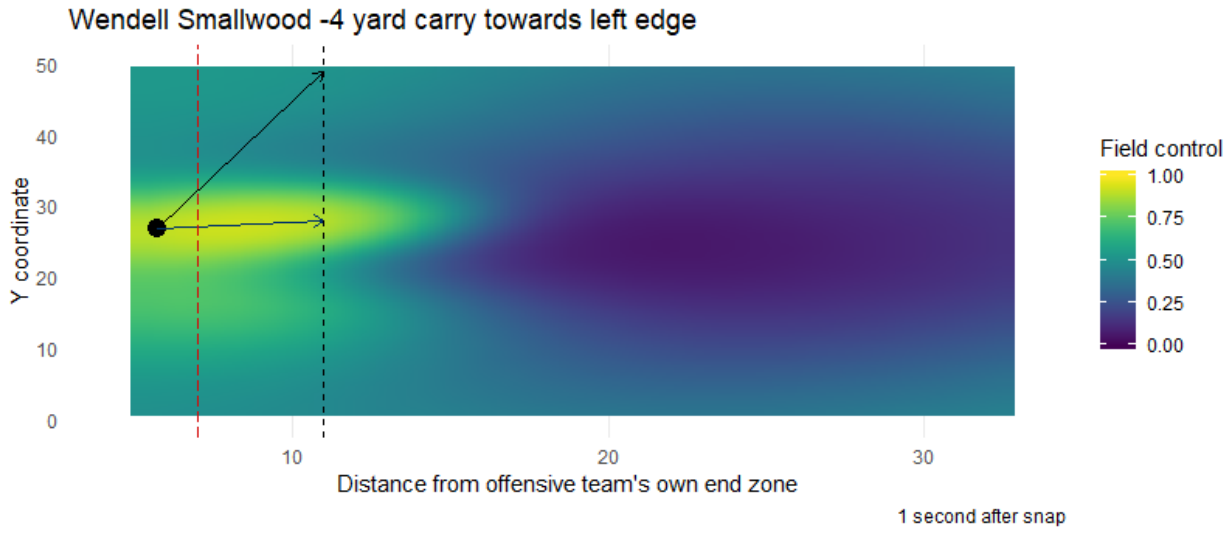

Figure 7: Expected and ideal path towards the LOS for sample play.

Given the expected (p_e) and ideal (p_i) points, the distance between the player's actual target and ideal is easily found by $d_{ei} = |y_e - y_i|$, given that $x_e, x_i = 0$ since both are on the LOS. Aditionally, the length of the expected and ideal paths are $l_e = d(p_{player}, p_e)$ and $l_i = d(p_{player}, p_i)$, respectively, where the function $d(p_1, p_2)$ computes the Euclidean distance between the input points.

Finally, the field control at points p_e and p_i must be found. Instead of recomputing these using the technique described in Section 3, I leveraged the already computed values for each play. However, I had only computed field control values for each play for integer coordinates. This is not a problem for p_i, given that—due to the way it is defined—it will always be an integer pair of coordinates, but it is for p_e, which is calculated as a product of the runner's angle of motion and position, both on continuous scales. The control $I(p_i, t)$ is thus already defined, but $I(p_e, t)$ is not, and must be either recomputed or approximated. To avoid incurring additional computational costs, I approximated values for p_e using the fact that, for any given point on the LOS $p = (0, y)$, the closest points where field control values are available must be $p_1 = (0, \lfloor y \rfloor)$ and $p_2 = (0, \lceil y \rceil)$. The field control at point p_e at time t can thus be approximated as follows, where d_i is the Euclidean distance between p_e and p_i:

$$I(p_e, t) \approx (1 - \frac{d_1}{d_1 + d_2})I(p_1, t) + (1 - \frac{d_2}{d_1 + d_2})I(p_2, t)$$

In other words, the field control at p_e is an average of the field control at the 4 closest points, inversely weighted by the distances between each point. In practice, $d1 + d2$ will always equal one, simplifying the

[8] An extension of this approach could consider a small range around p_i, as opposed to the single point with "best" control.

[9] The $min(y_e, 50)$ and $max(\theta, 0)$ operations are used to avoid plays overflowing past the boundaries of the field, such as if the player's angle of motion points vertically up or away from the LOS.

calculation. Using this approach, I approximated the field control $I(p_e, t)$ for each play. I then used these variables, along with the player's speed, to fit a linear regression model predicting the yardage gained. In short, the predictors used are:

- s, the player's speed
- FC_e, the field control $I(p_e, t)$ at the expected point of intersection
- FC_i, the field control $I(p_i, t)$ at the ideal point of intersection
- d_{ei}, the distance between p_e and p_i
- l_e, the length of the expected running path
- l_i, the length of the ideal running path
- y, the yards gained on the play,

with interactions between d_{ei} and FC_i, as logically the relevance of the ideal gap depends on the runner's ability to arrive at this gap, and between s and l_e and l_i, as speed dictates how long a path would take. The linear regression model thus takes the form

$$y = \beta_0 + \beta_1 s + \beta_2 FC_e + \beta_3 FC_i + \beta_4 d_{ei} + \beta 5 l_e + \beta_6 l_i + \beta_7 d_{ei} \cdot FC_i + \beta_8 s \cdot l_e + \beta_9 s \cdot l_i$$

I fit two linear models were using this specification, one on the entire dataset, and the other only on plays where the runner gained at most 10 yards. I did this to validate that the results observed in the first were not due to entirely to outlying plays where the running back breaks free of the defense for a massive gain. The regression results for the first and second model are shown in Tables 2 and 3, respectively.

