Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

SCORIGAMI: SIMULATING THE DISTRIBUTION AND ASSESSING THE RARITY OF NATIONAL FOOTBALL LEAGUE SCORES

Liam Moyer[1], Jameson Railey[2], Andrew Daw[3], and Samuel C. Gutekunst[4]

[1]Dept. of Computer Science, Bucknell University, Lewisburg, PA, USA
[2]Dept. of Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA
[3]Dept. of Data Sciences and Operations, University of Southern California, Los Angeles, CA, USA
[4]Depts. of Computer Science and Mathematics, Bucknell University, Lewisburg, PA, USA

ABSTRACT

NFL Scorigamis have a cult-like following, and occur whenever a game ends in a new, never-before-seen score. While substantial research has gone into simulating and predicting NFL game scores, most work has been in relation to winner prediction and betting spreads. We analyze a Poisson random variable model for the distribution of NFL game scores and show that it fails to incorporate important game dynamics. Through an analysis of extensive play-by-play data, we extend this to a non-stationary, state-dependent Poisson process model. This latter model more closely fits real NFL score data, and we use it in NFL score simulations to forecast likely future Scorigamis.

1 INTRODUCTION

Over the past decade, National Football League (NFL) *Scorigamis* have attracted a sizeable fanbase. A Scorigami occurs in the NFL whenever a game ends in a new score that has not previously happened in the league's history. Most recently, on 12/31/2023, an NFL game ended 56-19 (Baltimore vs. Miami). Since that score had never occurred in the 100+ year history of the NFL, it was dubbed a Scorigami and reported as such; see Breech (2023), among others.

In many sports, scores that are "numerically close" occur with similar frequencies. For instance, of the 64 games in the 2022 World Cup, the most frequent final game score (not accounting for penalty kicks) was 2-0; 12 games ended 2-0. The next two most frequent final scores were each one goal away: 2-1 (occurring 11 times) and 1-0 (occurring 10 times). In contrast, American football's scoring distribution is decidedly less regular. The most frequent NFL final score is 20-17, with 290 professional football games (of 17,665 in history) ending 20-17. However, just eight games have ended 20-18, two games have ended 19-18, and zero games have ended 18-18. Similarly, 177 different games have ended with a 24-17 score; 25-18 has never occurred. See Figure 1 for the full distribution of final scores.

The term *Scorigami* was coined by sports journalist Jon Bois. The concept rose to prominence in a 2016 YouTube video (Bois 2016), and it quickly took on a cult-like following: the original YouTube video now has almost 4 million views, the @NFL_Scorigami X (formerly Twitter) account has over 450,000 followers, the nine new Scorigamis in the 2023-24 season were each reported by press, and DraftKings even offered a prop bet on the Super Bowl ending in a Scorigami. Jon Bois used the Scorigami distribution to identify a "good game" region of competitive games with a good balance of scoring events. The Scorigami concept has since been extended to baseball (Bennett 2018) and even weather via daily minimum/maximum temperatures (Kahl 2023). "Scorigami" itself is a portmanteau of "score" and "origami," which Bois proposed because of the way the sport's unique point values fold together to create the disparate rarity among game outcomes that might otherwise appear similar. In the NFL, a team can only score 1 point if they had scored 6 points immediately before; the other base scoring events are worth 2 and 3 points. Hence, NFL scores are typified by strange sums.

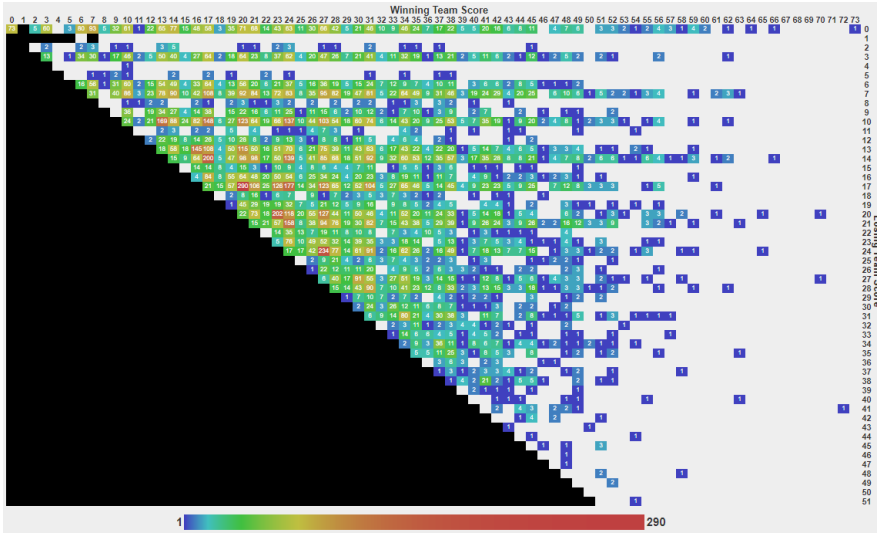


Figure 1: The distribution of NFL final scores (by winning and losing team's score) from the 1920 season through the 2023 season, from https://nflscorigami.com/. The winning team's score corresponds to the column, and the losing team's score, the row.

There has been some simulation work modeling American football scores and games. In general, however, previous work has not been focused on understanding the exact distribution of scores. Wilson (2005) simulates college football seasons for use in analyzing methods of ranking teams; to simulate game outcomes, they draw a *score differential* (the winning team's score minus the losing team's score, sometimes called margin of victory) from a Normal distribution. Glickman and Stern (1998), Warner (2010), Blanc et al. (2016), and Mohsin and Gebhardt (2024) similarly aim to predict score differentials, typically as they pertain to sports betting markets, using simulation and machine learning. Baker and McHale (2013) predicts exact NFL scores, but is again motivated by (and primarily benchmarked against) outperforming betting markets for predicting game results. In these applications, the exact score distribution is generally irrelevant or not part of how the simulations are benchmarked. In the realm of popular work, in 2017 Dave Mattingly created the @NFL_Scorigami account to track Scorigamis. Based on a game's current score and remaining time, it also estimates the probability that games ends in a Scorigami. Merriman (2024) created a Scorigami GitHub repository, which cites that the prediction algorithms implemented are by Mattingly. While the model is not precisely defined, the code appears to model each NFL game by computing independent scores for each team, where each team's score is based on five mutually independent Poisson processes (one for each possible scoring event).

