SMU Data Science Review

Volume 5 | Number 1 Article 8

2021

Analysis of Individual Player Performances and Their Effect on Winning in College Soccer

Angelo Bravo
Southern Methodist University, aebravo@smu.edu

Thomas Karba
Southern Methodist University, tkarba@smu.edu

Sean McWhirter
Southern Methodist University, sean.s.mcwhirter@gmail.com

Billy Nayden
Southern Methodist University, wnayden@sum.edu

Analysis of Individual Player Performances and Their Effect on Winning in College Soccer

Billy Nayden[1], Sean McWhirter[1], Thomas Karba[1], Angelo Bravo[1], Kevin Hudson[2], Tyler Heaps[3]
[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
[2] Men's Soccer, Southern Methodist University, Dallas, TX 75275 USA
[3] U.S. Soccer Federation, Chicago, IL 60616 USA
{wnayden, smcwhirter, tkarba, aebravo, khudson}@ smu.edu
theaps@ussoccer.org

Abstract. – This study describes the process of modernizing the approach of the Southern Methodist University (SMU) Men's Soccer coaching staff through the use of location and tracking data from their matches in the 2019 season. This study utilizes a variety of modeling and analysis techniques to explore and categorize the data and use it to evaluate the types of plays that are most often correlated with victories. This study's contribution to college soccer analytics includes the implementation of a model to determine individual players' performance, the production of team-level metrics, and visualizations to increase the efficiency of the coaching staff's efforts. This research can serve as a blueprint for college soccer programs to utilize data science in their coaching.

1 Introduction

While analytics in sports such as baseball and basketball have developed into cottage industries in the United States, the world of American soccer analytics remains relatively nascent in its progression. Some professional teams and the national teams have explored analytics concepts and now actively employ data science teams, but the college soccer world has lagged analytically. This study seeks to examine current cutting-edge concepts in soccer analytics, including models, metrics, and visualization, and apply them to SMU's data for their nationally ranked soccer team. This study's goal is to apply these concepts by the coaching staff to improve outcomes on the field.

This research is intended to bridge the gap between the work being done with European club teams, Major League Soccer (MLS) teams, and many national teams, with the world of college soccer. While college soccer has some analytics, they do not have the statistical resources necessary for the kind of research that has enabled many of the world's largest soccer programs to take the next step in their style of play and methods of coaching. However, there is access to event and location level data in college soccer, specifically with SMU, to craft the necessary models that can enact this type of change. The aim is to use this data with the existing statistical modeling

techniques available in the world of soccer and the Southern Methodist University data science curriculum to craft a model that will effectively measure the contributions each player makes to both scoring and preventing goals. In addition to the individual player contribution modeling, this study seeks to provide the SMU coaching staff and collegiate soccer in general with proven analytics and visualizations.

The history of soccer analytics can be traced to the 1950s when Charles Reep became upset at his team's seemingly weak scoring ability; this inspired him to begin recording trends and significant observations he noticed throughout games in a notebook (Luzum, 2019). Reep became employed as an analyst for several soccer teams for a large part of the 1950s until one of the teams he was advising took a turn for the worse. According to Luzum (2019), the main advice Reep would give to teams based on his analytical findings was to reduce passing and to encourage less long passes to propel the ball down the field; this was based on his observation that most goals are scored on plays which included less than four passes (Luzum, 2019). Overall, the advice he gave turned out to be flawed, as he would predominantly prefer watching the long-ball game, which gave him a preconceived bias that the long-ball strategy must be the best strategy. In further developments of soccer analytics, it has since been accepted that more passing is tied to more scoring (Luzum, 2019).

Since Reep's initial insights, soccer analytics has been overshadowed by other sports such as football and baseball. Sukumar et al. (2019) attribute soccer's lag in analytics to the generalized notion that due to the relatively low number of scoring events in soccer, players' effectiveness is mainly subjective. The ability to capture more data has rapidly changed that notion. In the past ten years, many studies have been conducted in the field of soccer analytics, thanks to this new data. This new data has allowed researchers to quantify aspects that have historically been unquantified, such as the acceleration index, which indexes a team's velocity at which the attacking team reaches the opponent's goal (Pappalardo et al., 2019). In addition, Huang (2018) and Whitaker et al. (2018) provided network and Bayesian theory-based approaches, creating stronger ties between the data science and soccer analytics communities. Some of the most promising works that tie many of the reviewed concepts together are the Expected Contribution to the Outcome of a Match (ECOM) rating from Bransen et al. (2019), a focal point of this study.

This study is limited to data provided by the sports analytics company InStat, which provides spatial and temporal data, but lacks data on the team not in possession of the ball. The lack of real-time opponent data limits this study's ability to classify the difficulty of actions. Passes are an example of this limitation. Generally, passes that bypass opposing players are considered more difficult, but this study will need to rely on available metrics, such as the pass's distance and direction.

Despite this limitation, promising models such as the ECOM rating (Bransen et al., 2019) provide a framework that fits this study's available data. Using possessions, this metric examines pass sequences leading to a goal and provides a score to each individual involved. The score each individual receives is based on the classified difficulty of the pass as well as the shot itself that scored the goal. While the ECOM rating proposed by Bransen et al. (2019) is designed to weight passes that lead to goals higher than passes that do not, this study aims to build upon the model to incorporate defensive contributions to some extent as well.

In addition to individual metrics, this study provides team-level analysis to help the SMU coaching staff better allocate their efforts. Whitaker et al. (2018) provide a road map to help quantify the attacking sequences of both SMU and the opposing teams. The attack plots created with the InStat data provide the success of each team's attacking progression by breaking that progression up into four different points: the attacking third, the 18-yard box, the goal zone, and whether or not a shot on goal took place. Each checkpoint of the attack progressions visualizes the attack end location as well as the percentage of attacks that proceeded beyond that point relative to the previous one.

The individual contribution model and other analytical models and insights are intended to be used as a blueprint to facilitate the analytics conversations with the SMU coaching staff and change the way the coaches think about their coaching style, tactics, and strategy.

2 Literature Review

While college soccer presents some unique rules and challenges, largely the analytics concepts that exist at the forefront of the soccer analytics world can be applied to the college game to improve on-field outcomes.

While high-revenue sports for the United States, such as baseball and football, have received a large amount of statistical analysis since the 1960s, professional soccer remains a step behind (Morgulev et al., 2018). Moreover, college soccer struggles behind the professional game and currently attempts to play catch-up through largely rudimentary statistics like goals, assists, and shots. Sukumar et al. (2019) categorize soccer's analytics capabilities to be "rather primitive," but sees promise as new technologies arise to capture more data points than what has been captured historically (Sukumar et al. 2019, p. 22).

