Beyond EPA

Paul Sabin

Background of EPA

Expected Points Added (EPA) has emerged from an obscure stat from quants to the mainstream, occasionally making television broadcasts and as evidence for and against TV personality's talking points analyzing football performance. *Background lit review on history of EPA models from Carter & Machol, to Burke, to Baldwin, to Brill.*

One advantage to using EPA as opposed to yards in football analysis is that it accounts for contextual information about the play that impacts the value of yards. For example 7 yards on 3rd and 5 is more valuable than on 3rd and 10 because the former results in a first down while the latter does not.

In team analysis, EPA allows division of credit among different units of a football team. If the offense throws a turnover in its own territory and the defense subsequently allows a field goal, EPA can tell us that the defense likely was a positive on the net scoring margin while the offense was negative in that sequence despite the three points allowed typically being assigned to the defense.

Similarly, if a punter is able to flip the field, the special teams unit will get credit for making the opposing offense have to go further to score.

Player Credit

EPA is used often to compare player performances. In fact total EPA gained on the season for quarterbacks mirrors closely the results of the MVP voting (*source).* In recent years, several attempts have been made to assign credit to players based on how that player's teams unit performs in terms of EPA (*Sabin Plus-Minus, Eager WAR, Yurko nflWAR, Baldwin EPA+CPOE composite, Kevin Cole WAR).*

Barring extreme situations where a player intentionally gives himself up instead of scoring at the end of the game to preserve possession and ensure victory, the goal of each player on a football team on each play is to advance the ball as far as possible towards the end-zone as possible (*for the offense)* and likewise to prevent that from happening for the defense.

This begs the question, do we use EPA as the preferred metric because it's the best metric to encapsulate individual and team performance on a play, or do we use it simply because it isn't *yards*?

Flaws of EPA

This paper will examine three main flaws that EPA has.

1. The large jump in EPA at the 1st down line vs 1 yard short of the first down line while the difference in player and team performance can largely be attributed to chance.
2. EPA does not follow a symmetric unimodal distribution, and depending on the situation it can skew left, skew right, be bimodal, unimodal, and more! This means that EPA per play for a team, unit, or player is affected largely by situation before the results of the play occur.
3. Selection Bias. Until recently all EPA models were built off observed plays. Better teams have more plays in the opposing territory, biasing the expectation towards better teams in those situations. Brill & Wyner 2024 use catalytic priors in an effort to adjust expected points models for this selection bias.

Yards vs EPA

No one with a knowledge of football can argue that 10 yards on 3rd and 10 and 10 yards on 3rd and 20 are worth the same to the offensive team. EPA accounts for this problem.

Let's consider the hypothetical situation, it is 3rd and 10 at exactly midfield in the 1st quarter of a 0-0 game.

Result A: The team gains 9 yards.

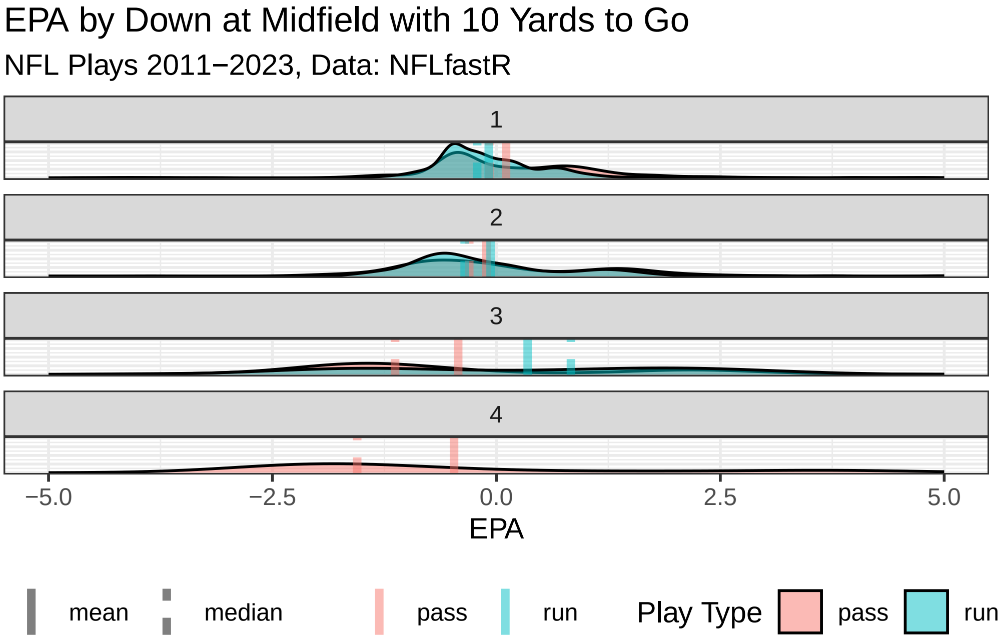
Result B: The team gains 10 yards.