In this paper, we focus on the distribution of NFL scores. Our primary motivation is the increasingly-pervasive interest in NFL Scorigamis: what Scorigamis are most likely to occur next, how likely is a game to end in a Scorigami, and how often should we expect to see Scorigamis? Since Scorigamis originated as a novelty fan concept, this pursuit may – at first glance – seem more niche and academic than operational. There is, however, a case to be made that Scorigamis are a proxy for the NFL's core business: entertainment (E.g., in 1940 the NFL Rules Committee wrote "each game should provide a maximum of entertainment insofar as it can be controlled by the rules and officials." See, e.g., NFL Football Operations (2024).) It is unlikely that league decision makers have considered Scorigamis when making updates to the game rules; nevertheless, Figure 2 shows that major changes to the rules have been preceded by low rates of Scorigamis. Inherently, Scorigamis (or the lack thereof) measure the freshness (staleness) of scores: when Scorigamis become increasingly rare, game outcomes are by definition more routine. Hence, an additional consequence of our simulation study is a manner by which to measure the future novelty of the game.

Table 1: Scorigami forecasts by model with the probability of being the next new Scorigami, given that a Scorigami occurs. As of the end of the 2023-2024 NFL season, the probabilities of a Scorigami occurring are 5.58% and 5.52%, respectively, according to the random variable and point process models.

Rank r	r-th Most Likely Score: Poisson R.V. Model	Conditional Probability	r-th Most Likely Score: N.S. Poisson Process Model	Conditional Probability
1	36-23	1.759%	32-26	1.737%
2	32-26	1.437%	36-23	1.394%
3	40-31	1.109%	40-31	1.279%
4	47-20	1.015%	36-29	1.203%
5	39-16	1.014%	38-18	1.126%
6	32-19	0.952%	25-18	1.012%
7	40-19	0.916%	32-22	0.974%
8	46-17	0.807%	18-9	0.897%
9	46-20	0.776%	41-40	0.897%
10	43-30	0.721%	20-11	0.879%

Our main contributions are twofold. First, we study a model based on Poisson random variables in the same spirit as prior reasoning about Scorigami. While this model is amenable to calculations and can be used to analytically compute the probability (according to the model) that a game ends in a certain score, it does not accurately reflect the distribution of NFL scores. For instance, it substantially underpredicts competitive games where the winning team's score is a 7 point touchdown or a 3 point field goal higher than the losing team's. The Poisson model also does not account for many features of the game that empirically effect score distributions: differing team strength, strategic decision making based on the score differential, and non-stationarity of scoring rates as time passes in the game. It is not tractable to add these features to an analytical model, so our second main contribution is to provide and analyze a new simulation-based model incorporating these game dynamics. Our proposed model is similarly Poisson in nature, but we generalize from the random variable model to a point process representation that is both time-varying and state-dependent. We end by using this model to predict future Scorigamis and interpret the distribution of the NFL scores not yet seen. These results, which are also contrasted to the random variable model, are summarized in Table 1.

Extensive data is available on NFL games. The number of scoring events of each time (touchdown, field goal, etc) in a given game can be found on statmuse and Pro Football Reference. More detailed information on games since 1999 is available in the Python package nfl_data_py, which records play-by-play data for NFL games. This is a large data set. For example, nfl_data_py records 49,664 plays in just the 285-game 2023 season, with roughly 400 columns giving extremely detailed information about each play. This data is also readily available in other programming languages, such as in R, where open-source sampling methods are available (Williams et al. 2023). Our analysis and simulation studies are reliant upon these data sources, and our simulation code is available in the Github repo (Moyer 2024).

2 BACKGROUND

The current NFL rules allow the following scoring events. See Goodell (2023) for more details:

- A *touchdown* is worth 6 points, which is awarded for bringing the ball into the defending team's endzone. It is the most common scoring event. After scoring a successful touchdown, a team has the opportunity for a *Try* to score 1-2 additional points: they can attempt to kick a field goal at the 15-yard line (worth one point) or to score another touchdown starting at the 2-yard line (worth two points). If they fail at either attempt, they only get the 6 points from a touchdown. A *defensive two point conversion* is an exceptionally rare event allowed since the 2015 season. This occurs when the defending team forces a turnover on a Try and then effectively scores a touchdown in

the opposite endzone 100 yards away; in this case, the team scoring the original touchdown gets 6 points and the defending team gets 2 points. There have been 12 defensive two point conversions in the 9 seasons since the introduction of the rule. (In addition to the simultaneous $(6,2)$ points awarded by the defensive two point conversion to the respective scoring and defending teams, it is also possible, by rule, for a $(6,1)$ outcome to be awarded. Essentially, this would happen when the offense attempting a Try is tackled in their own endzone at the opposite end of the field, much like a safety. It has never occurred in over a century of NFL seasons, and thus it is absent from both our data analysis and simulation study.)

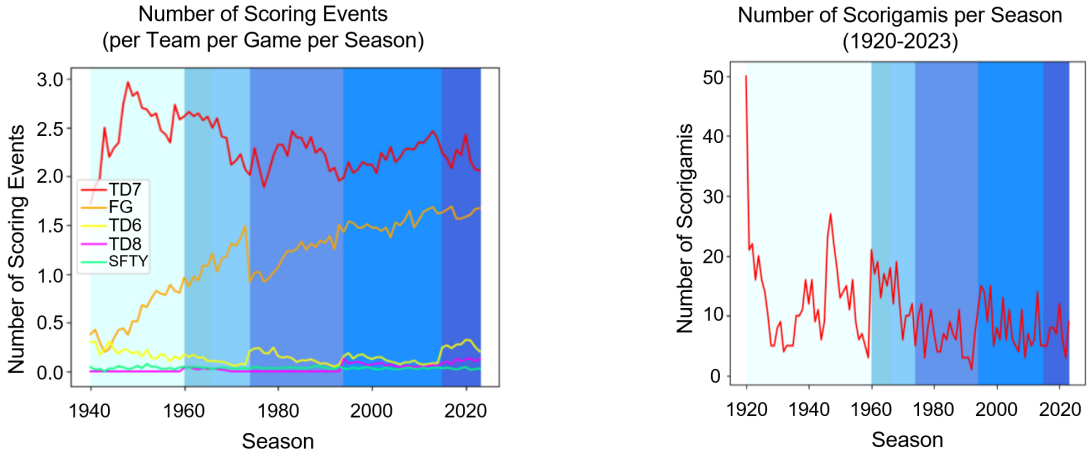
- A *field goal* is worth 3 points, which is awarded for successfully kicking the ball through an upright frame at the end of the field, beyond the endzone.
- A *safety* is worth 2 points, which is awarded to the defending team upon tackling the team currently possessing the ball inside their own endzone.

