PROGRAM

October 1 – 22, 2021

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

Acknowledgments: We wish to thank everyone who helped to make NESSIS possible. Special thanks go to Kevin Rader, Luke Bornn, Katy McKeough, Mingfei Li, Karen Barkow, Michelle Monestime, the Harvard University Department of Statistics, the Section on Statistics in Sports of the American Statistical Association, the Harvard Sports Analytics Collective, and the Sports Analytics Lab at Harvard University for their parts in helping with the symposium.

Looking forward to seeing everyone in person in 2023.

2021 New England Symposium on Statistics in Sports

October 1, 2021 - Introduction

12:45pm – 1:00pm: Mark Glickman and Scott Evans - Welcome Address

October 1, 2021 - Analyses with Tracking and Broadcast Data

1:00pm – 1:30pm: Sam Gregory, Victoria University and Inter Miami CF
"Pace and Power: Removing Unconscious Bias from Soccer Broadcasts"

1:30pm – 2:00pm: Ethan Baron, University of Toronto
"Predictive Value of Off-Target Shots in Soccer"

2:00pm – 2:30pm: Craig Fernandes, University of Toronto
"A Markov Process Approach to Untangling the Relationship between Intention vs. Execution in Tennis"

October 8, 2021 - Causation, Bias and Transformations

1:00pm – 1:30pm: Sean Fischer, University of Pennsylvania
"Causal Effect of Playing Time in the NBA"

1:30pm – 2:00pm: Jonathan Che, Harvard University
"Athlete Rating in Multi-Competitor Games with Scored Outcomes via Monotone Transformations"

2:00pm – 2:30pm: Luke Benz, Harvard University; and
Michael Lopez, NFL and Skidmore College
"Estimating the Change in Soccer... Home Advantage During the COVID-19 Pandemic"

October 8, 2021 - Probability Modeling and Decision-Making

3:30pm – 4:00pm: Nathan Sandholtz, Brigham Young University
"An Inverse Optimization Analysis of the Fourth Down Decision in Football"

4:00pm – 4:30pm: Chancellor Johnstone, Air Force Institute of Technology
"Which Teams Would Have Won the 2020 NCAA Men's and Women's Basketball Tournaments?"

4:30pm – 5:00pm: Brook Russell and Sydney Newman, Clemson University
"Developing Multivariate Extreme Value Methods to Explore NFL Prospect Data"

October 15, 2021 - Tactics in Soccer and Australian Football

9:00am – 9:30am: Jeremy Alexander, Victoria University
"Quantifying Congestion in Australian Rules Football"

9:30am – 10:00am: Jirka Poropudas, SportIQ
"Extended Model for Expected Threat in Soccer"

10:00am – 10:30am: Marius Oetting, Bielefeld University
"A Copula-based Hidden Markov Model for Classification of Tactics in Football

October 15, 2021 - Novel Methods in Soccer

1:00pm – 1:30pm: Daniel Daly-Grafstein, University of British Columbia
"Quantifying League-Independent Scoring Ability in Soccer"

1:30pm – 2:00pm: Devin Pleuler, Toronto FC
"Player Masks: Encoding Soccer Decision-Making Tendencies"

2:00pm – 2:30pm: Jackson Weaver, Harvard University
"The Statistics of Spin in Soccer"

October 22, 2021 - Panel Discussion

1:00pm – 2:30pm *"Sports Analytics Courses for the Next Generation of Analysts and Researchers"*

Moderator: Zachary Binney – Oxford College, Emory University

Panelist: Mark Broadie – Columbia Business School
Panelist: John Draper – The Ohio State University
Panelist: Konstantinos Pelechrinis – University of Pittsburgh
Panelist: Felesia Stukes – Johnson C. Smith University
Panelist: Tim Swartz – Simon Fraser University

Oral Presentation Abstracts

QUANTIFYING CONGESTION IN AUSTRALIAN RULES FOOTBALL

Alexander, Jeremy† (1); Robertson, Sam (1); Bedin, Timothy (2); Jackson, Karl (2)

(1) Institute for Health and Sport (IHES), Victoria University; (2) Champion Data Pty Ltd, Melbourne, Australia

† E-mail: *jeremyalexander60@gmail.com*

Contemporary Australian Rules Football (AF) has experienced a gradual decline in scoring, with defensive strategies and increased player congestion limiting the efficacy of opposition team performance. Preliminary investigations revealed that congestion has increased significantly over the last 15 years, which has generated disapproval amongst fans and AF lawmakers.