Now in terms of EPA, according to the NFLFastR model the Expected Points before the play is 1.71 while the EPA for Result A is -0.09 compared to result B of 2.02.

While in yards the team gained 10% more in situation B than A, in EPA one play is drastically different than the other. That is a massive difference for players with essentially a coin flip difference in result (*reference to Spatial proximity in causal inference*).

Changing Distributional Shape of EPA Outcomes

Let's look at the distribution of EPA for a pretty neutral situation at midfield for each down and play type (run & pass).

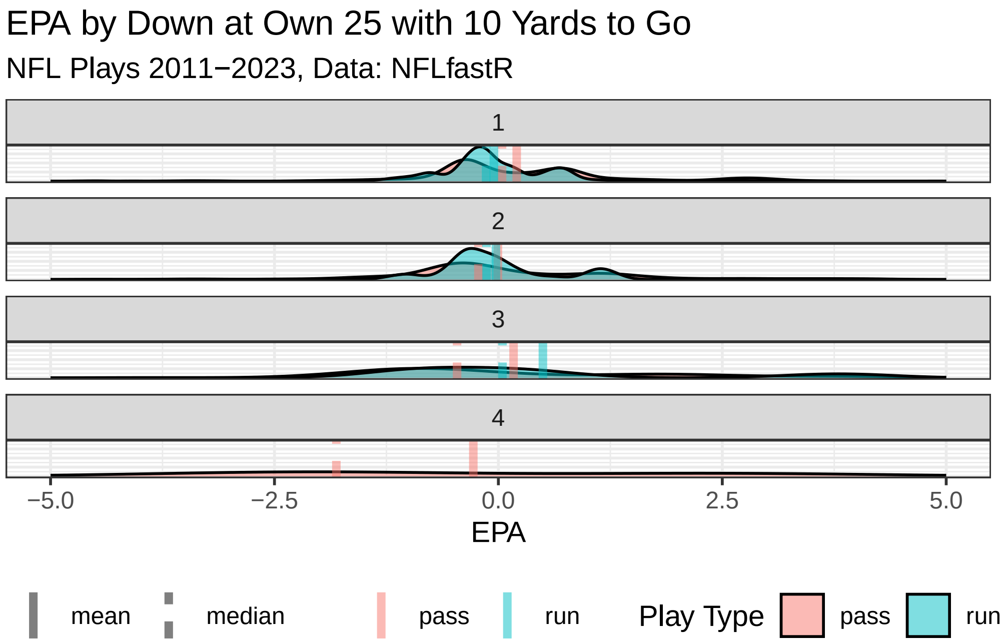


EPA by Down at Midfield with 10 Yards to Go

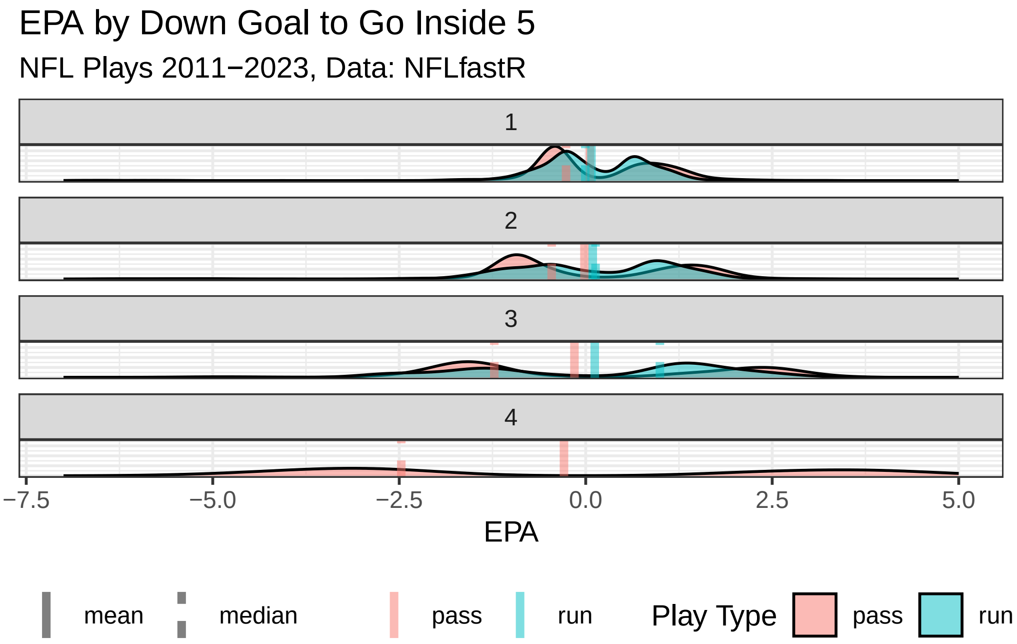
NFL Plays 2011−2023, Data: NFLfastR

For each down there is a bimodal distribution of outcomes which gets larger variance and more skewness for each additional down. This matters because the means are pulled away from the center of the distribution towards the side of the skew. Players and teams that are in this situation will get credit or blame for in terms of EPA for simply performing at the median or 50th percentile. For example a 50th percentile performance for a run on 3rd down and 10 at the 50 is worth 0.83 EPA despite it being simply the median outcome!

We can look at the most common starting field position of the 25 yard line (usually after a kickoff) as well.



We can look at an even more extreme example of this, for plays inside the 5 yard line in "goal to go" situations.