Largely, Sukumar et al. (2019) attribute the primitive analytics capabilities in soccer to three primary phenomena. The first is the analytic obsession with goal scoring and a few other factors that may or may not contribute to winning a game. Secondly, there are fewer quantifiable events in soccer as compared to other sports. Lastly, is the idea that soccer is more qualitative than quantitative, or the idea that what makes a player "good" is often unmeasurable.

Despite this, the advancement of soccer analytics has pushed forward over the past decade with exciting advancements taking place. Rein & Memmert (2016) find promise in the future of soccer analytics, stating:

> "Exciting times are emerging for team sports performance analysis as more and more data is going to become available allowing more refined investigations" (p. 1266).

This is due to the emergence and public availability of tracking data, utilizing both Global Positioning Systems (GPS) and radio waves to effectively track players' actions and movements on the field of play. Tracking data has both improved analysis and increased data awareness within the soccer space, similar to the sabermetrics revolution that occurred in baseball in the late 1990s (Rein & Memmert, 2016).

Pappalardo et al. (2019) capitalized on this publicly available data. They took a spatial-temporal approach to investigate emergent measures such as the invasion index

or acceleration index (i.e., the opponent's velocity reaches the goal) (Pappalardo et al., 2019).

Not only has the massive volume of this data provided analysts with insight, but it also has caught the eye of the teams' medical staff as a way to track the optimal training load (Passfield & Hopker, 2017). As this data becomes more available and more accepted in the traditional thought centers of the soccer community, it is reasonable to assume that analysis will improve.

The conclusion reached by Rein & Memmert (2016) additionally requires that soccer research will have to become more cross-functional to benefit from the plethora of new data available. The literature indicates that this cross-pollination has indeed taken effect, producing analysis based on approaches that utilize machine learning, network analysis, and probabilistic models.

Brooks et al. (2016) meaningfully applied machine learning to analyze passing events in soccer games to classify teams and rank players. Although limited to the amount of data the team had access to, Brooks et al. (2016) were able to obtain an accuracy score of 87% when classifying 20 different teams based on the teams' passing distributions (Brooks et al., 2016). While useful in its own right, the team calls for further research to determine if this method could be useful in constructing team-specific models (Brooks et al., 2016), something that could be immediately impactful to this study's stakeholders.

Network analysis is also a field showing promise in the arena of soccer analytics. In less of a study and more of a gift to soccer analytics, Huang (2018) provides a practical way to quickly and effectively implement network analysis via coding to efficiently create a network graph based on teams' activity captured throughout games. While not a dramatically useful tool on its own, this method could prove instrumental in exploratory data analysis or determine different subcomponents of teams to inspect more thoroughly. Additionally, it provides a blueprint for the centroid analysis described by Rein and Memmert (2016) and the neural network-based self-organizing map approach of Memmert et al. (2017) as potential methods for evaluating tactics of an individual team.

Whitaker et al. (2018) propose a Bayesian Inference approach using only attacking events as a means to quantify a team's goal-scoring opportunities. Whitaker et al. (2018) quantify a team's ability to create scoring opportunities using multiple methods. Firstly, the players' assist locations are clustered by game via K-means. Once key locations are clustered, those labels are used to create continuous measurements via Gaussian mixture models (Whitaker et al. 2018). Finally, a hierarchical Bayesian model is used to fit the Gaussian mixture models of each player to estimate each individual's probability to create a goal-scoring chance (Whitaker et al. 2018).

At the forefront of soccer analytics research is American Soccer Analysis. Kullowatz (2018), from American Soccer Analysis, provides a model used to determine the odds ratio of the success of a given pass based on variables such as pass type (e.g. cross, header, throw-in, etc.), player position (e.g. kick-off, goalkeeper, etc.), distance, and game state. While the model itself is not where Kullowatz wants it to be, this avenue of investigation shows promise as an applicable model.

Muller (2018) provides a much more ambitious model to produce the number of expected goals based on the shot's timing and location. The output of the expected goals model, referred to as *xG*, is a straight percentage representing that particular shot's

likeliness resulting in a goal (Muller, 2018). The logistic regression model takes three main variables into account: shot location, goal mouth placement, and passing. The *xG* model has grown in complexity since its unveiling. It started with six distinct zones for the shot location variable, but not takes into account continuous measurements of the distance from the goal and the angle the player has on the goal (Muller, 2018). The model also breaks down into three separate applications. There is a model for teams, a model for the shooter themselves, and a model for goalkeepers.

Muller (2020) adds onto the expected passing model of Kullowatz (2018) with the G+Boost model, which takes more into account than the pass itself. As Muller (2020) explains, "A pass's other qualities, such as whether it moves the ball away from your goal and closer to the opponent's, can also change that balance. Sometimes that matters more than which team winds up with the ball" (p. 1). With the models becoming more comprehensive, such as G+Boost, there has been sufficient groundwork laid to extrapolate and explore alternative approaches that could prove to be beneficial.

The G+Boost model is especially useful because it provides the passing analysis of the expected passing model with the goal creation of the *xG* model. Of course, scoring goals is the desired result of offensive possessions in a given soccer game. Still, those results-based statistics are not particularly helpful for coaches, or players outside of forwards or offensively oriented midfielders. As Muller (2020) postulates, "Of course, we know not all passes are created equal," (p.1) thus describing the inherent usefulness of the model. As Sukumar (2019) previously describes the challenge of current soccer analytics, the singular focus on goal-scoring, the G+Boost model attempts to eliminate the statistical noise and focus on each action, and whether or not this directly contributed to a goal.

The primary problem that the G+Boost model seeks to solve is evaluating the possible duress of the attacking player, primarily the presence and positions of both their teammates and the defending team, through evaluation of what happens after the pass and what the expected goal differential is from each subsequent play (Muller, 2020). Thus, analysts can determine whether passes actively contribute to their team improving their chances to score a goal, rather than simply measuring pass completion or measuring the expected goals from the pass's direct action. This research will investigate how to apply this holistic modeling approach to the event-based outcome model of SMU's data.

While logistic regression is an obvious application given the data and expected utility of the analysis, purely probabilistic methods show great potential in the soccer analytics domain. Brefeld et al. (2018) provide a unique method for generating probabilistic movements based on spatial data using x, y coordinates, and time intervals and player velocity. Given the complexity of the calculations as well as the extremely small time window in which these insights would be applicable, the authors suggest that distributed computing and discretization of the calculations could speed the computations up to nearly real-time (Brefeld et al., 2018).