The game starts with one team on offense, where they attempt to score a field goal or touchdown, and the other on defense. The offense team performs a series of plays, either leading to them scoring a field goal or touchdown, or them failing to score and the other team switching to offense. This series of plays (whether successful or not) is referred to as a *drive*. Normally these scoring events occur during 60 minutes of play, split into four 15-minute quarters. If game is tied at the end of the fourth quarter, the game enters overtime.

The exact rules of overtime have changed over time. As of the 2023-24 season, the rules are different for the regular season (where the game can end in a tie) and the postseason (playoff and Superbowl games, where it cannot). In both cases, a coin is flipped to decide which team starts overtime on offense. During the regular season, there are up to 10 minutes of overtime. If the team that starts on offense scores a touchdown (or the other team scores a safety) on that first drive, the game ends after that scoring event. If they do not score on their first possession, the next team to get any score wins. If they score a field goal on their first possession, the other team gets one possession in which they win if they score a touchdown, they lose if they do not score anything, and the game enters sudden death if they score a field goal (where the next scoring event wins). The game ends after 10 minutes of overtime, even if both teams have not had a full possession or the game is tied, and in no case is there a try after a touchdown. During the postseason, if the team defending on the first possession scores a safety, that team automatically wins. Otherwise, both teams are guaranteed one possession. After both teams have one possession, the game ends as soon as one team has a higher score than the other, and until that point, overtime continues in 15 minute increments.

The amount of scoring events per game has changed over time, as both the game and rules have evolved. Figure 2a shows the number of scoring events per team, per game each season from 1940 to 2023. (We could not find reliable scoring event data for games from 1920-1939.) Field goal rates generally increased until 1974, when the NFL moved the goals posts 10 yards back (from the goal line to the end line), adding 10 yards to the distance required to kick any field goal. Before the 2015 season, the NFL changed the rules for a one-point Try: players previously attempted to kick the field goal from the 2-yard line but now do so from the 15-yard line. This made the 1-point try field goal more difficult: in 2014, 99.3% of one-point Trys were successful, whereas only 94.2% of one-point Trys were successful in 2019. This rule change increased the relative rate of two-point conversion attempts.

Figure 2a is divided into six epochs of time, based on major rule changes effecting the state of play. The first epoch began in 1920 with the formation of the NFL and extends through the 1959 season. Note that a rival league (the All-America Football Conference, AAFC) had its first season in 1946, and merged with the NFL in 1949; AAFC games from 1946-49 still count towards official Scorigami counts, as do games from another NFL Rival – the American Football League – which had its first season in 1960 and merged with the NFL in 1966. The second epoch is from the 1960 season through the 1965 season, when the AFL and NFL operated independently; the AFL allowed 2 point conversions, while the NFL did not, and the new scoring rule led to a spike in Scorigamis. The third epoch ranges from the 1966 season through the 1973 season, stopping in 1974 when the goalposts for field goals were moved 10 yards back and overtime rules changed. The fourth epoch spans from the 1974 season until the 1993 season: as mentioned above, in



(a) Average number of scoring events per team, per game each season by event.

(b) Number of new Scorigamis per season.

Figure 2: Scoring events and Scorigamis over time. Differently-colored regions indicate major rule changes.

1994, the distance for a successful one-point Try kick was moved back 13 yards, and the 2 point conversion was reintroduced to the NFL (it had previously only been present in the AFL). The fifth epoch spans from the 1994 season to the 2014 season, and the final current epoch spans from the 2015 season (when the kick distance for a one-point try was increased) through the most recent 2023 season. Figure 2b shows the number of Scorigami's per season, based on game data from Sports Reference LLC. (2024). Major spikes in the number of new Scorigamis occur with rule changes/mergers in 1960, 1994, and 2015. Throughout this paper, we use data and benchmark against scores from the *current epoch* (seasons 2015-2023).

During the current epoch, each team has scored an average of 2.56 touchdowns per game (during the 60 minutes of regular time play). Of these, 10.08% had a failed Try, 85.46% had a successful one-point Try, and 4.36% had a successful two-point try (and 0.095%, a defensive two point conversion). There were an average of 1.61 field goals per team per game before overtime, and an average of 0.028 safeties per team per game before overtime. Defensive two point safeties were extremely rare, with 0.0024 per team per game (during regular time). Thus, the most common scoring events contribute 7 and 3 points. The most frequent scores are generally small integer combinations of 7 and 3: teams most frequently ended with 20, 17, 24, and 27, points occurring 300, 279, 271, and 262 times, respectively.

3 PROBABILITY MODELS FOR SCORES IN NFL GAMES

3.1 A Poisson Random Variable Model

We now consider a first model, similar to Merriman (2024)'s implementation of Mattingly's Scorigami prediction algorithm. This simple probability model for NFL game scores, amenable to analytic calculation, assumes each team's score is independent and that the number of underlying scoring events are mutually independent and drawn from Poisson distributions. Such a model proceeds as follows.

Let λ_6, λ_7, and λ_8 respectively denote the average number of 6-point, 7-point, and 8-point touchdowns per team per game. Let λ_3 denote the mean rate of field goals per team per game, and λ_2 denotes the mean rate of safeties and defensive two point conversions (all before overtime). Here, λ_6 denotes the rate of touchdowns where either the 1- or 2-point Try is attempted but failed (and thus includes touchdowns that lead to defensive two point conversions). (Note that this model does not directly account for the joint scoring associated with defensive two-point conversions, since that creates dependence between teams. Instead, the model imposes artificial independence by treating the $(6,2)$ touchdown-with-defensive-two-point events as two distinct and not necessarily coincident events: a 6-point touchdown and a 2-point safety.) Then, for

Table 2: Overtime scores and frequencies across all games from the 2015-2023 seasons.

