PROGRAM

September 28, 2019

Harvard University
Science Center, Lecture Halls C and D
1 Oxford Street
Cambridge, Massachusetts 02138

Symposium Organizing Committee:

Mark E. Glickman, Department of Statistics, Harvard University – Co-chair

Scott R. Evans, Department of Biostatistics and Bioinformatics, George Washington University – Co-chair

Laurie Shaw, Harvard Data Science Initiative, Harvard University – Panel organizer

Sponsors:

- Harvard University Department of Statistics (`https://statistics.fas.harvard.edu/`)
- IOS Press (`http://www.iospress.nl/`)
- DeGruyter (`http://www.degruyter.com/`)
- ESPN Analytics (`http://www.espn.com/analytics/`)
- SIG (`http://www.sig.com/`)
- SportLogiq (`http://www.sportlogiq.com/en/`)
- SimpleBet (`https://www.simplebet.io/`)
- Section on Statistics in Sports of the American Statistical Association (`http://community.amstat.org/sis/home`)
- Boston Chapter of the American Statistical Association (`http://community.amstat.org/bostonchapter/home`)

Wifi access:
Free guest wifi access will be available for conference participants in the Science Center. Please follow these instructions:

- With your laptop or mobile device's wifi turned on, choose "Harvard University" Wireless Network (SSID).
- Open a browser and go to `http://getonline.harvard.edu`. Select "I am a guest" from that page.
- Register for guest access on the resulting page. You will be asked for your name and e-mail address, and a confirmation that you accept the terms of use.

Acknowledgments: We wish to thank everyone who helped to make NESSIS possible. Special thanks go to Kevin Rader, Tom Lane, Patrick Gravelle, Paul Marino, Luke Bornn, Devin Pleuler, Andrew Swift, Karen Barkow, Natalie McKennerney, Emilie Campanelli, Xiao-Li Meng, Sameer Deshpande, Michael Schuckers, Kristen Hunter, Lisa Ruan, and Regina Nuzzo for their parts in helping with the symposium.

2019 New England Symposium on Statistics in Sports

Breakfast and Registration: Foyer area

8:00am – 9:15am: Coffee, tea, pastries

Welcome Address

9:15am – 9:30am: Mark Glickman and Scott Evans

Morning Session: Lecture Hall C - Featured talks

9:30am – 10:00am: Barry Nalebuff, Yale University
"Measuring Competitive Balance Correctly (in Sports)"

10:00am – 10:30am: Richard Smith, University of North Carolina, Chapel Hill
"How Do Typical Runners' Performances Vary with Age and Gender?"

10:30am – 11:00am: Ron Yurko, Carnegie Mellon University
"Going Deep: Models for Continuous-Time Within-Play Valuation of Game Outcomes in American Football with Tracking Data"

Break: Foyer area

11:00am – 11:30am: Coffee and tea

Late-morning Parallel Sessions

11:30am – 1:00pm: Lecture Halls C and D

Lecture Hall C - Novel Use of Tracking Data

11:30am – 12:00pm: Neil Johnson, ESPN Analytics
"Extracting Player Tracking Data from Video Using Non-Stationary Cameras and a Combination of Computer Vision Techniques"

12:00pm – 12:30pm: Fan Bu, Duke University
"SMOGS: Social Network Metrics of Game Success

12:30pm – 1:00pm: Jacob Mortensen, Simon Fraser University
"Estimation of Player Load Metrics Using Broadcast-Derived Tracking Data"

Lecture Hall D - Innovations in Sports Applications

11:30am – 12:00pm: Katherine Evans, Toronto Raptors
"Treatment Effect Heterogeneity in MLB Bunting Strategies"

12:00pm – 12:30pm: Katy McKeough, Harvard University
"Plackett-Luce Modeling with Parametric Growth Curves for Predicting Career Trajectories"

12:30pm – 1:00pm: Brian Macdonald, ESPN Analytics
"Analyzing Player Performance in eSports"

Lunch break: Foyer area

1:00pm – 2:00pm: Sandwiches, beverages, snacks

Afternoon Parallel Sessions

2:00pm – 3:30pm: Lecture Halls C and D

Lecture Hall C - Unsupervised Analysis of Tracking Data

2:00pm – 2:30pm: Laurie Shaw, Harvard University
"Classifying and Analyzing Team Strategy in Professional Soccer Matches"

2:30pm – 3:00pm: Dani Chu, Simon Fraser University
"Route Identification in the National Football League"

3:00pm – 3:30pm: Sam Gregory, Sportlogiq
"Unsupervised Run Type Detection"

Lecture Hall D - Evaluation of Players and Team Lineups

2:00pm – 2:30pm: Devan Becker, University of Western Ontario
"Comparing NHL Players' Shots and Goals by Algorithmically Decomposing Shot Intensity Surfaces"

2:30pm – 3:00pm: Nathan Sandholtz, Simon Fraser University
"Measuring Spatial Allocative Efficiency in Professional Basketball"

3:00pm – 3:30pm: Sarah Mallepalle, Carnegie Mellon University
"A Naive Bayes Approach for NFL Passing Evaluation Using Tracking Data Extracted from Images"

Poster Session: Foyer area

3:30pm – 5:00pm: With snacks and beverages

Panel Discussion: Lecture Hall C

5:00pm – 6:30pm: *"The State of Soccer Analytics"*

Moderator: Seth Walder – ESPN

Panelist: David Eccles – ChyronHego
Panelist: Tyler Heaps – US Soccer
Panelist: William Spearman – Liverpool FC

Post-NESSIS Get-Together

7:00pm – 9:30pm: The Heights, Smith Campus Center, 10th Floor
1350 Massachuseets Ave., Cambridge, MA 02138
(between Dunster St. and Holyoke St.)
`https://commonspaces.harvard.edu/theheights`

Appetizers provided; Cash bar only.
Conference participants will need to show ID at security desk and wear their name badges to get access to the venue.

Oral Presentation Abstracts

COMPARING NHL PLAYERS' SHOTS AND GOALS BY ALGORITHMICALLY DECOMPOSING SHOT INTENSITY SURFACES

Becker, Devan G.† (1); Woolford, Douglas G (1); Dean, Charmaine B (2)

(1) The University of Western Ontario, London, ON, Canada; (2) University of Waterloo, Waterloo, ON, Canada

† E-mail: *dbecker7@uwo.ca*

Spatial point processes have been successfully used to model the relative efficiency of shot locations for each player in professional basketball games. These analyses are possible because each player makes enough baskets to reliably fit a point process model. Goals in hockey are rare enough that a point process cannot be fit to each player's goal locations, so we must employ novel techniques to obtain measures of shot efficiency for each player.