Soccer analytics has been gaining momentum in the analytics space in recent years. *The Machine Learning Journal* (Springer) hosted the 2017 Soccer Prediction Challenge called on data scientists from around the country to predict future matches based on a dataset from 200,000 matches from leagues around the world. Before this, the most notable discovery was in evaluating soccer actions, the expected goals model, which predicted probabilities of scoring a goal based on Euclidean distance to the goal and

angle (Damour & Lang, 2015). In the challenge, the methods for predicting wins utilized Poisson models, Bayesian models, and rating systems (Berrar et al., 2018). The two goals of the contest were to test the limits of predictive methods and further soccer analytics conversation. The winning model was created with manually engineered features and an extreme gradient boosted tree-based algorithm (Berrar et al., 2018). Other placed models included hybrid Bayesian networks and dynamic performance rating and hierarchical log-linear Poisson model (Berrar et al., 2018). Though this study's main goal is not to predict wins for SMU Mustangs Soccer Team, the article details two main areas of sports analytics that are of interest. Those are the aforementioned evaluating actions as well as Identifying tactics and strategy, which most interestingly seek to discover how effectively a team is at creating scoring opportunities, and to find normally occurring events that lead to shots on goal, or passing behaviors (Berrar et al., 2018). These are most of interest to the coaching staff of SMU, and provide leads on further research.

The ECOM rating or the expected contribution to the match's outcome detailed by Bransen et al. (2019) is a successful metric that has provenly outperformed past metrics such as historical distributions and pass accuracy. The metric aims to quantify players' contribution to a match in passes of possession sequences that either leads to a goal-scoring or threat opportunity or do not. This process has been invented due to soccer, which is a low scoring game and is not as quantifiable as other sports such as baseball and basketball (Bransen et al., 2019). In their method passes are grouped into possession sequences and labeled according to the probability of scoring a goal in the possession's finality. The passes are grouped either by a cell of field or cluster label. These grouped passes are then compared with a function of Euclidean distances of pass origins and destinations, incorporating some data points such as timing which are unable to mimic, but generally available to us. For each pass per grouping enters a calculation with other passes belonging to the same grouping, weighted by the scoring probability (Bransen et al., 2019). The result is normalized for 90 minutes of play per player and gives us the ECOM rating for a player. The paper's experiment demonstrates that the ECOM metric surpasses the four previously established baselines for predicting outcomes of matches and can potentially convey useful information for assigning a market value to players (Bransen et al., 2019). With a lower logarithmic loss in the prediction than other metrics, this will be the inspiration for the SMU Mustangs Soccer Players.

The progress made over the past two decades in soccer analytics has supplied momentum that this study seeks to carry forward. Its young, yet rich history shows the effectiveness of data science applications to soccer. Pappalardo et al. (2019) and Memmert et al. (2017) provide useful models to assist in accomplishing the goal of providing translatable and easily digestible analytics to the SMU coaching staff. Additionally, the ECOM put forward by Bransen et al. (2019) combines multiple facets of the research previously covered to produce one of the most effective evaluation metrics. Building upon the ECOM rating to evaluate individual player contribution and their impact on winning the match more effectively is the primary area where this study seeks to advance the 'SMU Men's Soccer program's analytics capabilities.

3 Methods

As previously discussed, there is a multitude of ways to explore and analyze soccer; a variety of which are in competition to be employed in professional sports. How this study best achieves its goal is best determined by the SMU coaching staff's needs and the possibilities inherent in the data. As a nationally ranked team, the Mustangs need to be able to make decisions regarding the line-up, coaching methods, tactics and strategy, and potentially even opposition scouting. Furthermore, professional football clubs and their scouts will want to know which players outperformed the group on an objective basis. Ideally, the SMU coaching staff would supply this information and earn trust as being an organization comparable with analytics teams behind many football clubs in professional soccer.

The methods which are beneficial to these points are divided up into four sections. Firstly, the data used for these methods will be discussed. Next, how individual player performances will be expanded. Thirdly, team-level metrics that show potential to the SMU coaching staff will be discussed. Finally, the individual player results will ideally frame a determination to the outcome of games. The outcome of a match, realistically, is an examination of the accuracy of the method to analyze player performance and not a future tool or predictor of wins or losses because of the limitations of available data. If SMU were to invest in other teams' data, prediction could be possible.

3.1 Data

The data in this study utilized for analysis was provided by InStat, a company that provides data and tools for individual and team sports analytics. The data was gathered by iterating through a list of match I.D.s from the 2019 SMU Mustangs soccer season and making requests to the InStat football API. The final dataset was a composition of all 21 matches throughout the 2019 season. The dataset is mainly composed of the actions of each player throughout the match with a timestamp and x/y coordinates of the player for each action. The origin of the coordinates is the goalkeeper's corner of the team of the player who performed the action. The dimensions are calculated in meters and the field is 105x68 meters (using a horizontal layout). There is a range of 119 unique actions in the dataset (e.g. goal, tackle, shot blocked, etc.). Many of the 78 explanatory features in the dataset are metadata which encapsulate a noteworthy action of a player (e.g. the main action is a goal, and variables pos_x and pos_y tell the location of the player from where the shot was taken). Figure one summarizes the most important features in the dataset.

Table 1. Data points provided by InStat and their descriptions.

Data Point Name	Data Point Description
team_name	Name of the team which the player who performed the action belongs to.
half	Indicates which half of the match the action took place.
second	Indicates the time (in seconds) which the action took place.
player_name	Name of the player who performed the action.
position_name	Position of the player who performed the action (e.g. goalkeeper, midfielder, etc.)
opponent_name	Name of the immediate opponent of the player who performed the action (if no opponent is applicable to the action, the opponent_name entry will be "N.A.").
opponent_team_name	Name of the opponent's team.
opponent_position_name	Name of the opponent's position.
zone_name	Indicates which zone on the field the which action took place.
zone_name_dest	Indicates the destination zone on the field of the action (if no destination zone, the entry for this variable will be "N.A.").
Pos_x	x-coordinate on the field of where the action took place.
Pos_y	y-coordinate on the field of where the action took place.
Pos_dest_x	x-coordinate on the field of the destination of the action.
Post_dest_y	y-coordinate on the field of the destination of the action.

Overall, the dataset is very comprehensive. Out of the 65,505 actions recorded throughout the 2019 season, most of the observations contain all imperative metadata surrounding an action. The plots below demonstrate the proportion of missing values for each feature in each observation.