Overtime Scores	3-0	6-0	6-3	0-0	3-3	2-0
Total Occurrences	74	49	13	7	2	0

each $i \in I = \{2,3,6,7,8\}$ and $j \in \{1,2\}$, let $X_{i,j}$ be mutually independent Poisson(λ_i) random variables for the number of scores worth point i scored by team j in a given game (before overtime). Then, the total score for team j during regular time is given by the weighted sum of the Poisson random variables and their associated points: $T_j = \sum_{i \in I} i \cdot X_{i,j}$. If $T_1 \neq T_2$, then the ordered pair, $(T_{(1)}, T_{(2)})$ with $T_{(1)} > T_{(2)}$, is the final score outcome of the game. However, if $T_1 = T_2$, then the game heads to overtime, and the pair of random variables will be further updated.

The possible additional points in overtime can be succinctly enumerated. In fact, in the current NFL rules as of the 2015 season, there are only six possible scoring pairs that can occur. Table 2 shows the counts of overtime scoring events that have happened from 2015-2023. Let $OT = \{(0,0),(2,0),(3,0),(3,3),(6,0),(6,3)\}$ be the set of possible overtime scoring outcomes. For $(i,j) \in OT$, let $o_{i,j}$ be the probability that an overtime game ends with (i,j) points added to the final score (e.g., based on Table 2), and let T^O be a random variable with the $o_{i,j}$ probabilities on the OT sample space. T^O is mutually independent with T_1, T_2, and the underlying $X_{i,j}$ random variables. Then, if $T_1 = T_2$, the final score of the game will be $(T_1, T_2) + T^O$, with the sum of pairs taken to be element-wise. In general, letting $\mathbf{1}\{\cdot\}$ be the indicator function, the final score (F_1, F_2) in this Poisson random variable model is

$$(F_1, F_2) = (T_1, T_2) + T^O \cdot \mathbf{1}\{T_1 = T_2\}.$$

While this model can be easily simulated, it is also tractable to compute the distribution numerically. First, to compute $\mathrm{P}(T_j = k)$ for team j and a number of regular time points $k \in \mathbb{N}$ (the natural numbers), we sum over all non-negative integer combinations of 2, 3, 6, 7, and 8 that add to k. For instance, there are 3363 such combinations that add up to 100, so to compute $\mathrm{P}(T_1 = 100) = \mathrm{P}(T_2 = 100)$, we would add the probabilities of the 3363 combinations of scoring events that would lead to a team scoring 100 points total in regular time. Formally,

$$\mathrm{P}(T_j = k) = \sum_{\substack{i_2,i_3,i_6,i_7,i_8 \in \mathbb{N} \\ 2i_2+3i_3+6i_6+7i_7+8i_8=k}} \mathrm{P}(X_{2,j} = i_2) \cdot \mathrm{P}(X_{3,j} = i_3) \cdot \mathrm{P}\left(X_{6,j} = i_6\right) \cdot \mathrm{P}(X_{7,j} = i_7) \cdot \mathrm{P}(X_{8,j} = i_8).$$

Notice that $\mathrm{P}(X_{i,j} = x)$ is simply the probability mass function (PMF) of a Poisson random variable with mean λ_i. We call the set we sum over

$$\{(i_2,i_3,i_6,i_7,i_8) \text{ such that: } i_2,i_3,i_6,i_7,i_8 \in \mathbb{N}, 2i_2+3i_3+6i_6+7i_7+8i_8 = k\}$$

the set of *score partitions of* k. The set is known to grow in size quasi-polynomially (Baldoni et al. 2013).

It is straightforward to compute all score partitions of k for reasonable scores (say, up to $k = 200$) by brute force. Then, for $w, \ell \in \mathbb{N}$, the final game score probabilities can be computed as follows:

$$\mathrm{P}(F_1 = w, F_2 = \ell) = \begin{cases} 2 \cdot \mathrm{P}(T_1 = w) \cdot \mathrm{P}(T_1 = \ell) + \displaystyle\sum_{\substack{t \in OT: \\ w - t_1 = \ell - t_2}} \mathrm{P}(T_1 = w - t_1)^2 \cdot o_t, & w \neq \ell, \\ \displaystyle\sum_{\substack{t \in OT: \\ t_1 = t_2}} \mathrm{P}(T_1 = w - t_1)^2 \cdot o_t, & w = \ell. \end{cases} \tag{1}$$

Notice that throughout Equation (1) we have used only the PMF of T_1 in the notation, which is without loss of generality due to the presumed independence and identicality of the teams. For example, a score of 21-18 can happen if the game ends 21-18 in regular time, or if the game ends 18-18 in regular time and

scoring event (3, 0) happens in overtime, or if a game ends 15-15 in regular time and scoring event (6, 3) happens in overtime. Thus,

$$\mathrm{P}(F_1 = 21, F_2 = 18) = 2 \cdot \mathrm{P}(T_1 = 21) \cdot \mathrm{P}(T_2 = 18) + \mathrm{P}(T_1 = 18)^2 \cdot o_{3,0} + \mathrm{P}(T_1 = 15)^2 \cdot o_{6,3}.$$

3.2 Data for Time-Varying and State-Dependent Game Phenomena

We now evaluate the model and computing structure. Figure 3 shows the rates of each type of scoring event for games from 2015-2023, fit with a Poisson distribution having the same mean. From the perspective of these marginal distributions, the Poisson random variable model seems well-motivated and likely to succeed.

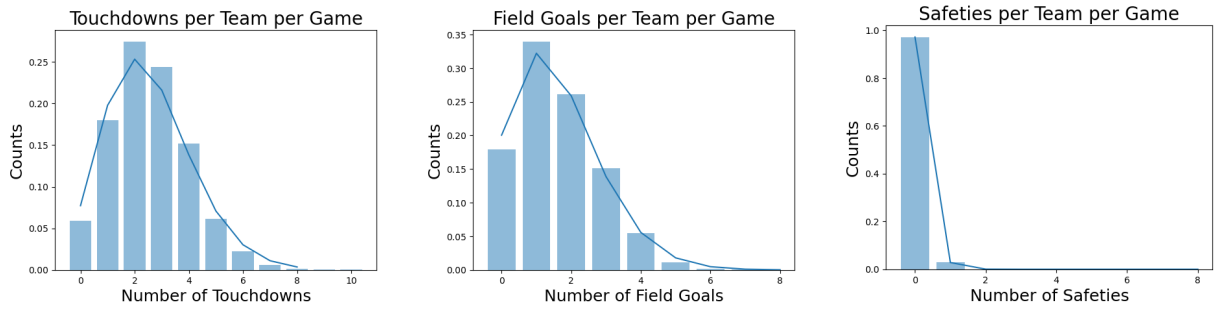
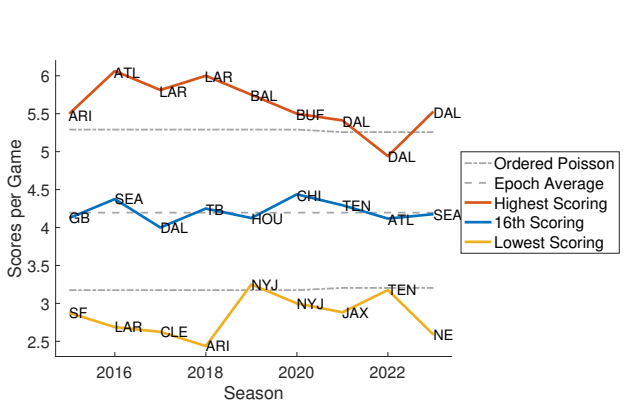