Methodology

This work proposes a method to quantify performance of a team, player, and unit for each and every play by how well they performed in the distribution of outcomes in similar situations. Since each situation has varying distribution shapes and properties, values are first quantified as percentiles then mapped via various link functions to well-known distributions.

Modeling Distribution of Play Outcomes

In a given football play for the offense the offense either runs a play that results in a yardage gain or loss, or ends in a score (touchdown or safety).

The following multi-stage model encapsulates what could happen as a result of an offensive play: 1. P(T) where event T is a turnover 2. P$(y \in 0, 1, 2, \ldots, 100|T)$ where y is the possessing team yardline at the end of the play (0 is a touchdown and 100 is a safety). 3. P$(y \in 0, 1, 2, \ldots, 100|T^c)$

Then the complete distribution of the outcome of the play follows:

$$P(y) = P(y|T)P(T) + P(y|T^c)(1 - P(T)).$$

Then we feed the resulting yardline, down, possessing team into the NFLFastR expected points model such that $z_{i,t,p+1} = E(x_{i,t,p+1}) - E(x_{i,t,p})$ where x is the eventual points scored at $i \in 0, 1, \ldots, 100$ yardline outcome, $t \in 0, 1$ for turnover outcome for play p. The resulting value of z is the corresponding Expected Points Added for each possible turnover and yardage outcome of the play.