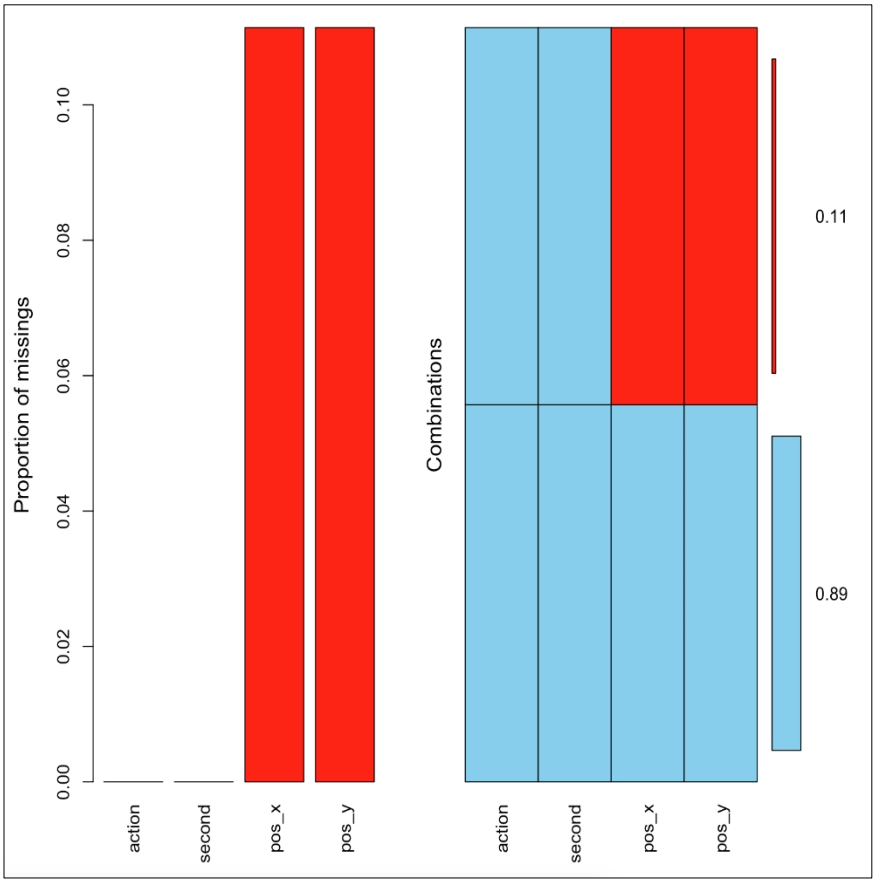

Fig. 1. Proportions of missing Data

Figure one shows that not one single action is missing a timestamp. Only 11% percent of the actions logged are missing a coordinate value for the action's location on the field; this is due to some actions not about a specific player or actions which are done by players on the field. Some of the actions not pertaining to a coordinate location on the field include player formations (e.g. 4-4-2 classic, 4-3-3 down) and events such as player substitutions and a declaration of the beginning of the second half.

Aside from missing values, the number of important actions directly impacts the usefulness of the dataset. Given the composition of the ECOM rating, as well as the actionable insights needed to provide useful analytics to the coaching staff, actions that involve passing and shots/goals are instrumental to this study. Table two and three provide the count of each related activity, as well as their percentage of the overall observations. Although this study only includes a single season of gameplay, the total number of passing, shooting, and scoring observations provides sufficient data to conduct a meaningful analysis.

Table 2. Data points relating to passing activity.

Action Name	Count of Observations	Percentage of Observations
Non attacking pass accurate	10216	15.65%
Attacking pass accurate	975	1.49%
Accurate key pass	609	0.93%
Attacking pass inaccurate	349	0.53%
Inaccurate key pass	122	0.19%
Pass into offside	106	0.16%
Non attacking pass inaccurate	85	0.13%
Pass interceptions	60	0.09%
Extra attacking pass Assist	53	0.08%
Extra attacking pass accurate	48	0.07%
Inaccurate extra attacking pass	26	0.04%

Table 3. Data points relating to shooting/scoring activity.

Action Name	Count of Observations	Percentage of Observations
Shot into the bar/post	2685	4.11%
Shots blocked	551	0.84%
Goal	477	0.73%
Shot on target	237	0.36%
Wide shot	168	0.26%
Misplaced crossing from set piece with a shot	136	0.21%
Shot blocked	112	0.17%
Accurate crossing from set piece with a shot	8	0.01%
Shot blocked by field player	6	0.01%

3.2 Individual Player Performances

When a soccer match is in play it can be difficult for a coach, standing from the sidelines, to support a player or group of players without giving vague directions. Also, remembering the course of actions performed by individual players over each 45-minute half, with no timeouts and eleven players to keep track of, can be confusing.

Every action within each possession and whether that possession resulted in a goal attempt can be reviewed with video after the match. The players of that possession and the difficulty of their actions need to be evaluated objectively and recorded. A metric tailored for individual player performance during a match and over the season can achieve this, giving a coach a measure of the contributions much deeper and quicker than a visual review of each game video. Furthermore, the contribution values will be relative to the pass's success and distance in a sequence, thus being granted more points for difficulty.

To evaluate the contributions of players, first, the result of the possession needs to hold a value that communicates beyond the binary result of a goal. Whether the goal was made or not, a value of 0 or 1, does little except evaluating the possessions that were a success and would not take into account the extensive work done by a team to wear down an opponent and, in a way, experiment with offensive drives to find a weakness in the opponent's defensive strategy. Those who watch soccer understand the importance of these runs and how they develop throughout a match. One experimental run may test the open spaces in the opponent's side and make shots that are on target, not strenuous for the opponent goalkeeper or miss the goal entirely.

The greater value for these results is already widely used in Soccer Analytics and is known as xG or expected goals (Muller, 2018). The expected goals value can take a variety of dependent variables including shot type, shot origin, goal angle, and shot distance. Because shot type is not recorded by the available InStat data it is missing from the tabular columns or features for the Mustangs. The independent variable or the event that is recorded in a pass/fail scenario is the outcome of the shot; does it result in a goal or not? Together these independent and dependent variables can be used in a functions, above, to model the probability or expected goals.

$$s_i = B_0 + B_1 origin + B_2 distance + B_3 angle \quad (1)$$

$$if\ Action\_Name_i = Goal\ then\ LF = \frac{\exp(s_i)}{1+\exp(s_i)} \quad (2)$$

$$if\ Action\_Name_i = Shot\ (No\ Goal)\ then\ LF = \frac{1}{1+\exp(s_i)} \quad (3)$$

L.F. is representation of the likelihood factor or the probability, achieving the expected goals value for possessions that are labelled as attack.

The definition of a possession is a sequence of actions by which the SMU Mustangs retain control over the ball, recorded as consecutive records with column "team_name" as "SMU Mustangs" until "team_name" shows as the opponent. Over the entire season these possessions are numbered and marked as either "attack" or "play" if the possession ends in a shot, regardless of the outcome. The actions within each possession

are numbered in increments. This allows for a kind of primary key, a unique identifier, for each action which is useful to pair actions with transformations with the metric's functions. The labels of the xG are distributed across all actions within the same possession. For example, the result of possession 123 had a shot with an expectation of a goal 25% of the time so all previous passes within the same possession receive the same expectation as a preliminary contribution, regardless of field position.