Figure 3: Scoring events per team per game (before overtime), by scoring event. The bars show the true distribution of data, and the line show a Poisson distribution having the same mean.

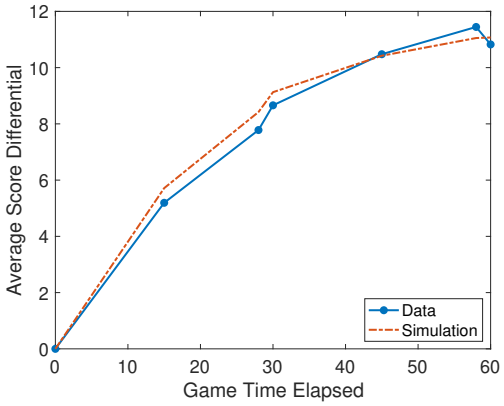
However, if we look more closely, we find that the fit seen at this score level may be a facade. Figure 7 benchmarks the final scores from the Poisson random variable model against those from the current epoch by comparing the distributions of score differentials. While the shapes are generally similar, they differ in key ways. For instance, 14.7% of NFL games in the current epoch ended with a score differential of 3, while the this model predicts that only 7.67% of games would end with this score differential; 8.89% of NFL games in the current epoch ended with a score differential of 7, while the Poisson model predicts that only 5.67% of games would end with this score differential. Thus, the Poisson model underpredicts "competitive games," where the game is decided by a single common scoring event. Consequently, this Poisson model leads to an inflated score differential: the mean score differential across games in the current epoch is 11.11, while the Poisson random variable model leads to a mean score differential of 13.32.

While this model is amenable to analytic calculation, it does not incorporate several key factors of the game. We focus on three that motivate our point process simulation model. First, it assumes that all teams are equally skilled. In practice, teams have different strengths, even over large periods of time. Figure 4a shows that, although the middle-of-the-pack teams are quite close to the model average, there is a significant difference among the top and bottom teams in the league, and that this difference exceeds what would be expected by the corresponding ordered statistics of the underlying Poisson random variables. The orange and yellow lines in Figure 4a show the number of scoring events per game by the teams that score the most and least, respectively, for each season in the current epoch. The gray dash-dot lines show the expected minimum and maximum values among 32 independent samples from a season's worth of independent and identically distributed (i.i.d.) Poisson random variables with mean $\lambda_2 + \lambda_3 + \lambda_6 + \lambda_7 + \lambda_8$. The orange and yellow lines almost are each more extreme than their corresponding model values for all but one season. Hence, this model of independent and identical teams is under-estimating the true spread, and, if a model is to be based on Poisson components, *there must be heterogeneity amongst teams.*

Furthermore, Figure 4b (which also includes simulation results that will be discussed in the following section) shows how the score differential evolves during the game: At the start of the game, scores are tied. By the end of the first quarter, one team is, on average, ahead by 5.19 points. The score differential grows

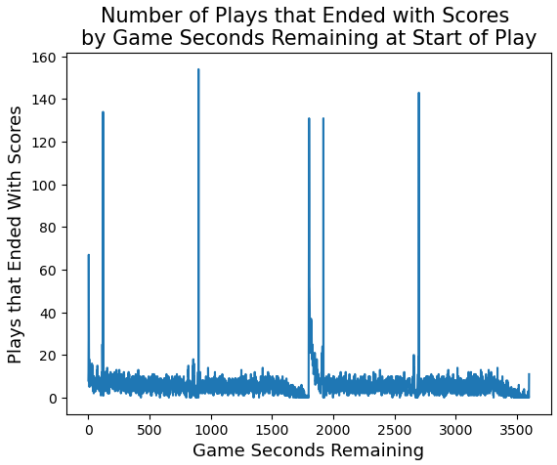


(a) Scoring rates per game for the best, middle, and worst teams each season.

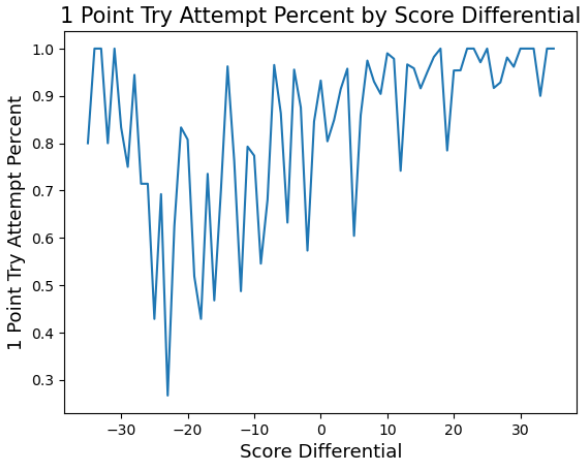


(b) Scoring differential by time remaining at the start of each period.

Figure 4: Scoring rates by season and in-game scoring differential.



(a) Scoring events vs seconds remaining.



(b) Try attempt strategy by score differential.

Figure 5: Game time and score differential dynamics.

less quickly during the second half, and even drops in the last two minutes of the game. For example, a team that is winning by a substantial amount may not play as competitively in the last two minutes where there is little risk, while a team that is losing may push for one final scoring event to close the gap and lose by a smaller margin. This suggests that the effects of team strength become less pronounced throughout the game. Such detail is missed by a model that only samples scores at the very end of the game.