We use a Log-Gaussian Cox Process to model all shot locations, including goals, of each NHL player who has taken at least 500 shots in the last 8 years. Each player's LGCP surface is treated as an image and these images are then used in an unsupervised machine learning algorithm that decomposes the pictures into a linear combination of spatial basis functions. The coefficients of these basis functions are a very useful tool to compare players.

To incorporate goals, the locations of all shots that resulted in a goal are treated as a "perfect player" and used in the same algorithm (goals are further split into perfect forwards, perfect centers, and perfect defense). These perfect players are compared to other players as a measure of shot efficiency. This analysis provides a map of common shooting locations, identifies regions with the most goals relative to the number of shots, and demonstrates how each player's shot location differs from scoring locations.

SMOGS: SOCIAL NETWORK METRICS OF GAME SUCCESS

Bu, Fan†; Xu, Sonia; Heller, Katherine; Volfovsky, Alexander

Duke University, Durham, NC, USA

† E-mail: *fb75@duke.edu*

We propose a novel metric of basketball game success, derived from a team's dynamic social network of game play. We combine ideas from random effects models for network links with taking a multi-resolution stochastic process approach to model passes between teammates. These passes can be

viewed as directed dynamic relational links in a network. Multiplicative latent factors are introduced to study higher-order patterns in players' interactions that distinguish a successful game from a loss. Parameters are estimated using a Markov chain Monte Carlo sampler. Results in simulation experiments suggest that the sampling scheme is effective in recovering the parameters. We also apply the model to the first high-resolution optical tracking data set collected in college basketball games. The learned latent factors demonstrate significant differences between players' passing and receiving patterns in a loss, as opposed to a win. Our model is applicable to team sports other than basketball, as well as other time-varying network observations.

ROUTE IDENTIFICATION IN THE NATIONAL FOOTBALL LEAGUE

Chu, Dani†; Reyers, Matthew; Thomson, James; Wu, Lucas

Simon Fraser University, Vancouver, BC, Canada

† E-mail: *danic@sfu.ca*

Currently in football many hours are spent watching game film to manually label the routes run on passing plays. Using tracking data, each route can be described as a sequence of spatial-temporal measurements that varies in length depending on the duration of the play. We demonstrate how model-based curve clustering using Bernstein polynomial basis functions (i.e. Bézier curves) fit using the Expectation Maximization algorithm can cluster route trajectories. Each cluster can then be labelled to obtain route names for each route and create route trees for all receivers. Using the receiver route trees, we devise receiver metrics that account for receiver deployment. The resulting route labels can also be paired with film to enable streamlined queries of game film.

UNSUPERVISED RUN TYPE DETECTION

Gregory, Sam†

Sportlogiq, Montreal, QC, Canada

† E-mail: *sam@sportlogiq.com*

One of the major pitfalls of traditional event data in soccer is the inability to identify what players are doing off the ball. How are players finding space? How are they providing options for the player in possession?

Using player tracking data we are able to discretize player actions into off-ball runs. By identifying periods of acceleration and deceleration, we can pinpoint when players are making intentional off the ball runs. Building off of pre-existing work in basketball, we then employ Bezier curves to align individual runs in space and time, comparing similar runs and clustering them into distinct groups using functional clustering techniques. A second modelling approach uses an autoencoder

to compress the representation of off-the-ball runs into latent vectors, which are better suited to common clustering algorithms than the full representation.

This information can be used to identify players who act similarly, which run types are the most dangerous or effective against particular teams, and compare different off the ball run patterns that teams employ. Being able to identify and group these runs has applications across both pre- and post-match analysis and recruitment.

TREATMENT EFFECT HETEROGENEITY IN MLB BUNTING STRATEGIES

Lopez, Michael (1); Evans, Katherine† (2)

(1) The National Football League, New York, NY, USA; (2) Toronto Raptors, Toronto, ON, Canada

† E-mail: *causalkathy@gmail.com*

Bunting is down in Major League Baseball recently. This decline is generally attributed to simply looking at the change in expected runs for typical bunting scenarios. However, run expectancy varies greatly from batter to batter and across various potential bunting scenarios. We aim to better understand the heterogeneous treatment effect of bunting. Heterogeneity comes with respect to different game scenarios (runners on which bases, number of outs, inning, score difference etc) and types of hitters (OPS, speed, bunting ability, etc). We estimate the effect of bunting among those who bunted using Bayesian Additive Regression Trees (BART) as well as propensity score methods (matching and inverse weighting). We show that there are certain scenarios where bunting is advantageous even if the overall change in run expectancy is negative.

EXTRACTING PLAYER TRACKING DATA FROM VIDEO USING NON-STATIONARY CAMERAS AND A COMBINATION OF COMPUTER VISION TECHNIQUES.

Johnson, Neil†

ESPN Analytics, Bristol, CT, USA

† E-mail: *neil.johnson@espn.com*

Given the impact player tracking data has had on basketball and other sports as well as recent technological innovations in computer vision methodologies, the ability to extract player tracking data from non-stationary camera video feeds such as game broadcasts is now very accessible. This presentation will go over the key components and methods necessary to extract as much tracking data as possible, give optimal approaches for filling in the gaps and to tie everything together.

Using the broadcast video for a given game, we can first parse game context from the on-screen scoreboard and map it to each frame as well as each frame's timestamp. Then we can apply multi-person pose estimation using an open source library such as AlphaPose, which can detect each player on the court and use that to determine their center-of-mass and orientation. Then, using the distinct features of the court itself we can track the features positions frame-by-frame using a simple method such as OpenCV's Template Matching function. From there, we can apply smoothing, estimation, and interpolation filters to come up with a consistent set of player coordinates that can be translated to a standard 2-dimensional coordinate system using the distinct court features coordinates. Finally, we can attempt to bridge the absence of source video during essential live game segments using a variety of prediction techniques and assumptions inherent to the nature of the sport.