The next step in quantifying this metric requires returning to a view of the passes during the game about their positional arguments. To compare similar passes, labeled with respective possession's scoring probability, a cluster analysis must be performed. Because the data is limited to a single season for a single team this is preferable to a more exact cell-based approach. Dividing a field into cells, compartmentalizing passes into origin cells and destination cells has greater relevance, but this would greatly limit the number of comparable passes for the metric. A more concise approach is clustering, using the K-means algorithm. The K-means method is best for the limited field approach. In action, it partitions the pass vectors into k clusters, with each pass consequently belonging to its kth cluster because of its mean or centroid. The formula adjusts its clusters to minimize the within-cluster sum of squares to define groups of the most passes. The passes within the groups will be compared with one another.

$$K\ means\ Objective\ Function = \sum_{j=1}^{k}\ \sum_{i=1}^{n}\ ||x_i^{(j)} - centroid_j||^2 \quad \textbf{(4)}$$

Once the proper k has been deduced from cluster validation the formative formulas of the metric and pass-xG comparison can be computed. First, each pass within each cluster needs to be assigned a count, again as a kind of primary key to keep track of each pass. The reason for this is because within a cluster every pass will be compared to every other pass and will consider the distance between the origin of the two passes, the distance between the destination of the two passes, and the xG label of the second pass as weight to the first pass. The first pass is compared to all other passes within the cluster in this fashion and summed based on a parameter of success. This equation is denoted below as value Ve.

$$if\ pass_i\ has\ xG > 0;\ V_e(pass_i) = \sum_{j=1}^{n}\ s(pass_i, pass_j) \cdot xG(pass_j) \quad \textbf{(5.1)}$$

$$if\ pass_i\ has\ xG = 0;\ V_e(pass_i) = 0 \quad \textbf{(5.2)}$$

The secondary value Vs is the average of all xG labels of other passes within its respective cluster, as denoted below.

$$V_s(pass_i) = \frac{\sum_{j=1}^{l} xG(pass_j)}{l} \quad \textbf{(6)}$$

Finally, the difference between these two values, Ve and Vs, gives the value V. This value V is the score for the record containing a single pass. The idea is that a pass holds a value that is contingent upon its distance and a weighted average of similar passes, less the average outcome of other passes.

$$V = V_e - V_s \quad \textbf{(7)}$$

This value (V) can then be summed or averaged in a grouping of games or players to create real, digestible values for coaching staff in minutes, compared to hours of video inspection and notation. This contribution metric also creates other opportunities for understanding the Mustangs 2019 season. More will be discussed in Results.

3.3 Offensive and Defensive Attack Plots

One of the key components in gaining momentum over the opposing team in soccer is to achieve as many shots on goal as possible. In order to place effective shots on goal, it is necessary to get into proper position. The progression of the attacking team's positional advantage can be separated into four different checkpoints. The first checkpoint is the attacking third. The attacking third is defined as an attack succeeding past the opponents third of the field, or past the 70-yard mark. The second checkpoint is called the 18-yard box (or penalty box). For an attack to successfully breach the 18-yard box, the attack must end past the 91.7-yard mark on the x-axis, and between the 18.9-yard and 49.1-yard marks on the y-axis. The goal zone is the third checkpoint. The goal zone area is the same 91.7-yard mark on the x-axis, but the y-axis boundaries are narrowed to 27.2 and 40.8 yards. The final checkpoint is whether or not the attacking team was able to execute a shot on goal. A shot on goal is classified as a shot on goal, or goal, from the goal zone.

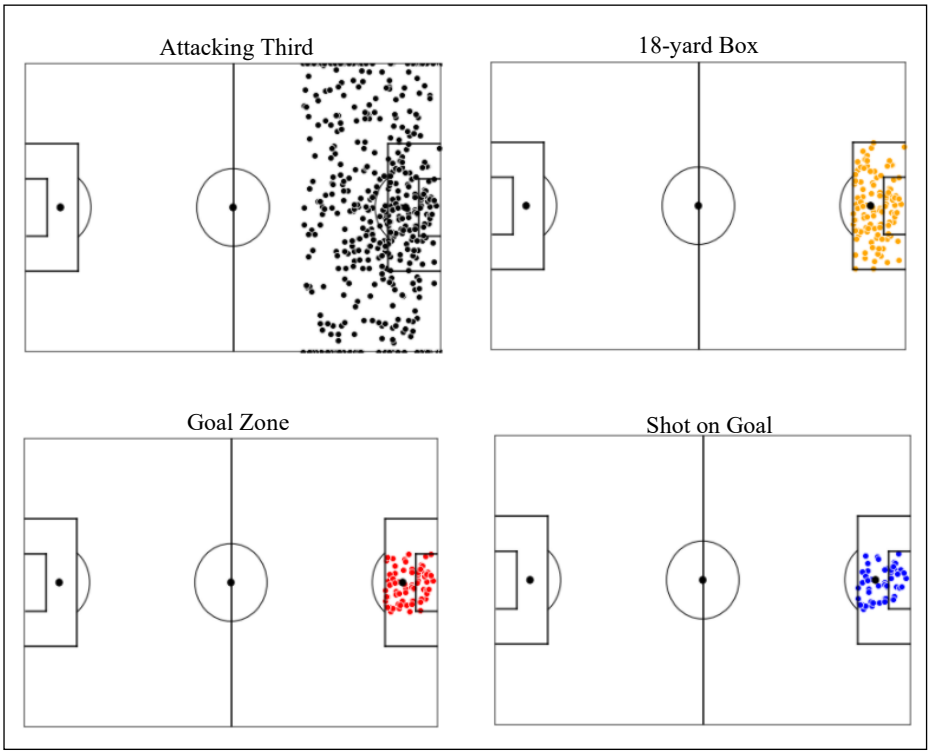

Fig. 2. The four measurement points of the attack plots. Top-left shows the attacking third, top-right the 18-yard box, bottom left the goal zone, and bottom right the final measurement space of shot on goal.

Attack plots are produced for the SMU offense and defense by game as well as by season total. While not exceptionally technical from a machine learning standpoint, the attack plots provide effective and easily interpretable insight into strengths and weaknesses in terms of offensive positioning as well as defensive success. The ability to dissect whether or not a shot was taken, or a goal was scored, into more detail allows the coaching staff to pinpoint areas in the attack sequence that need the most attention.

4 Results

The analysis contained a contribution metric that, because it relied heavily on popular areas on the field and scoring probabilities more opportunities to score higher in the metric, were on the offensive side because it was more likely that an attack on goal would happen. The metric was biased to the player's actions in the offensive final third of the field. This was evident in the plots and brought to attention by the coaching staff. The chances were higher to contribute to an attack the greater the position x toward the opponent's goal.

The coaching staff asked for a more balanced view of each game. This leads to a motion to rebalance the metric. The best idea was to first standardize the metric values twice, once in each half of the field actions. Secondly, to scale the values from 0 to 100 and reformat them to be integer values. Luckily, this did not significantly alter the main application of the metric, a sum of values per player and discover which players contributed the most to the season. The final summed contribution metric for each player can be viewed in the table below.