We can also see that scoring event rates change throughout the game. Figure 5a shows how many scoring events happened in the current epoch based on the seconds remaining of regular time in the game when the play started. There is a small spike at 3,600, corresponding to games where a scoring event happened on the opening kick off or very first play (and whereas every game may not have a play starting with 3,599 seconds remaining, every game has a play starting at 3600 seconds). A larger spike occurs with 2700 seconds remaining, on the first play of the second quarter; this is larger than the spike with 3600 seconds remaining because teams tend to be closer to their scoring zone at the start of the second quarter (in the middle of their possession period) than at the first. Final spikes in the first half occur with 1920

seconds remaining, at the two minute warning, and right at the end of the first half. After the two minute warning, scoring rates are generally higher: possession switches at the start of the second half (regardless of how close a team is to scoring at the end of the first half), and the team on offense may attempt a field goal to end their possession with a score if a touchdown seems out of reach. Similar patterns can be seen in the second half. Because of the designed stoppage of time and strategic clock management, scoring rates should be non-stationary across game time.

Finally, a team may also make strategic decisions based on the current score differential. This is particularly relevant after a touchdown, when the team decides to go for a 1 or 2 point try. Figure 5b shows how a team's choice depends on their score differential (their score minus their opponents score, at the time of the decision after getting 6 points from the touchdown). A team is substantially more likely to go for a 2 point try, e.g., when they are behind in score. Hence, a model that treats Try outcomes as independent events is missing a crucial strategic detail impacting the exact distribution of scores.

3.3 A Non-Stationary, State-Dependent Poisson Process Model

We will now define a model that is effectively a Poisson process model rather than a random variable model. In fact, each team's score will be composed of six different underlying Poisson processes. Furthermore, these stochastic processes will be non-stationary and state-dependent, meaning that the overall model's underlying scoring rates will update throughout the game as time passes and as scoring occurs.

These models will take a collection of input parameters. First, for the base scoring event set $E = \{\sigma, \phi, \tau\}$ (where these elements respectively represent safety, field goal, and touchdown, with all Try outcomes includes in this category) and the scoring-rate-rank index set $r \in \{1, \ldots, 32\}$, let $\mu_r^e(t)$ be the rate of scoring at time $t \in [0, 60]$ having elapsed in the game. Following the observations from Figure 5a and to enable efficient estimation from data, we will assume this function is piecewise constant between the change points in the list $\mathscr{T} = (0, 15, 28, 30, 45, 58, 60)$. Respectively, these times mark the start of the first quarter, start of the second quarter, two minute warning before half, start of the third quarter, start of fourth quarter, two minute warning before the end of regulation, and the end of regular time. With slight abuse in notation, we will let $\mathscr{T}(t) = \sup\{z \in \mathscr{T} \mid t \geq z\}$. Then, let us also define the additional parameters $\gamma, \Delta > 0$, which will be used to account for the score differential effects.

For convenience in notation, let us additionally define the shorthand $\mu(t) = \frac{1}{32}\sum_{r=1}^{32}\sum_{e\in E}\mu_r^e(t)$, which is the average scoring (accounting for all event types) rate at time t, averaged over teams. Furthermore, let $\mu^e(t) = \frac{1}{32}\sum_{r=1}^{32}\mu_r^e(t)$ be the average scoring rate for each event type e at time t per team.

Now, we consider the stochastic model. Updating the notation from the random variable model, we will construct the pair of scores at time t elapsed in regulation, $T_1(t)$ and $T_2(t)$. We will assume that two distinct ranks, $r_1, r_2 \in \{1, \ldots, 32\}$, are drawn uniformly at random. For $e \in \{\sigma, \phi\}$, let us first define the non-stationary Poisson processes $N_1^e(t)$ and $N_2^e(t)$ with time-varying intensities given by

$$\hat{\mu}_1^e(t) = \bar{w}_{r_1}(t, T_1(t), T_2(t)) \cdot \mu^e(t),$$
$$\hat{\mu}_2^e(t) = \bar{w}_{r_2}(t, T_1(t), T_2(t)) \cdot \mu^e(t).$$

Here, $\bar{w}_{r_1}(t, T_1(t), T_2(t)) = w_{r_1}(t, T_1(t), T_2(t)) \,/\, ((w_{r_1}(t, T_1(t), T_2(t)) + w_{r_2}(t, T_1(t), T_2(t)))$ for weight functions defined

$$w_{r_1}(t, T_1(t), T_2(t)) = \begin{cases} e^{-\gamma\mathscr{T}(t)} \cdot \left(\sum_{e\in E}\mu_{r_1}^e(t)\right) + (1 - e^{-\gamma\mathscr{T}(t)}) \cdot \mu(t) \cdot (1-\Delta), & T_1(t) > T_2(t) \\ e^{-\gamma\mathscr{T}(t)} \cdot \left(\sum_{e\in E}\mu_{r_1}^e(t)\right) + (1 - e^{-\gamma\mathscr{T}(t)}) \cdot \mu(t), & T_1(t) = T_2(t) \\ e^{-\gamma\mathscr{T}(t)} \cdot \left(\sum_{e\in E}\mu_{r_1}^e(t)\right) + (1 - e^{-\gamma\mathscr{T}(t)}) \cdot \mu(t) \cdot (1+\Delta), & T_1(t) < T_2(t) \end{cases} \tag{2}$$

and likewise for $w_{r_2}(t, T_1(t), T_2(t))$. These intensities maintain a consistent overall scoring rate in each time frame for the safety and field goal scoring events, but shift the weight of which team is more likely to be scoring based on their ranks, the current score, and the time remaining. In particular, notice that

Equation (2) has the two weights regress away from the specific team strength weights and towards a score differential-based weight as time progresses. Hence, at the beginning of the game, the model respects the inherent strengths of each team, but late in the game, the model favors the team with the lower score (regardless of rank), which may capture either the spirit of competition or so-called "garbage time" scoring.