ANALYZING PLAYER PERFORMANCE IN ESPORTS

Clark, Nicholas (1); Macdonald, Brian† (2); Kloo, Ian (1)

(1) United States Military Academy, West Point, NY; (2) ESPN Analytics

† E-mail: *bmacnhl@gmail.com*

There are many similarities between analyzing player performance in a team eSports game like Defense of the Ancients 2 or League of Legends and in a traditional sport like basketball or hockey. Each team has 5 players who work together towards a common goal, while accumulating individual and team statistics that can be used to evaluate and analyze in-game performance. As opposed to having primarily isolated 1-on-1 matchups and discretizable events, these games are free-flowing with more many-on-many matchups, and determining a player's individual contribution to his or her team's success is difficult. We first discuss these similarities, as well as some of the main differences, in analyzing players in eSports and traditional sports. We then propose a Bayesian hierarchical regression model for assessing player performance that is similar to adjusted plus-minus models used in traditional sports like basketball, hockey, and soccer, and compare the results of the regression model with several "box score" type statistics. Finally, we explore the role that heroes, the in-game characters or avatars that competitors choose to play with, have on the measurable performance of a player, and options for how those differences can be accounted for.

A NAIVE BAYES APPROACH FOR NFL PASSING EVALUATION USING TRACKING DATA EXTRACTED FROM IMAGES

Mallepalle, Sarah† (1); Yurko, Ronald (1); Pelechrinis, Konstantinos (2); Ventura, Samuel L. (1)

(1) Carnegie Mellon University, Pittsburgh, PA, USA; (2) University of Pittsburgh, Pittsburgh, PA, USA

† E-mail: *sarahmallepalle@gmail.com*

The NFL collects detailed tracking data capturing the location of all players and the ball during each play. Although the raw form of this data is not publicly available, the NFL releases a set of aggregated statistics via their Next Gen Stats (NGS) platform. They also provide charts that visualize the locations of pass attempts for players throughout a game, encoding their outcome (complete, incomplete, interception, or touchdown). We present next-gen-scraPy: a publicly available framework designed to help close the gap between what data is available privately (to NFL teams) and publicly, and our contribution is twofold. First, we introduce an image processing tool designed specifically for extracting the raw data from the NGS pass chart images. We extract the outcome of the pass, the on-field location, and other metadata. Second, we analyze the resulting dataset and examine NFL passing tendencies and the spatial performance of individual quarterbacks and defenses. We introduce a generalized additive model for completion percentages by field location, and use a Naive Bayes approach for adjusting the 2-D completion percentage surfaces of individual teams and quarterbacks based on the number of their pass attempts. We find that our pass location data matches the NFL's official ball tracking data provided by the Big Data Bowl.

PLACKETT-LUCE MODELING WITH PARAMETRIC GROWTH CURVES FOR PREDICTING CAREER TRAJECTORIES

McKeough, Katy†; Glickman, Mark

Harvard University, Cambridge, MA, USA

† E-mail: *katy.mckeough@gmail.com*

Predicting the performance of an athlete from multi-competitor sports, such as track or skiing events, is an age-old question amongst sports analysts, coaches, and fans. Modeling athlete's abilities from the results of multi-competitor sports is a challenge because of the time-varying nature of abilities, but also because parametric models for rank orderings are typically more difficult to analyze than, for example, models for head-to-head competition. We propose a novel model to analyze time-varying multi-competitor game outcomes. Our model assumes that outcomes follow a Plackett-Luce model for ranks conditional on model parameters. The Plackett-Luce parameters are assumed to be a mixture of parametric growth curves, with the number of mixture components chosen through model selection. The growth curve model component is a flexible, non-linear mixed effects model that incorporates time, player age and other time-dependent, player-specific covariates.

An advantage of modeling time variation in abilities through growth curves is the ability to make predictive statements of future athlete ratings. We apply this method to professional women's luge and show how it is easily generalizable to other sports.

ESTIMATION OF PLAYER LOAD METRICS USING BROADCAST-DERIVED TRACKING DATA

Mortensen, Jacob† (1); Bornn, Luke (1,2)

(1) Simon Fraser University, Burnaby, BC, Canada; (2) Sacramento Kings, Sacramento, CA, USA

† E-mail: *jmortens@sfu.ca*

The introduction of optical tracking data across sports has given rise to the ability to dissect athletic performance at a level unfathomable a decade ago. One specific area that has seen substantial benefit is sports science, as high resolution coordinate data permits sports scientists to have to-the-second estimates of acceleration load and density–metrics traditionally used to understand the physical toll a game takes on an athlete. Unfortunately, collecting this data requires installation of expensive hardware and paying hundreds of thousands of dollars in licensing fees to data providers, restricting its availability. Algorithms have been developed that allow a traditional broadcast feed to be converted to x-y coordinate data, making tracking data easier to acquire, but coordinates are available for an athlete only when that player is within the camera frame. Obviously, this leads to inaccuracies in player load estimates, limiting the usefulness of this data for sports scientists. In this research, we use games for which both full optical tracking data and broadcast data are available to develop models that predict offscreen player load metrics. We compare the performance of various approaches, including a simple scaled estimator and two models that simulate offscreen movement: an auxiliary regression model in conjunction with B-splines, and a nonstationary Gaussian process model. Our work is the first to measure the utility of broadcast feeds in estimating physical load metrics across soccer and hockey, demonstrating situationally the strengths and weaknesses of broadcast-derived tracking data for understanding these metrics.

MEASURING COMPETITIVE BALANCE CORRECTLY (IN SPORTS)

Doria, Matthew (1); Nalebuff, Barry† (2)

(1) National Basketball Association, New York, NY, USA; (2) Yale University, New Haven, CT, USA

† E-mail: *barry.nalebuff@yale.edu*

In a well-balanced sports league, teams are evenly matched, game are exciting, and championships are hard to predict. The most commonly used measure of competitive balance in a given season – what we call static competitive balance – is the Noll-Scully score. Contrary to its design, this score does not accurately reflect the relative competitive balance of leagues with different season lengths. The basic error is that the score artificially inflates the measured imbalance for leagues with long seasons (e.g., MLB) compared to those with short seasons (e.g., NFL). We provide a new score that is simple to compute and, true to the motivation of Noll-Scully, is neutral to the season length. The result of using the new score is a reversal of commonly held views regarding which sport leagues have the greatest level of static competitive balance: the NFL goes from having the most balance to being tied for the least, while MLB becomes the sport with the most balance. While our new measure provides an unbiased comparison of the underlying variance in team win probabilities across sports leagues, like Noll-Scully it does not provide direct insight into competitive balance at the game level – when a team that wins 60% of its games faces a rival that wins 40%, the stronger team wins more than 60% of the time. This leads us to a new measure of competitive balance based on variance at the game-level. We find that variance at the game level is almost double that at the team level. To measure competitive balance at the season-level requires a different measure, one that takes into account season length. We look at the disparity between the different teams' chances to come out on top at season end. Here the NBA uniquely stands out for having the most predictable results and hence the least amount of full-season competitive balance.