Nonetheless, a variable or metric that continuously provides a description of congestion remains absent. Counting the number of players in a given radius from the play, as is current practice, is inadequate to determine the comparative degree of congestion a player if confronted with when disposing of the ball and also infeasible. Using player tracking data, this study developed a measure of continuous congestion in Australian football using density-based clustering (OPTICS). Match events were manually labelled to three levels (inside congestion, congestion nearby, outside congestion) based on current practice to provide a ground truth from which a supervised machine learning model could predict the same classification from a range of spatiotemporal features. The model showed higher results for inside congestion, precision 0.88 and recall 0.90, and outside congestion, precision 0.98 and recall 0.88, compared to nearby congestion, precision 0.76 and recall 0.86. Overall model accuracy was 88.2% and the AUC was 0.85.

This information can be used to understand how congestion develops with the introduction of rule changes and determine the influence of contextual variables, such as field position, time of the match, and event outcomes.

PREDICTIVE VALUE OF OFF-TARGET SHOTS IN SOCCER

Baron, Ethan† (1); Chan, Timothy CY (1); Pleuler, Devin (2); Sandholtz, Nathan (3)

(1) University of Toronto, Toronto, ON, Canada; (2) Toronto FC, Toronto, ON, Canada; (3) Brigham Young University, Provo, UT, USA

† E-mail: *eth.baron@mail.utoronto.ca*

Measuring shooting skill in soccer is a challenging analytics problem for several reasons. Firstly, shots are rare events, so sample sizes are limited. Secondly, the goal-scoring probability of a shot depends on many contextual variables independent of player skill. The availability of tracking data enables analysts to better account for these variables via pre-shot expected goal models. Likewise, a shot's trajectory can be translated into a goal-scoring probability via a post-shot expected goal model. Comparing the values from these two models for a particular shot, sheds light on the finishing ability of the shot-taker.

In post-shot expected goals models, off-target shots are universally zero-valued since they have no probability of scoring. Despite this fact, we posit that a non-negligible shooting skill signal is contained in off-target shots. For example, all else being equal, a player's shot that narrowly misses the upper corner of the frame as opposed to another player's shot that misses wildly tells us something about their respective shooting skill.

In this project, we aim to extract the signal contained in off-target shots, thus providing a more accurate valuation of player shooting skill. We first develop a data-generating process for shots. Then, we attach non-zero value to observed off-target shots by connecting shots to their underlying data-generating process, simulating many shots from this process, and aggregating their post-shot expected goals values. We propose a new metric for player shooting skill which incorporates this approach and compare it against existing metrics.

ESTIMATING THE CHANGE IN SOCCER... HOME ADVANTAGE DURING THE COVID-19 PANDEMIC

Benz, Luke[†] (1); Lopez, Michael (2,3)

(1) Harvard University Department of Biostatistics, Boston, MA, USA; (2) National Football League, New York, NY, USA; (3) Skidmore College, Saratoga Springs, NY, USA

[†] E-mail: *lukesbenz@gmail.com*

In wake of the Covid-19 pandemic, 2019-2020 soccer seasons across the world were postponed and eventually made up during the summer months of 2020. Researchers from a variety of disciplines jumped at the opportunity to compare the rescheduled games, played in front of empty stadia, to previous games, played in front of fans. To date, most of this post-Covid soccer research has used linear regression models, or versions thereof, to estimate potential changes to the home advantage. However, we argue that leveraging the Poisson distribution would be more appropriate, and use simulations to show that bivariate Poisson regression reduces absolute bias when estimating the home advantage benefit in a single season of soccer games, relative to linear regression, by almost 85 percent. Next, with data from 17 professional soccer leagues, we extend bivariate Poisson models estimate the change in home advantage due to games being played without fans. In contrast to current research that suggests a drop in the home advantage, our findings are mixed; in some leagues, evidence points to a decrease, while in others, the home advantage may have risen. Altogether, this suggests a more complex causal mechanism for the impact of fans on sporting events.

A MARKOV PROCESS APPROACH TO UNTANGLING THE RELATIONSHIP BETWEEN INTENTION VS. EXECUTION IN TENNIS

Chan, Timothy CY (1); Fearing, Douglas S (2); Fernandes, Craig† (1); Kovalchik, Stephanie (2)

(1) Department of Mechanical and Industrial Engineering, University of Toronto, Toronto,Ontario, Canada; (2) Zelus Analytics, Austin, Texas, USA

† E-mail: *craig.fernandes@mail.utoronto.ca*

Value functions are used in sports applications to determine the optimal action players should employ. However, most literature implicitly assumes that the player can perform the prescribed action with known and fixed probability of success. The effect of varying this probability or, equivalently, "execution error" in implementing an action (e.g., hitting a tennis ball to a specific location on the court) on the design of optimal strategies, has received limited attention. In this paper, we develop a novel modeling framework based on Markov reward processes and Markov decision processes to investigate how execution error impacts a player's value function and strategy in tennis. We power our models with hundreds of millions of simulated tennis shots with 3D ball and 2D player tracking data. We find that optimal shot selection strategies in tennis become more conservative as execution error grows, and that having perfect execution with the empirical shot selection strategy is roughly equivalent to choosing one or two optimal shots with average execution error. We find that execution error on backhand shots is more costly than on forehand shots, and that optimal shot selection on a serve return is more valuable than on any other shot, over all values of execution error.