Table 4. Expected Contribution Metric for Individual Player Results

Player Name	Value
Brandon Terwege	9415
Cole Rainwater	0
DJ Williams	987
Eddie Munjoma	13743
Gabriel Costa	9289
Garrett McLaughlin	5954
Grant Makela	1149
Henrik Bredeli	7436
Henry Smith-Hastie	1993
Jacob Cohen	136
Jon-Talen Maples	4672
Joshua Berney	518
Knut Ahlander	14287
Lane Warrington	5050
Luke Thompson	717
Nick Taylor	3401
Nicky Hernandez	10410
Noah Hilt	8068
Philip Ponder	7283
Shane Lanson	349
Thomas Haney	3716
Tobin Shanks	528
Wyatt Priest	887

The only player which, according to coaching staff, had a value above what was expected was Brandon Terwege. Though, the authors are confident there are no mistakes and the data is simply stating an outcome that had not been realized by coaching staff. All other players compare favorably to a perceived contribution.

4.1 Application of Contribution Metric Values

An analysis of individual player contributions is useful for crafting a game roster, but soccer is a team sport. The contribution metric can be as a lens through which to view the game in greater scopes-common actions and possessions.

There are over 100 different action names to describe play. Which actions are the success-drivers for the Mustangs motivated the graphic below.

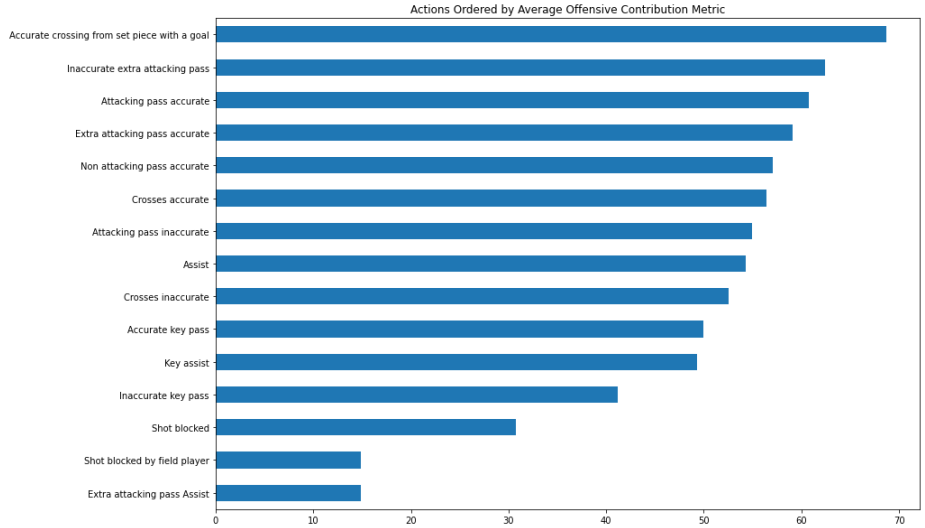

Fig. 3. Top actions by Mustang players according to contribution metric (with shot action names removed)

According to Figure 3 the contributing actions in descending order show the strength of the metric. The order of the actions is generally the order one would assume a higher or lower contribution to a match to be. This snapshot of the team actions shows where the greatest strength lies and hints at actions where the team did not comprehensively succeed.

The second prominent application of the contribution metric is with possessions. In total there are over 4700 possessions by the Mustangs in the season. The question arises, which possessions racked up the most metric points and what do they look like? Since the metric is made up of expected scoring as well as spatial harmonic areas by the centroids the top 25 plays show areas of tried-and-true success. Success defined as a shot taken, whether it was a goal or missed the target entirely.

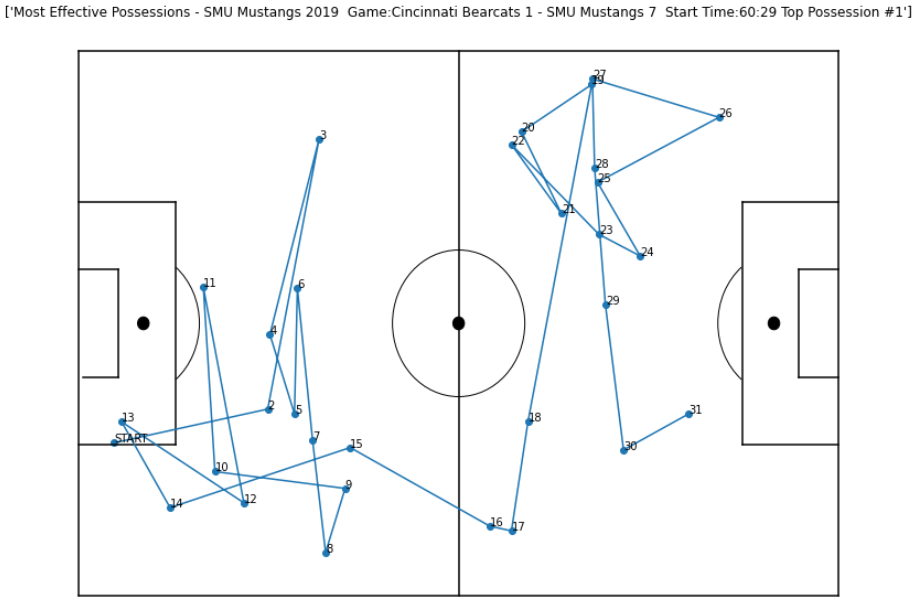

Fig. 4. #1 Top possession according to combined contribution metrics for possession actions

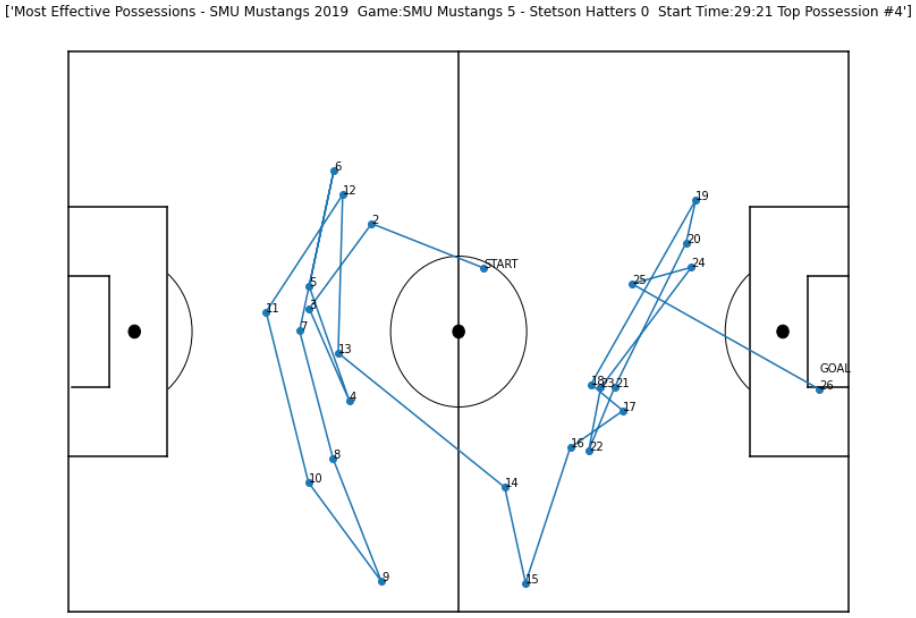

Fig. 5. #4 Top possession according to combined contribution metrics for possession actions

25 visualizations were prepared including the game-clock as reference to view the play unfold on video for more detail to assist the coaching staff in digesting this information.