MEASURING SPATIAL ALLOCATIVE EFFICIENCY IN PROFESSIONAL BASKETBALL

Sandholtz, Nathan† (1); Mortensen, Jacob (1); Bornn, Luke (1,2)

(1) Simon Fraser University, Burnaby, BC, Canada; (2) Sacramento Kings, Sacramento, CA, USA

† E-mail: *nsandhol@sfu.ca*

Implicit in any discussion of shot efficiency in professional basketball is the fact that inefficient players have a negative impact because basketball is a team sport; a given player's inefficient shots take higher value shot opportunities away from teammates. This aspect of efficiency–the allocation of shots within a lineup–is the primary focus of our work. Allocative efficiency is fundamentally a

spatial problem because the distribution of shot attempts within a lineup is highly dependent on court location. The main idea behind our approach is to compare a player's field goal percentage (FG%) to his field goal attempt (FGA) rate in context of both his four teammates on the court and the spatial distribution of his shots. To this end, we build Bayesian hierarchical models to estimate player FG% and FGA rates at every location in the offensive half-court using publicly available data from the National Basketball Association. Then, by pairing a player's lineup-specific FGA rankings with his corresponding FG% rankings, we can detect areas where the lineup exhibits inefficient allocation of shots, estimate and visualize the points that are consequently lost, and identify which players are responsible. We estimate uncertainty in our metrics using posterior draws from the FG% surfaces. Lastly, we analyze the impact that deviations from optimality have on a team's overall winning potential by incorporating the these metrics within an adjusted plus-minus model, elucidating the relationship between a lineup's shot selection optimality and its per-possession production.

CLASSIFYING AND ANALYSING TEAM STRATEGY IN PROFESSIONAL SOCCER MATCHES

Shaw, Laurie[†]

Harvard University, Cambridge, MA, USA

[†] E-mail: *laurie.shaw@cfa.harvard.edu*

One of the most important tactical decisions that a soccer manager must make is to determine the spatial configuration of the team (i.e., formation) during different phases of a game, such as while defending, attacking or following transitions. The selection of formation influences how aggressively a team attacks, where they focus their attacks, and their overall playing style. In this talk we present a new technique for measuring, classifying and studying team formations in professional soccer matches. Using player tracking data from a season's worth of matches, we measure the relative positioning of each team's players, both in and out of possession of the ball, over successive intervals within each match. Applying hierarchical agglomerative clustering - using the Wasserstein metric to compare formations - we identify the distinct categories of offensive and defensive formations that were employed. We use the learned categories, in combination with Bayesian model selection criteria, to classify the formations adopted by teams during their matches, providing a tactical summary of each match. We explore each team's preferred formations, investigate how team strategy varied according to the opponent, and study how managers reacted tactically to key events during their matches. Finally, we discuss how formation choices relate to playing style, and discuss other potential applications of our methodology.

HOW DO TYPICAL RUNNERS' PERFORMANCES VARY WITH AGE AND GENDER?

Smith, Richard L[†]

University of North Carolina, Chapel Hill, NC, USA

[†] E-mail: *rls@email.unc.edu*

All runners get slower as they age, and in the vast majority of events, women are slower than men. But how should one quantify the differences? One widely used method is age-graded times, but being based on world record performances they may not correspond to performances by ordinary runners. Furthermore, many large races (especially, the Boston Marathon) impose qualifying times for guaranteed entry, but the standards are not based on detailed comparisons between age groups. This study (based on data from the Boston Marathon) aims to quantify the age-gender discrepancies based on typical runners' performances. A mixed effects model is proposed to estimate the time-age curves for both men and women using random effects to account for differences among runners. The results show marked discrepancies from both age-graded curves and from the age-gender relationships that are implicit in the Boston Marathon standards. However, the data are limited and more detailed analysis of a larger dataset is proposed to validate these conclusions.

GOING DEEP: MODELS FOR CONTINUOUS-TIME WITHIN-PLAY VALUATION OF GAME OUTCOMES IN AMERICAN FOOTBALL WITH TRACKING DATA

Yurko, Ronald[†] (1); Matano, Francesca (1); Richardson, Lee F (1); Granered, Nicholas (2); Pospisil, Taylor (1); Pelechrinis, Konstantinos (2); Ventura, Samuel L (1)

(1) Carnegie Mellon University, Pittsburgh, PA, USA; (2) University of Pittsburgh, Pittsburgh, PA, USA

[†] E-mail: *ryurko@andrew.cmu.edu*

Continuous-time assessments of game outcomes in sports have become increasingly common in the last decade. In American football, only discrete-time estimates of play value were possible, since the most advanced public football datasets were recorded at the play-by-play level. While measures like expected points (EP) and win probability (WP) are useful for evaluating football plays and game situations, there has been no research into how these values change throughout a play. In this work, we make two main contributions: First, we provide a general framework for continuous-time within-play valuation in the National Football League using the Next Gen Stats player and ball tracking data. Our framework incorporates several modular sub-models, so that other recent work involving player tracking data in football can be easily incorporated. Second, we construct a ball-carrier model to estimate how many yards the ball-carrier will gain conditional on the locations and trajectories of all players. We test several modeling approaches, and ultimately use a long

short-term memory recurrent neural network to continuously update the expected end-of-play yard line. This prediction is fed into between-play EP/WP models, yielding a within-play value estimate. The framework is modular, so that existing models, eg for pass attempt outcomes or quarterback decision-making, can be applied within this framework. Finally, the fully-implemented framework allows for continuous-time assessment of all 22 players on the field, which was never before possible at such a granular level.

Poster Presentation Abstracts

HIERARCHICAL MODELING TO PREDICT SHOT OUTCOME IN NBA GAMES

Siegel, Spencer; Barlow, Daniel†; Bury, James; Smith, Richard

University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

† E-mail: *dcbarlow@live.unc.edu*

Knowing the probability of a shot being made is essential in basketball. With the amount of raw information available from proprietary NBA play-by-play data, we saw the potential to quantitatively analyze shot quality. There are many different factors that can affect the probability of a shot going in such as distance, nearest defender distance, shot type, time left in the quarter, home/away, etc. In this paper, we look at trying to predict the probability of a shot being successful for 122 players during the 2014-15 NBA regular season. Since a subset of our predictors affect each player differently, we used a mixed effects random model to improve predictions based on unique skill sets of players. Because the best initial results utilized decision trees, we adapted our mixed effects model to include indicator variables corresponding to specific shot distance ranges and closet defender distances. These ranges were chosen by modeling shot outcome using a standard decision tree.