ATHLETE RATING IN MULTI-COMPETITOR GAMES WITH SCORED OUTCOMES VIA MONOTONE TRANSFORMATIONS

Che, Jonathan†; Glickman, Mark

Harvard University, Cambridge, MA, USA

† E-mail: *jche@g.harvard.edu*

Sports organizations often want to infer athletes' time-varying abilities from game outcomes. Existing methods for doing so typically rely on rank outcomes, such as win/loss in head-to-head games or the athlete's placement in multi-competitor games. In many games, however, athletes earn scores or times, which provide additional information about their performances. To rate athletes using game score data, we propose a Bayesian dynamic linear model that allows for transformations of the score outcomes. Our model learns nonlinear monotone transformations of the outcomes to account for non-normality in athlete performances and can be easily fit using standard regression and

optimization routines. Using simulated data, we show that our model outperforms conventional methods that only use rank outcome data. We demonstrate our method on several Olympic sports, including biathlon, rugby, and diving.

QUANTIFYING LEAGUE-INDEPENDENT SCORING ABILITY IN SOCCER

Daly-Grafstein, Daniel[†]

University of British Columbia, Vancouver, BC, Canada

[†] E-mail: *daly-grafstein@stat.ubc.ca*

Soccer is a worldwide sport where teams often trade-for or recruit players from different leagues. Evaluating these players objectively is difficult because it is hard to predict how a player's skill will translate from one league to another. Using goal-scoring data from the past 20 years of play in the Big-5 European Leagues, we aim to estimate players' true goal-scoring skill independent of the competitiveness of the league where they play in order to more accurately predict how players will perform when they move between leagues. We construct a hierarchical state-space model where we treat individual scoring ability as a latent variable. We utilize league, team, and position level random effects to share information between players as well as leverage autocorrelation in player skill and league competitiveness to improve player scoring estimation. Additionally, we include nonparametric age functions to account for changes in player scoring ability over time. Overall, we find incorporating time-varying league, team and player random effects allows us to better predict how players will perform when moving between leagues. We conclude by examining league-specific covariates to compare the relative scoring difficulty between leagues over time.

CAUSAL EFFECT OF PLAYING TIME IN THE NBA

Fischer, Sean[†]

Annenberg School for Communication, University of Pennsylvania

[†] E-mail: *sf585978@gmail.com*

Recent shifts in professional basketball have led teams to place more urgency in drafting as well as possible. Draft picks must play out their initial years under team-friendly contracts that provide teams with increased salary cap flexibility. Yet, while this urgency has led to widespread discussion and research of how to improve teams' draft decisions, little attention has been given to identifying what teams can do to maximize the performance and potential of their draft picks once they are added to their roster. However, theories of learning and ecological psychology suggest that giving

young players as much playing time as possible should lead to concrete improvements in their development and future performance. In this study, I test this causal theory by evaluating the relationship between the minutes a player receives in their first two seasons in the NBA and their fourth-year performance using a novel method of propensity score weighting that enables weighting for continuous treatment variables. I find that players who receive more minutes in their first two seasons have better fourth seasons and make larger jumps from their first two seasons to their fourth season, controlling for a broad set of potential confounders. These results have important implications for teams as they develop organizational strategies for the short- and medium-term.

PACE AND POWER: REMOVING UNCONSCIOUS BIAS FROM SOCCER BROADCASTS

Gregory, Sam† (1, 2); Pleuler, Devin (3), Daly-Grafstein, Daniel (4), Liu, Yang (5), Marchwica, Paul (5)

(1) Victoria University, Melbourne, AUS; (2) Inter Miami CF, Miami, FL; (3) Toronto FC, Toronto, CAN; (4) University of British Columbia, Vancouver, CAN (3), Sportlogiq, Montreal, CAN (5)

† E-mail: *gregoryd.sam@gmail.com*

This research challenges common stereotypes in professional soccer based on race and gender by employing computer vision based match recreations. Using player location and body pose information we recreate match broadcasts with player skeletons so that the viewer can analyze a match without any visual information about the players' race or gender. We ask a series of questions about a video segment involving teams of different racial backgrounds and video segments from a men's game and a women's game. We use the broadcast recreations as a control group and show a second group the original broadcasts. We find that when viewers are able to identify player race they are much more likely to attribute athletic or physical characteristics to black players. We also find evidence that when viewers can tell they are watching a women's game they may identify it as of a lower quality than when they are watching the recreation.