4.2 Statistical Viability of Contribution Metric

The null hypothesis is that the mean metric values are equal in games that are won and lost. The alternative hypothesis is the mean metric values are different in won and lost games. Using a pooled standard deviation due to unequal variances, the test statistic is 8.38. The probability of observing by random chance, metric values as extreme or more extreme than what was observed was recorded. The null hypothesis was rejected.

In summary, this experiment provides strong evidence that higher metric values are indicative of positive match outcomes (p-value = 2e-17 from a two-sample Welch's t-test). The estimated gain in contribution metric is 4.1. (95% confidence interval for the winning effect on the metric is [3.15, 5.03] for the average metric across all possessions in the 2019 season.)

While a metric mean may show a difference in quality of attack possessions, another perspective of the quantity of non-zero values between games won, lost, and tied are of interest. The mean count of number of actions in each game outcome type also supports the metric.

Table 5. Average Actions per Game with Non-Zero Metric Values in 2019 Season

Outcome	Actions
Won	97
Tied	65
Lost	62

5 Discussion

The ability to quantify and realize potential outcomes from detailed geospatial and event-based data in college soccer games is a possible value add. Coaching staff can utilize data on non-scoring events throughout the game to adjust game plans and practices that will lead to more shots on goal. There are, however, points to be made about the metric that must be considered for its application and utility.

The traditional approach for analyses with model building and numeric outcomes defer to real metrics that are essential in grading said models using error measurements. This was not the purpose of the metric. It uses models to concretize a value for the valueless. In other words, there are no set values to grade an action by which to provide a metric for evaluation. The metric is derivative to other measures through statistical algorithms. For being derivative, the application is not limitless.

College soccer, being a revolving door of players, comes at odds with the metric that relies on repeated data to quantify a set of results for comparisons. From season to season, the metric becomes less about the scoring of individual players and more about the training and strategy of the coaching staff. Also, if a player is injured the values could be compromised by the sparsity of data. The metric needs as much data as possible, and a season with a set roster may be the limit for the metric. A season-to-season comparison is equitable, but differences in player skills from season to season (specifically the strikers) will create an imbalance that would undermine the respective metric values.

Another consideration that limits the use of the metric is the failure of a team. The metric depends on goal-scoring opportunities, specifically an array of different shots, by which to label the supporting possession sequences. If a team were to play poorly game after game, there is little the metric can do to evaluate its players' contributions. Winning seasons create winning metrics. Losing seasons create sparse data which will add fewer values to assess.

Though the SMU Mustangs failed to clinch a title, they were still in contention. By the metric, the 2019 season is capable of being evaluated for future seasons. More importantly, the metric creates a basis for comparing plays that assist the coach's research of past games. In this sense, it was a success.

The greatest confirmation the metric received was in the presentation of Table 4 to the head coach. He affirmed the ranking of players aligned with his expectations. Also, the top-scoring players Eddie Munjoma and Knut Ahlender are now under contract with Major League Soccer (MLS). Another tangible result were the heat maps of the metric for winning and losing games, further evidence beyond numbers that it certainly works.

The delivery to SMU Mens Soccer's coaching staff was a file (csv) of all the possession sequences with identifiable timestamps, ECOM metric averages, ECOM metric totals, and play styles as defined by the director of U.S. Soccer. This result is the document that will be used in conjunction with more visual analysis of gameplay of the 2019 Season.

The metric's greatest challenge was the meticulous manipulation of data points and results to achieve an understandable outcome value. The computational process of achieving this metric was arduous and possible through leveraging of many lessons in the Data Science program.

Areas of deeper analyses using a similar metric would be contingent on the supply of incoming data. For the deliverable to the coaching staff, this study leveraged nearly all of the data through the inStat API. Other soccer programs may have more data points recorded, such as shot type and timings of possessions. The process for the metric would only be improved for these additions as reference. Shot type can be utilized to create likelihood factors (L.F.) or labels for possession sequences. Timings can be used in conjunction with the metric value to normalize and rebalance player actions based on statistical measures of quickness or danger for the length of player possession.

The authors expect that this work will inspire more scientific analyses ending in relative measures of success. With caution and awareness of implications, statistical metrics can inspire derivative outcomes in other applications. A scarcity of definable values should not hinder scientific research. The results may not always align with expectation, but the general results should provide a pleasant surprise, as it has here, if statistical models are applied reasonably.

5.1 Future Research

There are two primary areas where this research could be advanced: evaluating substitutions and physical fitness barometers.

College soccer uses an unlimited substitutions rule that is distinctly different from most other levels of the game, which makes substitutions a more important part of the game. Hill (2019) investigates these substitute players' warm-up patterns, which this research could combine with the ECOM metric to determine the best way SMU and others can utilize their substitutes.

Additionally, this paper could look at physical baseline data from a variety of physical tests outlined in Saward et al.'s research (2020). This research can help determine players' physical development and better help coaches understand expectations, especially for younger players. Additionally, future research can look at research similar to De Beeck (2020) and better assess players' athletic load and overall wellness to ensure coaches are getting the most out of players physically.

5.2 Ethics

There are many different ethical concerns when it comes to implementing any numerical scoring metric, particularly one that ranks the players' estimated contribution. Sufficient consideration should be given to the implementation of this metric and the weight coaching staff put on the metric as part of a holistic approach to guiding their players to success.

Firstly, an area of concern is how an implementation of this metric would impact the dynamic between the players. There is a potential risk that too much emphasis will be given to this metric. As discussed, this metric is not meant to provide a definitive conclusion or final result, but is based on probabilities and provides a similar likelihood to scoring opportunities. If this distinction is not made, and too much stock is put into this metric without any other factors, potential issues could arise.

With each player having a prescribed numerical value based on their expected contribution toward scoring opportunities, there is potential for some players to feel inadequate or left out if they are not among the high-scoring group of players. Growing animosity between teammates could weaken the bond of the team, leading not only to decreased performance but also mental and psychological issues for the players.