AN EXAMINATION OF TIMEOUT VALUE, STRATEGY, AND MOMENTUM IN NCAA DIVISION 1 MEN'S BASKETBALL

Benz, Luke†

Yale University, New Haven, CT, USA

† E-mail: *lukesbenz@gmail.com*

Fans watching a basketball game often believe that they can sense when one team has "momentum." Coaches seem to take timeouts when their team is on a negative scoring run, feeling pressure to

stop an opponent's quick flurry of scoring. This work examines how timeouts are used in NCAA Division 1 men's basketball and whether there is any truth to the notion that timeouts stop opponent momentum by decreasing the rate of opponent scoring or swinging the rate of scoring in favor of the timeout-calling team. Additionally, this work attempts to quantify the value of a taking a timeout throughout the course of the game. Overall, this work yields an estimate that on average, teams perform between 1.5–2.2 points better in five minute intervals following called timeouts compared to intervals of equal length preceding the timeout.

IN OR OUT? THE NEW FLAGSTICK DILEMMA FOR PUTTING IN GOLF

Bilder, Christopher R[†]

University of Nebraska-Lincoln, Lincoln, NE, USA

[†] E-mail: *bilder@unl.edu*

The United States Golf Association adopted rule changes in 2019 to simplify and speed up the game of golf. Before 2019, golfers were assessed a one-shot penalty if a putted ball made contact with the flagstick. Now under Rule 13.2(a), golfers have the option to either remove the flagstick or leave it in the hole without penalty when putting on a green. No consensus on the better strategy has emerged, even among professional golfers. The Edoardo Molinari Golf Academy recently performed an experiment to solve this dilemma, with its analysis relying solely on comparing the observed proportion of holed putts. The purpose of our presentation is to examine their data using statistical modeling and inference methods. With our analysis, we agree with their conclusion that it is better for the flagstick to be out for a ball approaching the hole at a moderate speed. However, we do not necessarily agree with their conclusion that leaving the flagstick in is better for a ball approaching the hole at a fast speed. While the observed proportions suggest a possible benefit with leaving it in, the data does not provide sufficient evidence that this can be generalized beyond the sample.

ESTIMATING (FOUR) FACTOR VALUES IN THE NBA: A SEEMINGLY UNRELATED REGRESSION ANALYSIS

Bosch, Jonathan[†]; Speakman, Dax; Sanders, Shane

Syracuse University

[†] E-mail: *jbosch@syr.edu*

We consider the four factors model of basketball output constructed by Oliver (2004). Using data from stats.nba.com, we construct player-level factor performance data on factor performances for each NBA free agent from 2012 through 2018. Importantly, this data source contains pioneering (public) data on player-level shot defense such that an NBA player's factor value is now fully

observable. We also collect free agent contract data for the same period using spotrac.com. From this, we are able to estimate the marginal effect of units of (player) factor improvement upon the score margin per 100 possessions. Using seemingly unrelated regression (SUR), we also estimate the effect of a unit score margin improvement per 100 possessions upon a player's subsequent free agency salary. As in SUR ordinary least squares regressions, these two equations are estimated simultaneously as a system of (error-term related) equations. On average, we estimate that offensive factor improvements are approximately 2.5 times as valuable on the free agency market as are equal (in terms of score margin implication) defensive factor improvements. We also find considerable and significant heterogeneity within the implied salary returns within the set of offensive factors and within the set of defensive factors. Subsequent computations support the conclusion that a win-maximizing team can engage in win-maximization arbitrage on the NBA free agency market, whereby players whose win value arises from relatively expensive factors are shed in favor of those whose win value arises from relatively inexpensive factors.

THE PATHS WE TAKE: A PLAYER TRACKING ANALYSIS OF NHL SHOOTOUTS

Czuzoj-Shulman, Nick[†]

Sportlogiq, Montreal, QC, Canada

[†] E-mail: *nick@sportlogiq.com*

In the NHL, the shootout takes place if a tie remains at the end of overtime by pitting a single shooter against the opposing goalie in a mano a mano bid to seal the game. In the 2018-2019 NHL Regular season, there were nearly 400 shootout attempts, with shooters scoring 31% of the time. Shootout wins for this season accounted for 7%, on average, of a teams' total wins and given how important each point is in the standings, finding a competitive advantage in these end of game scenarios could be the difference between making the playoffs and ending the season early.

Research on shootouts until now has measured shooting success rates based on observable aspects such as shot types, handedness, shot placements, and distances from the net. With the advances in player tracking data, we consider the above along with the paths that players take leading up to their attempt, in addition to changes in speed and acceleration prior to the shot. In observing these tracks and speeds we find a new understanding of shootouts that was not previously possible. For instance, left-handed shooters who take a right to left path prior to their shot score more than 45% of the time, whereas right-handed shooters skating left to right score only 26% of the time. With these new findings, we introduce a shootout expected goals model as well as more optimal strategies for teams and goalies to consider when the clock runs out on overtime.

UNSUPERVISED METHODS FOR IDENTIFYING PASS COVERAGE AMONG DEFENSIVE BACKS WITH NFL PLAYER TRACKING DATA

Dutta, Rishav† (1); Yurko, Ronald (1); Ventura, Samuel (1, 2)

(1) Carnegie Mellon University, Pittsburgh, PA, USA; (2) Pittsburgh Penguins, Pittsburgh, PA, USA

† E-mail: *rishavd@andrew.cmu.edu*

Analysis of player tracking data for American football is in its infancy, since the National Football League (NFL) released its Next Gen Stats tracking data publicly for the first time in December 2018. While tracking datasets in other sports often contain detailed annotations of on-field events, annotations in the NFL's tracking data are limited. Methods for creating these annotations typically require extensive human labeling, which is difficult and expensive. We begin tackling this class of problems by creating annotations for pass coverage types by defensive backs using unsupervised learning techniques, which require no manual labeling or human oversight. We define a set of features from the NFL's tracking data that help distinguish between "zone" and "man" coverage. We use Gaussian mixture modeling and hierarchical clustering to create clusters corresponding to each group, and we assign the appropriate type of coverage to each cluster through qualitative analysis of the plays in each cluster. We find that the mixture model's "soft" cluster assignments allow for more flexibility when identifying coverage types. Our work makes possible several potential avenues of future NFL research, and we provide a basic exploration of these in this paper.