For resulting touchdowns the expected points is assumed to be 6.96 and for safeties it is assumed as -2.0. The signs are negated if the result of the play is a touchdown or safety for the defensive team.

Classification Models

While theoretically the resulting yardline is a continuous value between 0 and 100, it is only recorded in public data as discrete integer values from 0 to 100. An xGBoost classification model is fit for the turnover probability and the two yardline models (one conditional on no turnover and the other conditional on there being a turnover).

Transitioning Classification Model to EPA Distribution

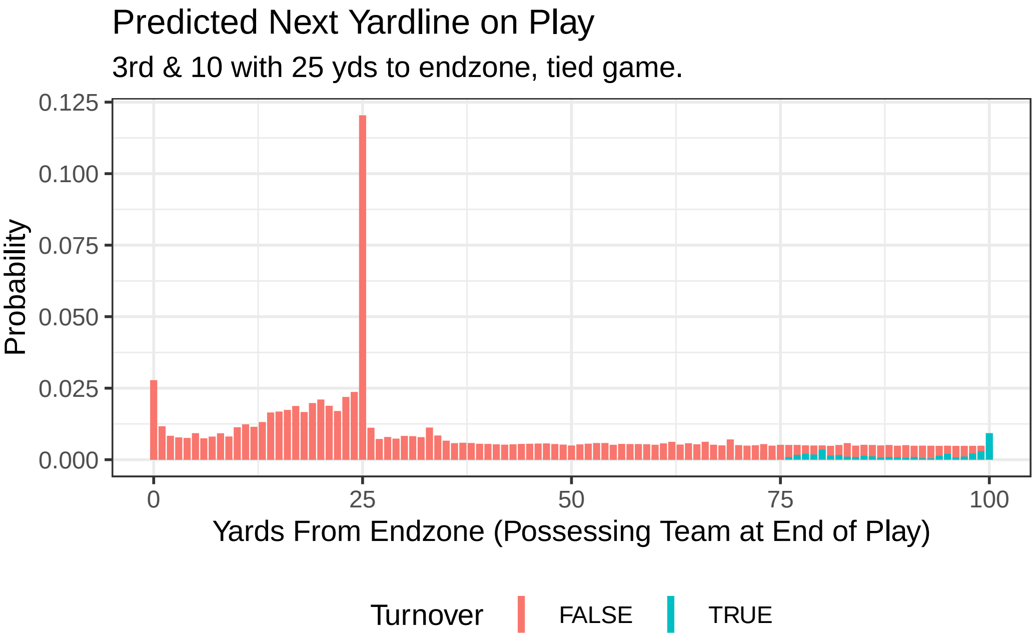
Using the 202 unique classification outcomes (101 for no turnover and 101 if there's a turnover), we can then convert the outcome of the play to asses what the new Expected Points of the next play will be. From there we can subtract from the expected points from before the play and get the 202 unique values of EPA that could result from the play outcome.

To keep things simple, we assume no timeouts remain the same before and after the play. We also assume exactly 5 seconds come of the clock on the play, or until the clock hits 0 on a quarter if there are less than 5 seconds left.