Let us now model touchdowns in a similar fashion but with one added level of detail. With $\theta(x \mid T_1(t), T_2(t))$ as the distribution of having Try outcome $x \in \{6,7,8,2\}$ (denoting no points, 1 added point, 2 added points, and a defensive 2-point conversion added to the touchdown) given the current score differential, $|T_1(T) - T_2(t)|$, let $N_1^{\tau,x}(t)$ and $N_2^{\tau,x}(t)$ be defined with intensities

$$\hat{\mu}_1^{\tau,x}(t) = \bar{w}_{r_1}(t, T_1(t), T_2(t)) \cdot \bar{\mu}^{\tau}(t) \cdot \theta(x \mid T_1(t), T_2(t)),$$
$$\hat{\mu}_2^{\tau,x}(t) = \bar{w}_{r_2}(t, T_1(t), T_2(t)) \cdot \bar{\mu}^{\tau}(t) \cdot \theta(x \mid T_1(t), T_2(t)),$$

for each $x \in \{2,6,7,8\}$. Through these rates, we add the dependence on the score differential for the particular yield of the touchdown and its associated Try.

Together, these scoring-event Poisson processes let us now define the live-score processes

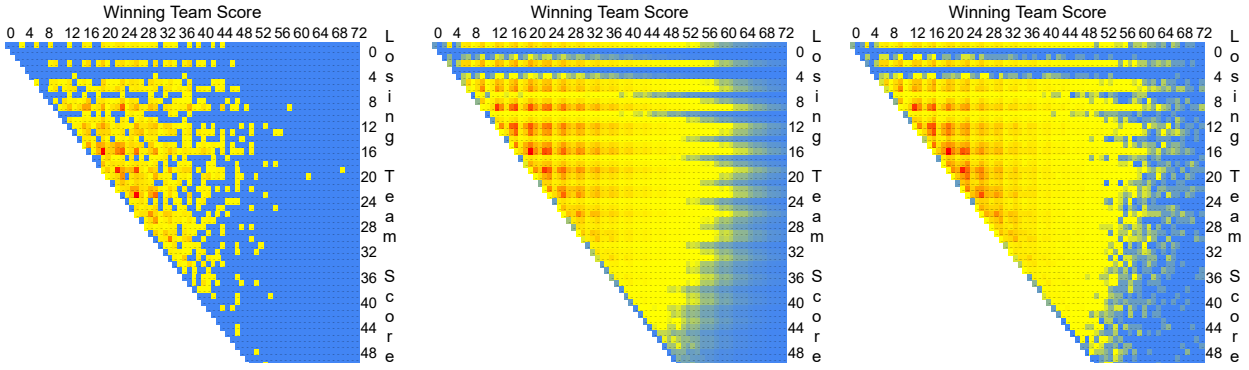
$$T_1(t) = 2\left(N_1^{\sigma}(t) + N_2^{\tau,2}(t)\right) + 3N_1^{\phi}(t) + 6\left(N_1^{\tau,6}(t) + N_1^{\tau,2}(t)\right) + 7N_1^{\tau,7}(t) + 8N_1^{\tau,8}(t),$$
$$T_2(t) = 2\left(N_2^{\sigma}(t) + N_1^{\tau,2}(t)\right) + 3N_2^{\phi}(t) + 6\left(N_2^{\tau,6}(t) + N_2^{\tau,2}(t)\right) + 7N_2^{\tau,7}(t) + 8N_2^{\tau,8}(t).$$

In addition to the multiple places that the rates depend on both scores, we can also clearly see that the live-scores are dependent on both teams underlying Poisson processes through the defensive two-point conversion. To complete the model, we use the overtime score random variable from the first model, defining the new random final scores as $(F_1, F_2) = (T_{(1)}(60), T_{(2)}(60)) + T^O \cdot \mathbf{1}\{T_1(60) = T_2(60)\}$.

4 SIMULATION RESULTS AND ANALYSIS

Naturally, this point process model is much more complex than the random variable model, and thus we evaluate it through simulation. We simulated 100,000 games using the above approach, with the underlying μ's estimated from the NFL play-by-play data. Additionally, we took $\gamma = 0.02$ and $\Delta = 0.4$. We choose these parameters after optimization to fit the mean score of a winning team (28.34) and mean score of the losing team (17.24) according to L_2 loss. Figure 6 compares Scorigami heatmaps of this non-stationary, state dependent model to the actual score distribution and to the original Poisson model. The non-stationary, state-dependent Poisson process model places more weight on competitive games close to the "middle diagonal" than the Poisson random variable model, mirroring the distribution of true NFL scores. The improved accuracy is also reflected in the comparison of root mean square errors (defined as $\left(\sum_{x,y}(P_M(x,y) - P_D(x,y))^2 \cdot P_D(x,y)\right)^{1/2}$ for $P_M(x,y)$ and $P_D(x,y)$ respectively as the model-based and data-based PMF at a given x,y score outcome). For the random variable model, this error is 0.00335, whereas the point process model error is 0.00310, which constitutes a 7.8% improvement.

Since we optimized our parameters to fit the mean winning and losing scores, it is not surprising that it produces game scores whose score differentials much more closely mirror reality: the mean score differential from the non-stationary, state-dependent Poisson process model is 11.25 (compared to the actual mean score differential of 11.11 of games in the current epoch, and 13.32 in the Poisson random variable model). Moreover, the non-stationary, state-dependent Poisson process model produces an output that allows us to see how scores and score differentials evolve over time, through the 6 major periods of an NFL game. The score differentials closely mirrors the diminishing effects of team strength seen throughout games; see Figure 4b. Finally, Figure 7 shows that this model more closely matches the overall distribution of score differentials. Interestingly, Figure 7a shows that the marginal winning team score distributions of both models are quite similar. While the fit to data could be improved in future work, we are pleased that, while tuning the point process model the match the mean of the winning score, we have actually found that it matches the random variable model beyond the mean.

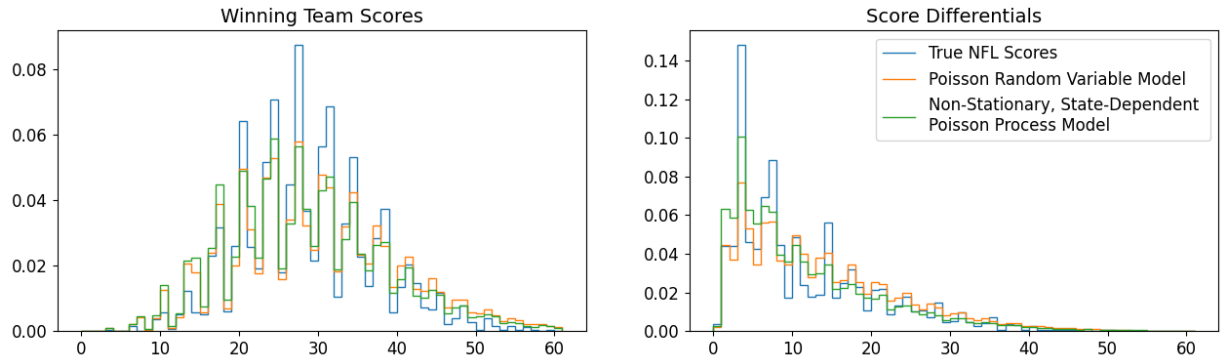


(a) True NFL scores (Seasons 2015-2023).