TRAP: A PREDICTIVE FRAMEWORK FOR TRAIL RUNNING ASSESSMENT PERFORMANCE

Fogliato, Riccardo†; Oliveira, Natalia Lombardi; Yurko, Ronald

Carnegie Mellon University, Pittsburgh, PA, USA

† E-mail: *rfogliat@andrew.cmu.edu*

Trail running is an endurance sport in which athletes face severe physical challenges. Due to the growing number of participants, the organization of limited staff, equipment, and medical support in these races now plays a key role. Determining when a runner needs medical help is a difficult task that requires knowledge of the terrain and of the runner's abilities. In the past, choices were solely based on the organizers' experience without reliance on data. However, this approach is neither scalable nor transferable. Instead, we propose a firm statistical methodology to perform this task, both before and during the race. Our proposed framework, TRAP, studies (1) the assessment of the runner's ability to reach the next aid station, (2) the prediction of the runner's passage time at the next aid station, leading to (3) the quantification of the likelihood that the runner needs medical help. To obtain data on the ability of runners, we introduce a Python package, scrapITRA,

to access the race history of runners from the International Trail Running Association (ITRA). We apply our methodology, using the ITRA data along with checkpoint and terrain-level information to the "holy grail" of ultra-trail running, the Ultra-Trail du Mont-Blanc (UTMB), demonstrating the predictive power of our methodology that can be implemented for improving the detection of medical assistance in trail running.

A MIXED EFFECTS MULTINOMIAL LOGISTIC-NORMAL MODEL FOR FORECASTING BASEBALL PERFORMANCE

Gerber, Eric AE† (1); Craig, Bruce A (2)

(1) California State University, Bakersfield, CA, USA; (2) Purdue University, West Lafayette, IN, USA

† E-mail: *gerber19@purdue.edu*

Prediction of player performance is a key component in the construction of baseball team rosters. Traditionally, the problem of predicting plate appearance outcomes for a season has been approached univariately; focusing on each outcome separately rather than modeling the collection of outcomes jointly. Recently, there has been greater effort to account for the correlations between outcomes. However, most of these state of the art prediction models are the proprietary property of teams or industrial sports entities and little is available in open publications.

This research introduces a joint modeling approach to predict season plate appearance outcome vectors using a mixed-effects multinomial logistic-normal model. This approach accounts for both positive and negative correlations between outcomes across and within players. The purpose of this methodology is to provide a theoretically sound approach to predict baseball player performance. It is applied to the problem of predicting performance for players moving between the Japanese and American major leagues.

Not only does this model represent a significant contribution to the area of joint outcome prediction but its application area is one that has not yet been addressed. Specifically, the direct application of prediction models to players moving between Japanese and American baseball leagues is an unexplored subject. We describe the modeling approach and apply the method to longitudinal multinomial count data of baseball player-seasons for players moving between the Japanese and American major leagues.

RETRODICTIVE MODELLING IN MODERN RUGBY UNION

Hamilton, Ian†

University of Warwick, Warwick, UK

† E-mail: *i.hamilton@warwick.ac.uk*

In modern rugby union tournaments points are awarded for wins, draws, and losses, but also for losing within a particular score margin, and for scoring a particular number of tries. This means a system of five result outcomes and an additional bonus point. In schools rugby union, matches take place based on geographical and historical ties, and thus do not have the features of a typical round robin tournament. Teams will only play a subset of the other teams in the tournament, giving fixture schedules of varying difficulty, varying size, and with no systematic home or away status. In this poster we present a retrodictive model applied to rugby union for the purpose of ranking teams when schedule strengths differ. In doing so we extend the famous Bradley-Terry model beyond the known win/draw/loss paradigm in a manner fully consistent with the full round-robin tournament norms. We investigate the use of a prior for the purpose of calibrating the ranking with respect to the number of matches played, as well as for enforcing connectivity. We conclude by applying the model to a schools rugby tournament in England, and using it to assess their current ranking methodology.

EVALUATING THE QUALIFYING TIMES OF THE BOSTON MARATHON

Hammerling, Dorit† (1); Albrecht, Laura (1); Ring-Jarvi, Ross (1); Smith, Richard (2)

(1) Colorado School of MInes, Golden, CO, USA; (2) University of North Carolina, Chapel Hill, NC, USA

† E-mail: *hammerling@mines.edu*

The Boston Marathon is one of the most prestigious running races in the world. From its inception in 1897, popularity grew to a point in 1970 where qualifying times were implemented to limit the number of participants. Currently, women's qualifying times are thirty minutes slower in each age group than the men's qualifying times equating to a 16.7% adjustment for the 18-34 age group, decreasing with age to a 10.4% adjustment for the 80+ age group. This setup counter-intuitively implies that women become faster with age relative to men. We present a data-driven approach to investigate this issue, working under the paradigm that "fair" qualifying standards should lead to an equal proportion of qualifiers in each age category and gender. Initial results, using a constant percentage adjustment instead of an absolute difference, show that an adjustment of 14-15% in every age category yields a more balanced field of qualifiers.

A BAYESIAN JOINT MODEL FOR SPATIAL POINT PROCESSES WITH APPLICATION TO BASKETBALL SHOT CHART

Jiao, Jieying†; Hu, Guanyu; Yan, Jun

University of Connecticut, Storrs, CT, USA

† E-mail: *jieying.jiao@uconn.edu*

The success rate of a basketball shot may be higher at locations in the court where a player makes more shots. In a marked spatial point process model, this means that the marks are dependent on the intensity of the process. We develop a Bayesian joint model of the mark and the intensity of marked spatial point process, where the intensity is incorporated in the model for the mark as a covariate. Further, we allow variable selection through the spike-slab prior. Inferences are developed with a Markov chain Monte Carlo algorithm to sample from the posterior distribution. Two Bayesian model comparison criteria, the modified Deviance Information Criterion and the modified Logarithm of the Pseudo-Marginal Likelihood, are developed to assess the fitness of different models focusing on the mark. The empirical performances of the proposed methods are examined in extensive simulation studies. We apply the proposed methods to the shot charts of four players in the NBA's 2017–2018 regular season to analyze the shot intensity in the field and the field goal percentage. The results suggest that the field goal percentages of these players are significantly positively dependent on their shot intensities, and that different players have different predictors for their field goal percentages.