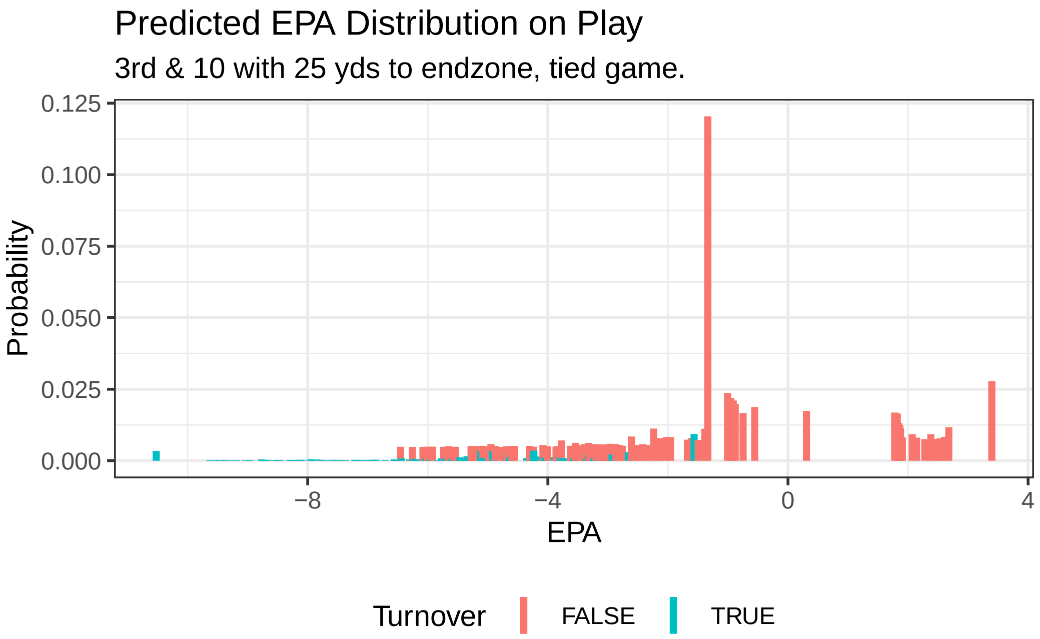
Plays that result in a touchdown (yardline = 0) are assumed to have expected points of 6.96 while plays that result in a safety (yardline = 100) are assumed to have -2 expected points.

In the example we have a 3rd and 10 at our other teams 25 yardline (25 yards from endzone), we use the classification models to calculate the probability of the next play resulting in any of the 202 possession & yardline combinations.

The current model would give an output that looks like this:



Then applying the expected points after each of the possible play outcomes we can develop a distribution of possible EPA outcomes:



Mapping Distributions

For each play in the dataset I saved the quantiles from the estimated EPA distribution from 0.005 to 0.995 in intervals of 0.005. I also calculated the observed EPA value and its quantile in the estimated distribution. For values greater than 0.995 and less than 0.005 I assumed those extremes to protect against extreme tail behavior.

From there I convert the observed quantiles for each play to different known distributions:

- Standard Normal Distribution
- Standard t distribution with 5 degrees of freedom (1 degree of freedom or *Cauchy* had erratic behavior so it was omitted)
- Standard laplace distribution
- Gamma distribution
  - The mean and variance of the gamma were decided to be equivalent to the observed mean and variance of EPA across all offensive plays after you subtract off the smallest observed EPA value so the support is positive like a Gamma distribution.

comp_var_1	avg_rank	pct_no_1	avg_correlation	sd_correlation
kgamma	2.165	0.577	0.4685426	0.0263641
knorm	2.618	0.215	0.4624901	0.0269240
kt5	2.583	0.104	0.4624740	0.0266177
klaplace	3.644	0.034	0.4598413	0.0266538
play_quantile	4.929	0.016	0.4494879	0.0276412
epa	5.763	0.030	0.4305393	0.0280615
yards_gained	6.720	0.024	0.4134349	0.0294258
adj_nya	7.581	0.000	0.4021861	0.0283624
wpa	8.997	0.000	0.3204819	0.0308437

stat	avg_rank	pct_no_1	avg_correlation	sd_correlation
knorm	1.649	0.409	0.5486108	0.0188296
play_quantile	2.294	0.346	0.5470875	0.0199097
kgamma	3.080	0.194	0.5418150	0.0173614
kt5	3.605	0.000	0.5393409	0.0183805
klaplace	5.024	0.000	0.5333888	0.0181580
epa	5.348	0.051	0.5205935	0.0174090

Simulation Study

One measure of the effectiveness of EPA has been how well correlated it is with itself on a per play basis for values intra-season or the next season. This is sometimes referred to as **stability**. The thought process is that if a trait or metric is stable for a team or player, it reflects somewhat a measure of skill or ability. This of course isn't always true as any arbitrary value would have a correlation of 1 with itself and human scouting bias may result in a higher stability since a human knows who they are watching.