Table 2: Regression outputs for model fit on full data.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.9341	0.5749	-1.62	0.1043	
s	0.9597	0.1083	8.86	>.001	***
FC_e	2.3496	0.5107	4.60	>.001	***
FC_i	0.1343	0.6468	0.21	0.8355	
d_ei	0.0111	0.0202	0.55	0.5830	
l_e	0.2487	0.0591	4.21	>.001	***
l_i	-0.0365	0.0259	-1.41	0.1581	
FC_e:d_ei	0.0300	0.0408	0.74	0.4619	
s:l_e	-0.0632	0.0128	-4.94	>.001	***
s:l_i	0.0020	0.0053	0.37	0.7125	

Adjusted $R^2 = 0.011$

Though neither models explain much of the variance in yards gained, there are several significant predictors. In both models, the player's speed, the field control at the expected point of intersection, and the interaction between player speed and the length of the expected path are highly significant ($p < .001$). For the first model, the length of the expected path is also highly significant ($p < .001$), and significant ($p < .05$) for the second model. In the second model, the distance between the expected and ideal points on the LOS and the length of the ideal path are also significant. In terms of coefficient magnitude, s and FC_e are the most impactful predictors. The coefficients for FC_e show that if the offense has complete control of the expected point of intersection ($FC_e = 1$), we can expect that play to go for about 2 yards more, on average, than it otherwise would. The coefficients for speed (s) indicate that for every unit increase in the player's speed, we expect to see, on average, an increase in the yardage gained between 0.5 and 0.8.

The third term highly significant in both models, the interaction between s and length of the expected path, is harder to interpret. In both cases, the individual effect of l_e is both significant and positive,

Table 3: Regression outputs for model fit on plays where yards gained <= 10.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.2061	0.2823	0.73	0.4654	
s	0.5648	0.0532	10.62	>.001	***
FC_e	1.8308	0.2509	7.30	>.001	***
FC_i	-0.0356	0.3189	-0.11	0.9112	
d_ei	0.0227	0.0099	2.30	0.0216	*
l_e	0.0685	0.0291	2.36	0.0183	*
l_i	-0.0312	0.0127	-2.47	0.0136	*
FC_e:d_ei	-0.0130	0.0200	-0.65	0.5141	
s:l_e	-0.0377	0.0063	-5.98	>.001	***
s:l_i	0.0021	0.0026	0.79	0.4268	

Adjusted $R^2 = 0.031$

suggesting longer paths to the LOS lead to higher yardage gain. This is only true, however, when the player speed is under 3.9 in the first model, and 1.8 in the restricted model. This indicates that the number of plays where l_e *actually* has a positive impact on the yardage gained is quite small, given that a speed of 1.8 is in the bottom 2% of the data. That the effect of l_e is positive even for speeds up to 3.9 in the long yardage case suggests long yardage plays could disproportionately come down the sideline, where the back has to travel farther to reach the LOS. It is also likely that this effect is partly produced in unpredictable ways by situations where the back changed direction, given that this effect is linked with slower speeds, which would necessarily make easier abruptly switching directions.

Interestingly, predictors capturing the ideal gap in the LOS are virtually irrelevant in both models, with only d_ei and l_i being significant only in the second model, and FC_i being completely insignificant in both cases. Additionally, no interactions with FC_i or l_i were significant. There are several possible reasons for this. First, it may be that the ideal spot is not relevant to the success of the play because the running back cannot consistently adjust to target it, and is instead locked in to their current path. Secondly, it may be that the variables capturing information about the ideal gap add little information that the FC_e predictor does not already provide. The predictor d_ei, for instance, is strongly correlated with FC_e, as it makes sense that the closer the expected point is to the ideal one, the higher the field control FC_e will be. I confirmed this by calculating the variance inflation factor (VIF) for each individual predictor to test for multicolinearity, which showed d_ei ($VIF = 7.48$) and l_i ($VIF = 6.39$) were correlated with other predictors, as a VIF of 5 is usually the boundary for the presence of multicolinearity.[10]

5 Conclusions and Future Work

In beginning this work, I set out to answer the following question: what makes a run effective? In short, an extensive analysis of how field control varies between succesful and unsucessful runs, alongside linear models predicting the outcome of runs using running back speed, expected running path, and field control at the LOS indicate that if you want a run to be succesful, then the running back better be pointed directly at a gap in the defensive line and be moving fast at the moment of the handoff. These two variables were significantly more impactful than how far the back would have to travel before reaching the LOS, and whether other, better gaps existed at the LOS. This suggests, for play design purposes, that runs work best when kept simple. Expecting plays where the back receives the ball before committing to a run is likely a recipe for poor runs, even if Le'Veon Bell's hesitation runs produce explosive plays here and there.

Naturally, there are significant limitations to the conclusions I found in this work. Given that data is only available for the moment of the handoff, it is impossible to determine precisely how any given play unfolded. Further exploration into this question would benefit greatly from full tracking data across entire

[10]Note that these values were calculated by fitting a third model without the interaction terms, as interactions necessarily exhibit high multicolinearity.

plays being made available. This would permit testing whether succesful runs do indeed tend to be straight shots, or whether the empirical findings warrant a more nuanced explanation. In a similar vein, complete tracking data would allow for the motion of running backs to be accurately treated as curves, as opposed to the obviously oversimplified direct line approach in this paper. Finally, complete tracking data would allow for pitch control to be measured and explored as the play evolves, opening the door for an unimaginable number of questions.

References

[1] Ben Baldwin. "Rushing Success and Play-Action Passing". In: (2018).

[2] Javier Fernandez and Luke Bornn. "Wide Open Spaces: A statistical technique for measuring space creation in professional soccer". In: (2018).

[3] Josh Hermsmeyer. "Can NFL Coaches Overuse Play-Action? They Havent Yet." In: (2019).

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org`.