USING SCOUTING REPORTS TEXT TO PREDICT NCAA→NBA PERFORMANCE

Maymin, Philip†

Dolan School of Business, Fairfield University, Fairfield, CT, USA

† E-mail: *philip@maymin.com*

I train nearly one thousand LOOCV individual random forests machine learning models, one for each collegiate NBA draft prospect who was in the ESPN Top 100 prospects list prior to the draft, using their college efficiency and production stats, combine measurement information, high school RSCI and NBA mock draft placements, handedness, ethnicity as estimated from a separate machine learning model based on their name (appeler/ethnicolr on github), and also scouting evaluations and raw scouting text scraped from nbadraft.net for 2006-2019 and processed through term frequency-inverse document frequency (TFIDF) then doubly dimension reduced. The forecast variable is an average of three win production measures (Win Shares, Wins Produced, and Estimated Wins Added) over each player's three years after the draft, or zero for years they did not play.

The correlation between predicted and actual NBA production is 63%. Every team except the Denver Nuggets would have benefited from drafting based on this model rather than the decisions

they actually made, even though the model does not have hindsight bias and can only draft NCAA prospects. The average model pick outperformed the actual pick by 70% and the average team lost out on $100 million worth of on-court production.

I conclude that NBA teams should incorporate their own internal historical scouting reports into their projection models using the techniques I outline here. For 2019, the model values Brandon Clarke and Grant Williams much higher than their mocks, suggesting they will be the sleepers of the draft.

A SHINY MARKOV MACHINE FOR DECISION-MAKING IN MAJOR LEAGUE BASEBALL

Osborne, Jason A† (1); Post, Justin B (1); Wen, Melody (2)

(1) North Carolina State University, Raleigh, NC, USA; (2) North Carolina School of Science and Mathematics, Durham, NC, USA

† E-mail: *jaosborn@ncsu.edu*

We present a shiny app enabling the user to carry out Markov Chain calculations to assist with decision-making in baseball. These decisions, made both during and before the game, use summary statistics to estimate transition probabilities. The Markovian assumption leads to the calculation of the entire probability distribution of runs scored in the remainder of a game. An example of such a decision is whether or not to attempt to steal a base. While conventional analysis has phrased this problem in terms only of the base stealer's chance of success, a more informed decision would take account of the sequence of hitters who follow the batter in the lineup (and all other available information). Other illustrations of this machinery include selection of batting order at the beginning of a game and deciding whether or not to attempt a sacrifice bunt or pinch-hit. The app allows users to select the following inputs: MLB team and year, an ordering of nine players from each team, inning, outs and runners on base.

SEMIPARAMETRIC SCORING RATE ESTIMATION IN EUROPEAN SOCCER

Ritchie, Robyn†; Leblanc, Alexandre

University of Manitoba, Winnipeg, MB, Canada

† E-mail: *ritchi12@myumanitoba.ca*

We analyzed five seasons of soccer across four different leagues including the English Premier League, Bundesliga, La Liga and Serie A between 2014-2019. Game event times, to the second, have been obtained from game commentaries through web scraping and lead to data on more than 20,000 goals

over a combined 7320 games. From these data, we study semiparametric scoring rate estimation (league-wide and team-specific) under the assumption that goal times follow a non-homogeneous Poisson process model. Specifically, scoring rates are decomposed into a scoring pattern which smoothly changes over the course of the game, which we then scale-up by the expected number of goals for each team/league based on different aspects of the game. We rescale all goal times over the course of a 90-minute plus additional time game to be between $[0, 1]$ and we look at specific questions related to team performance, including whether scoring patterns differ between home and away games and between the two halves. The approach can also be used to compare scoring rates between leagues and/or seasons and study scoring rates throughout extra time. It can also be expanded to examine the patterns of other events such as red cards, substitutions, or player specific scoring patterns and much more.

PLUS-MINUS MODELS FOR AMERICAN FOOTBALL

Sabin, Paul[†]

ESPN, Bristol, CT, USA

[†] E-mail: *paul.sabin@espn.com*

Plus-Minus models have long been used in Basketball (Rosenbaum (2004), Kubatko et al. (2007), Winston(2009), Sill(2010)) to establish each players value by accounting for those on the court at the same time. Few sports have the lineup changes and scoring frequency as basketball. More recent methods have found ways to incorporate plus-minus models in sports such as Hockey (Macdonald (2011)) and Soccer (Schultze and Wellbrock (2018) and Matano et al. (2018)). These models are especially useful in coming up with results-oriented estimation of each player's value. In American Football, it is difficult to estimate every player's value since many positions such as offensive lineman have no recorded statistics. While player-tracking data in the NFL is allowing new analysis, such data does not exist in other levels of football such as the NCAA. Using player participation data available for college football and the NFL, I provide a framework to create plus-minus models that estimate the value of every player per play in a football league over the course of a season. One by-product of this methodology is the model evaluating the importance of every position in football.

THE PATH TO SUCCESS ON THE LPGA TOUR – IDENTIFYING THE THREE IMPORTANT DIMENSIONS IN STROKE PLAY

Sen, Kabir C.[†]

Lamar University, Beaumont, TX, USA

[†] E-mail: *kabir.sen@lamar.edu*

The LPGA tour attracts a wide range of talent from across the globe and is highly competitive with a diverse group of champions. This paper identifies the three dimensions of a golfer's performance

which determine her success on the tour. Here, success is defined as the money earned per event. The three dimensions are based on the statistics available on the LPGA Tour website. These are: the average putts per greens hit in regulation, the average putts per greens missed in regulation and the average non-putting strokes used per hole. The results show that the last statistic has a high negative correlation with the percentage of greens hit in regulation (i.e., over 0.90). Regression analysis for the five years of the study (2014 through 2018) shows that each of the three measures has a significant influence on the money earned per event. However, the most influential elements appear to be the average non-putting strokes per hole and the average putts per greens hit in regulation. Standardizing each of the three elements of stroke play for each year provides an opportunity to evaluate comparative performance over time. An aggregate compilation of the three measures has a high rank correlation with the money earned per event (absolute value of the Spearman rank correlation > 0.90). The golfers who consistently score well on the composite measure are identified as the best LPGA golfers over the last five years.