Additionally, it is unknown how various coaching staffs would use this metric. With limited coaching resources, would the coaching staff focus solely on the highly-rated players to increase their effectiveness, or would they devote attention to the players with lower scores to create a more robust team as a whole? Similar to the intrateam animosity scenario, a coaching staff who neglects certain players could potentially lead to a decline in mental health for those overlooked players.

Lastly, there should be consideration around the best use of this geospatial and event-based data. There is potential to use this player-specific data to monitor and evaluate

the players' physical health, which presents opportunity cost for resource-constrained soccer programs (Saward et al., 2020). Suppose coaching staffs and soccer programs are limited in their data analysis capabilities. In that case, the use of this metric prior to any implementation concerning the health of the players could pose ethical concerns.

6 Conclusion

The ability to use InStat game data to produce the expected contribution metric and the offensive and defensive attack plots, provides a potentially high return on investment for the SMU coaching staff. The ability to capture this data for a relatively low cost (low thousands of U.S. dollars) and analyze it with open-source software would provide actionable insights without the potentially more expensive addition to the coaching staff. This study's application of ECOM, originally put forth by Bransen et al. (2018), shows that the approach is statistically significant in indicating the outcome of the match.

While this study was limited to data from the men's soccer program at SMU, it is feasible that this approach could be taken to the women's program and other amateur level programs such as high schools and travel clubs, both domestic and abroad. This would help the SMU coaching staff with recruiting players that mesh with what the coaching staff teaches on a daily basis. Furthermore, suppose this study was provided data for the rest of the conference, or the whole of collegiate soccer. In that case, this study could start to create baselines and leaderboards that would not only assist SMU with scouting, but also put players in better context within the greater collegiate soccer landscape.

The practical application of this model provides the SMU coaching staff and other collegiate programs the ability to participate in the ongoing data revolution with a relatively low barrier to entry. The SMU Soccer staff is now armed with the tools, terminology, and knowledge to make their first foray in the world of soccer analytics, offering a distinct advantage over other college teams which frequently do not have the same analytics expertise. Furthermore, because the cost of data is relatively low and the software is open source, this process is beneficial and sustainable for the SMU soccer program.

Acknowledgments. Jacquelyn Cheun, PhD. – Capstone Professor

References

1. Beéck, T. O., Jaspers, A., Brink, M. S., Frencken, W. G., Staes, F., Davis, J. J., & Helsen, W. F. (2019). Predicting Future Perceived Wellness in Professional Soccer: The Role of Preceding Load and Wellness. *International Journal of Sports Physiology and Performance, 14*(8), 1074-1080. doi:10.1123/ijspp.2017-0864

2. Berrar, D., Lopes, P., Davis, J., & Dubitzky, W. (2018). Guest editorial: Special issue on machine learning for soccer. *Machine Learning, 108*(1), 1-7. https://doi.org/10.1007/s10994-018-5763-8

3. Bransen, L., Haaren, J. V., & Velden, M. V. (2019). Measuring soccer players' contributions to chance creation by valuing their passes. *Journal of Quantitative Analysis in Sports, 15*(2), 97-116. https://doi.org/10.1515/jqas-2018-0020

4. Brefeld, U., Lasek, J., & Mair, S. (2018). Probabilistic movement models and zones of control. *Machine Learning, 108*(1), 127-147. https://doi.org/10.1007/s10994-018-5725-1

5. Brooks, J., Kerr, M., & Guttag, J. (n.d.). Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 9*(5). Retrieved September 10, 2020, from https://onlinelibrary-wiley-com.proxy.libraries.smu.edu/doi/full/10.1002/sam.11318

6. Damour, G., & Lang, P. (2015). Modelling Football as a Markov Process (Unpublished master's thesis). KTH ROYAL INSTITUTE OF TECHNOLOGY.

7. Hills, S. P., Barrett, S., Feltbower, R. G., Barwood, M. J., Radcliffe, J. N., Cooke, C. B., . . . Russell, M. (2019). A match-day analysis of the movement profiles of substitutes from a professional soccer club before and after pitch-entry. *Plos One, 14*(1). doi:10.1371/journal.pone.0211563

8. Huang, L. (2019, September 26). A Network Analysis of Soccer Passes: Parsing XML with Python to Visualize Sports in Gephi. Retrieved September 10, 2020, from https://sites.temple.edu/tudsc/2018/10/23/create-a-network-from-xml-and-visualize-it-a-soccer-player-passing-network/

9. Kullowatz, M. (2018, April 23). An Updated Expected Passing Model. Retrieved September 10, 2020, from https://www.americansocceranalysis.com/home/2018/4/19/an-updated-expected-passing-model

10. Luzum, N.Z, Michael M.M. (2019). The Soccer Analytics Revolution History and Background. Duke University. https://sites.duke.edu/socceranalyticsrevolution/history-and-background/

11. Memmert, D., Lemmink, K., & Sampaio, J. (2017). Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. Sports Medicine, 47(1), 1–10. https://doi.org/10.1007/s40279-016-0562-5

12. Morgulev, E., Azar, O., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5, 213–222. https://link-springer-com.proxy.libraries.smu.edu/article/10.1007/s41060-017-0093-7

13. Muller, J. (2018). XGoals Explanation. Retrieved September 10, 2020, from https://www.americansocceranalysis.com/explanation

14. Muller, J. (2020, July 28). G Boost: Measuring What Happens After the Pass. Retrieved September 10, 2020, from https://www.americansocceranalysis.com/home/2020/7/27/gboost-measuring-what-happens-after-the-pass

15. Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus, 5*(1). https://doi.org/10.1186/s40064-016-3108-2

16. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. Scientific Data, 6(1), 1–10. https://doi.org/10.1038/s41597-019-0247-7

17. Passfield, L., & Hopker, J. (2017). A mine of information: can sports analytics provide wisdom from your data? A Mine of Information: Can Sports Analytics Provide Wisdom from Your Data?, 12(7), 851–855. https://doi.org/10.1123/ijspp.2016-0644

18. Saward, C., Hulse, M., Morris, J. G., Goto, H., Sunderland, C., & Nevill, M. E. (2020). Longitudinal Physical Development of Future Professional Male Soccer Players: Implications for Talent Identification and Development? *Frontiers in Sports and Active Living, 2*. doi:10.3389/fspor.2020.578203

19. Sukumar, S. S., Davis, J. A., & Laczniak, R. (2019). Soccer Analytics & its future. *Iowa State University Digital Repository*. Retrieved September 10, 2020, from https://lib.dr.iastate.edu/creativecomponents/254/

20. Whitaker, G., Silva, R., & Edwards, D. (2018). Visualizing a Team's Goal Chances in Soccer from Attacking Events: A Bayesian Inference Approach. Big Data, 6(4), 271–290. https://doi.org/10.1089/big.2018.0071