(b) Poisson random variable model (calculated).

(c) Non-stationary, state-dependent Poisson process model (simulated).

Figure 6: Heatmaps of estimated NFL score probabilities using our models compared to the distribution of scores from the current epoch. Blue indicates low probabilities/counts, while red indicates high.



(a) Distributions of winning team scores.

(b) Distributions of score differentials.

Figure 7: Comparing winning scores and score differentials across models.

5 CONCLUSION

Section 4 show that our proposed point process adds depth and nuance to the popular random variable model. Future work could incorporate more complex details and dependencies, like the actual real-time position and possession of the ball. Nevertheless, these simulation results demonstrate the value of explicitly incorporating new details, such as score differences, time inhomogeneity, and diversity of team strengths.

To close, let us return to Table 1, which shows the top 10 most likely future Scorigamis according to each model. These models predict different scores as the "next most likely Scorigami:" 36-23 in the Poisson random variable model, and 32-26 in the nonstationary, state-dependent Poisson process model. The estimated probabilities of these scores (conditional on a game ending in a Scorigami) in their respective models are 1.759% and 1.737%. Both models predict the same top three most likely future Scorigamis, just in different orders: 36-23, 32-26, and 40-31. Other than these three, however, the remaining "top 10" scores are completely disjoint. The Poisson random variable model generally predicts higher-scoring Scorigamis. We add that the point process model hedges its bets a bit more, with the conditional probabilities being flatter among its top 10 than in the random variable model. We look forward to seeing how these predictions hold up over the coming seasons.

ACKNOWLEDGMENTS

We thank the Hymas fund at Bucknell University for supporting undergraduate research on this project. We thank the reviewers for thoughtful feedback.

REFERENCES

Baker, R. D. and I. G. McHale. 2013. "Forecasting Exact Scores in National Football League Games". *International Journal of Forecasting* 29(1):122–130.

Baldoni, V., N. Berline, J. De Loera, B. Dutra, M. Koeppe and M. Vergne. 2013. "Coefficients of Sylvester's Denumerant". *arXiv preprint arXiv:1312.7147*.

Bennett, Dana. 2018. "MLB Scorigami". https://scorigami.danaben.net/, accessed 19th September.

Blanc, G., E. S. Luxenberg, and S. C. Xie. 2016. "NFL Score Difference Prediction with Markov Modeling". *Bachelor's Project. Stanford University*. https://cs229.stanford.edu/proj2016/report/BlancLuxenbergXie-NFLScoreDifferencePredictionWithMarkovModeling-report.pdf, accessed 19th September.

Bois, J. 2016. "Every NFL Score Ever | Chart Party". https://www.youtube.com/watch?v=9l5C8cGMueY, accessed 19th September.

Breech, J. 2023. "Ravens' Blowout Win Over Dolphins Ends with Final Score that's Never Been Seen Before in NFL History". https://www.cbssports.com/nfl/news/ravens-blowout-win-over-dolphins-ends-with-final-score-thats-never-been-seen-before-in-nfl-history/, accessed 19th September.

Glickman, M. E. and H. S. Stern. 1998. "A State-Space Model for National Football League Scores". *Journal of the American Statistical Association* 93(441):25.

Goodell, R. 2023. *2023 Official Playing Rules of the National Football League*. NFL. https://operations.nfl.com/media/tvglh0mx/2023-rulebook_final.pdf, accessed 19th September.

Kahl, J. D. 2023. "Weathergami". *Bulletin of the American Meteorological Society* 104(10):E1790–E1798.

Merriman, A. 2024. "Scorigami". https://github.com/Merry3750/scorigami, accessed 19th September.

Mohsin, M. and A. Gebhardt. 2024. "A Stochastic Model for NFL Games and Point Spread Assessment". *Journal of Applied Statistics* 51(2):216–229.

Moyer, L. 2024. "Scorigami Modeling". https://github.com/liam-moyer/Scorigami-Modeling, accessed 19th September.

NFL Football Operations 2024. "Bent but not Broken: The History of the Rules". https://operations.nfl.com/the-rules/evolution-of-the-nfl-rules/, accessed 19th September.

Sports Reference LLC. 2024. "NFL Score Data". https://www.pro-football-reference.com/, accessed 19th September.

Warner, J. 2010. "Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line". https://www.cs.cornell.edu/courses/cs6780/2010fa/projects/warner_cs6780.pdf, accessed 19th September.

Williams, B., W. Palmquist, and R. Elmore. 2023. "Simulation-Based Decision Making in the NFL using NFLSimulatoR". *Annals of Operations Research* 325(1):731–742.

Wilson, R. 2005. "Validating a Division I-A College Football Season Simulation System". In *2005 Winter Simulation Conference (WSC)*, 2437 – 2442 https://doi.org/10.1109/WSC.2005.1574536.

AUTHOR BIOGRAPHIES

LIAM MOYER is a senior computer science major at Bucknell University. He is interested in software engineering and data science, and his email address is lcm021@bucknell.edu.

JAMESON RAILEY is pursuing a Master of Science in Computer Science at Stevens Institute of Technology. Prior to Stevens, he was a Buisness Analytics major at Bucknell University. He was on the men's soccer team at both universities. His email address is jkr008@bucknell.edu.

ANDREW DAW holds a Dean's Assistant Professorship of Business Administration at the University of Southern California. His research is in applied probability, with an emphasis on history-driven models. For better or worse, he is a passionate fan of the Jacksonville Jaguars. His email address is andrew.daw@usc.edu, and his website is https://faculty.marshall.usc.edu/Andrew-Daw/.

SAMUEL C. GUTEKUNST is the John D. and Catherine T. MacArthur Assistant Professor of Data Science at Bucknell University. His research focuses on combinatorial optimization and data driven decision-making. His email address is s.gutekunst@bucknell.edu, and his website is https://samgutekunst.com/.