Intra-season Stability

I wanted to see if any of these metrics perform better than EPA for both quarterbacks and for team offenses and defenses. For intra-season comparison, I perform a bootstrap simulation by sampling 8 weeks of the season and then comparing those 8 weeks per play numbers to the per play numbers for the remaining weeks.

Quarterbacks

Offenses

Defenses

stat	avg_rank	pct_no_1	avg_correlation	sd_correlation
knorm	1.649	0.409	0.5486108	0.0188296
play_quantile	2.294	0.346	0.5470875	0.0199097
kgamma	3.080	0.194	0.5418150	0.0173614
kt5	3.605	0.000	0.5393409	0.0183805
klaplace	5.024	0.000	0.5333888	0.0181580
epa	5.348	0.051	0.5205935	0.0174090

stat	avg_rank	pct_no_1	avg_correlation	sd_correlation
epa	1.536	0.744	0.4198002	0.0522222
adj_nya	4.642	0.033	0.3685498	0.0531352
wpa	4.877	0.114	0.3653028	0.0557913
klaplace	4.243	0.029	0.3606291	0.0591673
kt5	5.212	0.007	0.3559117	0.0598880
kgamma	5.191	0.027	0.3541468	0.0595235
yards_gained	5.781	0.026	0.3515248	0.0503577
knorm	5.938	0.006	0.3508226	0.0612303
play_quantile	7.580	0.014	0.3331898	0.0633355

Inter-season Stability

For year to year comparison, I perform a bootstrap where I sample 300 unique player x season or team x play type (run/pass) x season in each bootstrap. Then I calculate the Pearson Correlation Coefficent of the previous year's value to this seasons value. I perform 1000 bootstrap samples.

Quarterbacks

Offenses

Defenses

stat	avg_rank	pct_no_1	avg_correlation	sd_correlation
epa	2.250	0.663	0.4506052	0.0696085
knorm	3.110	0.042	0.4151039	0.0656606
play_quantile	3.440	0.166	0.4141272	0.0659080
kt5	3.729	0.020	0.4103262	0.0660543
kgamma	3.946	0.095	0.4079203	0.0648523
klaplace	4.525	0.014	0.4061664	0.0662686

stat	avg_rank	pct_no_1	avg_correlation	sd_correlation
epa	1.815	0.690	0.4229399	0.0698401
kgamma	2.567	0.188	0.3832876	0.0675689
play_quantile	3.330	0.116	0.3641721	0.0679347
knorm	3.686	0.002	0.3568953	0.0677368
kt5	4.322	0.002	0.3494288	0.0681030
klaplace	5.280	0.002	0.3421529	0.0685885

Discussion

For intra-season metrics, the standard normal transformed values tend to correlate with the other half of the season the best while EPA does quite poorly.

For season to season it tends to be almost the opposite where EPA has a higher stability than the others. This may be because EPA does best with larger sample sizes but struggles with smaller samples due to the "cliffs" that exist around the first down line.

When evaluating teams on the season EPA is still a very good metric, but for players perhaps there is something better. If we want to evaluate the value of a player on a play, should we really discount them several points because they gained 9 yards on 3rd and 10 instead of 10? Yes, 10 yards is worth a lot more, but the player got most of the way there and there is a lot (but maybe not complete) randomness between getting 9 yards and 10.

Another idea is to model directly the distribution of EPA on a play via mixture of normals, or a distributional Random Forest.

The idea ultimately is to be able to compare two plays in two completely different contexts and know which result was more impressive given expectation (accounting for play call such as run or pass) in a predictive sense. Otherwise using any play outcome to evaluate players who participated in that play is going to suffer from some bias of players being used in certain settings, some of which are more likely to have bigger plays than others (or vice-versa) and none of it being due to the ability (or lack thereof) of the player.