META-METRICS TO QUANTIFY PROPERTIES OF QUARTERBACK STATISTICS

Stiller, Julia† (1); Lopez, Michael (2)

(1) Skidmore College, Saratoga Springs, NY, USA; (2) The National Football League, New York, NY, USA

† E-mail: *jstiller@skidmore.edu*

Despite recent growth in public football analytics, there is limited research on how effectively certain metrics isolate player talent. For example, quarterback performance is generally analyzed by using box score summaries such as completion percentage and quarterback rating in combination with newer tools like win probability and expected points. However, it is both unknown if these complex modern metrics offer an advantage over traditional ones and if, when taken wholly, quarterback statistics can consistently discriminate player talent. Using the meta-metric framework of Franks et al. (2016), we aim to quantify properties of quarterback metrics to better understand how performance measures vary over time, between players, and within players. Results using game and season-level data from 2009-2018 suggest that expected-point based summaries tend to feature preferred statistical properties, yet these game-level metrics compare relatively poorly to those in other professional sports.

FROM GRAPES AND PRUNES TO APPLES AND APPLES: USING MATCHED METHODS TO ESTIMATE OPTIMAL ZONE ENTRY DECISION-MAKING IN THE NATIONAL HOCKEY LEAGUE

Toumi, Asmae† (1); Lopez, Michael (2)

(1) Columbia University, New York, NY, USA; (2) National Football League, New York, NY, USA

† E-mail: *asmae.toumi@columbia.edu*

Previous research in the National Hockey League has suggested that teams' decisions to gain the offensive zone with puck possession ("carry-ins") is preferred over dumping the puck in and chasing after it ("dump-ins"). However, standard comparisons of zone entry strategy are confounded by factors such as offensive and defensive talent, location on the ice, and shift time, each of which impact player choice. Indeed, contrasting carry-ins to dump-ins isn't exactly an apples-to-apples comparison; instead, it is more like studying grapes versus prunes. Using two matching methods – propensity score matching and Bayesian additive regression trees – we leverage player-tracking data to estimate the causal benefits due to zone-entry decisions. Both approaches better account for the variables that affect entry choice. We also highlight the wide-ranging potential of the causal inference framework with player tracking data in sports while emphasizing the challenges of using standard statistical methods to inform decision-making in the presence of substantial confounding.

HOME SWEET HOME: QUANTIFYING HOME COURT ADVANTAGES FOR NCAA BASKETBALL STATISTICS

van Bommel, Matthew†; Bornn, Luke; Chow-White, Peter; Gao, Chuancong

Simon Fraser University, Burnaby, BC, Canada

† E-mail: *matthew.vbommel@gmail.com*

Box score statistics are the baseline measures of performance for National Collegiate Athletic Association (NCAA) basketball. Between the 2011-2012 and 2015-2016 seasons, NCAA teams performed better at home compared to on the road in nearly all box score statistics across both genders and all three divisions. Using box score data from over 100,000 games spanning the three divisions for both women and men, we examine the factors underlying this discrepancy. The prevalence of neutral location games in the NCAA provides an additional angle through which to examine the gaps in box score statistic performance, which we believe has been underutilized in existing literature. We also estimate a regression model to quantify the home court advantages for box score statistics and compare the magnitudes to other factors such as increased number of possessions, and team strength. Additionally, we examine the biases of scorekeepers and referees. We present evidence that scorekeepers tend to have greater home team biases when observing men compared to women, higher divisions compared to lower divisions, and stronger teams compared to weaker teams. Finally, we present statistically significant results indicating referee decisions are impacted by attendance, with larger crowds resulting in greater bias in favor of the home team.

PREDICTING THE DIRECTION OF SERVE IN PROFESSIONAL TENNIS USING BY CONTEXTUAL INFORMATION AND MACHINE LEARNING

Yamamoto, Hiroyuki† (1, 2); Kudo, Kazutoshi (1); Buszard, Tim (3, 4); Reid, Machar (3, 4); Farrow, Damian (3, 4); Kovalchik, Stephanie A (3, 4)

(1) The University of Tokyo, Tokyo, Japan; (2) Japan Society for the Promotion of Science, Tokyo, Japan; (3) Victoria University, Melbourne, Australia; (4) Game Insight Group, Tennis Australia, Melbourne, Australia

† E-mail: *hyamamoto.mcml@gmail.com*

The purpose of this study was to examine whether machine learning model (i.e. random forest, logistic regression) using contextual information (CI) improved prediction accuracy of serve direction (to receiver's forehand (FH) or backhand (BH)). Several categories of CI variables were considered: player characteristics, player performance, score, environment. The data sample included 999 matches and 75538 points from male professional singles matches from 2013 to 2017. The servers in the sample included 25 players ranked in the top 30. Results show that there is a significant negative correlation between average aces in a match and BH proportions ($r = -0.57$, $p < .01$) and left-handed servers target the BH more than right handed servers ($p < .05$). Our prediction model increased prediction accuracy for the serve direction of twenty and seventeen players comparing to naive model without CI for Random forest model and Logistic regression model, respectively. We compared our model's prediction accuracy (Ao) against the accuracy of naive model (An) by Improvement index as $I = 100\% \times (Ao - An)/An$. Improvements were 7.97 and 6.72% for Random forest model and Logistic regression model, respectively. Machine learning model using CI can provide strategic advantage for predicting the service behavior of elite opponents.

ON THE UTILITY OF TRACKING DATA FOR AUGMENTING EVENT-BASED MODELS IN HOCKEY AND SOCCER

Keane, Evin; Yu, David†

Sportlogiq, Montreal, QC, Canada

† E-mail: *david.yu@sportlogiq.com*

Expected Goals (xGoal) and Expected Pass Completion (xPass) models have become standard metrics to benchmark player and team performance in both soccer and hockey. Current models are typically built using spatio-temporal event datasets that largely capture only the locations of on-ball/on-puck actions. These event datasets typically don't capture context surrounding the action such as the amount of defensive pressure on the ball/puck carrier, and the locations of defenders surrounding the target. Where concepts of pressure do exist, they are typically encoded as binary

variables and therefore fail to capture that the number of defenders applying pressure varies, as does the amount of pressure applied by each defending player.

By combining event and tracking data streams, we develop a tracking augmented event dataset that accounts for the positions of all nearby offensive and defensive players at the time of the event. This allows us to model of defensive pressure on both the shooter/passer and their targets on a continuous scale. We then build xGoal and xPass models using this augmented even dataset and show that these models outperform those built using standard event datasets.

For example, we demonstrate that the players widely considered the English Premier League's best passers correspond with those whose numbers are most improved by incorporating tracking data. Players whose passing ability seemed under-rated with event data but whose numbers improved significantly with the inclusion of tracking-based features include Eden Hazard, Jorginho, Mezut Özil and Raheem Sterling.