WHICH TEAMS WOULD HAVE WON THE 2020 NCAA MEN'S AND WOMEN'S BASKETBALL TOURNAMENTS?

Johnstone, Chancellor† (1); Nettleton, Dan (2)

(1) Air Force Institute of Technology, Wright-Patterson Air Force Base, OH; (2) Iowa State University, Ames, IA

† E-mail: *chancellor.johnstone@gmail.com*

The COVID-19 pandemic was responsible for the cancellation of both the men's and women's 2020 National Collegiate Athletic Association (NCAA) Division 1 basketball tournaments. With the end goal of identifying which college basketball teams might have won the tournament(s), we introduce a closed-form calculation for overall tournament win probabilities. Starting from the point at which the Division 1 tournaments, along with any unfinished conference tournaments, were cancelled, we deliver closed-form tournament win probabilities for the top men and women's college basketball teams in 2020, removing the inherent Monte Carlo error associated with simulation. We also introduce a new method for generating win probabilities associated with individual game outcomes through conformal prediction, named conformal win probability. We generalize conformal win probability to conformal event probability and provide multiple theoretical results to show its validity. We then compare conformal win probabilities to those generated through linear and logistic regression on five years of historical college basketball data. Conformal win probabilities are shown to be better calibrated than other methods resulting in more accurate win probability estimates, while making fewer distributional assumptions.

A COPULA-BASED HIDDEN MARKOV MODEL FOR CLASSIFICATION OF TACTICS IN FOOTBALL

Oetting, Marius† (1); Karlis, Dimitris (2)

(1) Bielefeld University, Germany; (2) Athens University of Economics, Greece

† E-mail: *marius.oetting@uni-bielefeld.de*

Driven by recent advances in technology, tracking data in soccer are available to nearly every professional team. However, for tasks such as analyzing future opponents, most teams do not use tracking data and instead still rely on video analysis, for example with regard to how goals were scored or a team's general style of play. Since video analysis is rather time-consuming, this talk provides a data-driven approach for automated classification of tactics in soccer using the publicly accessible tracking data provided by Metrica Sports at GitHub.

For our analysis, we consider hidden Markov models (HMMs) to analyze high-frequency tracking data, where the underlying states serve for a team's tactic. As space control in soccer has been considered a major driver of success, we focus on the effective playing space (EPS), which is the convex hull created by the players. In our HMMs, we jointly model the EPS of both teams using a copula to ensure that interactions between teams are captured.

Our fitted model can be used to obtain the decoded most likely sequence of a team's underlying tactics. These decoded states enable to investigate a team's style of play at each time point, which can be linked to events in a match such as scored goals. Furthermore, the decoded states are used to evaluate a team's attacking phases or to investigate their pressing play. These further analyses can be beneficial for team managers but are also of interest to soccer fans.

PLAYER MASKS: ENCODING SOCCER DECISION MAKING TENDENCIES

Devin Pleuler†

Toronto FC, Toronto, ON, Canada

† E-mail: *dpleuler@torontofc.ca*

We extend modern deep convolutional neural network architectures designed for analysis of spatio-temporal performance data in soccer by attaching a parallel network that introduces limited player-level information into the model training process which encodes individual decision making tendencies into latent space embeddings. These embeddings are reshaped into surfaces, or "Player Masks", which are used as additional channels for the prediction architecture. These embeddings and masks can be compared through traditional techniques to develop measures of player similarity, but a more interpretable measure of similarity is earned when comparing the spatial similarity of predicted player passing distribution. By applying this approach with widely available event data, it provides a new method for identifying players in the recruitment theater. This approach can also serve as a foundation for building player-parameterized agent-based models of game simulation for performance forecasting purposes.

EXTENDED MODEL FOR EXPECTED THREAT IN SOCCER

Poropudas, Jirka† (1); Inkilä, Ville-Pekka (2)

(1) SportIQ, Helsinki, Finland; (2) Football Association of Finland, Helsinki, Finland

† E-mail: *jirka.poropudas@gmail.com*

In soccer, expected goals (xG) models are used to estimate the probability of scoring when a shot is taken from a given location. Expected threat (xT) models combine xG models with soccer dynamics by introducing a stochastic model for the next ball event (move the ball or shoot). In xT models, moving the ball to another location by dribbling or passing is modeled as a Markov Chain. If a shot is taken, the scoring probability is computed as in xG models. Once the probabilities for both the Markov and xG model have been estimated from event data, the xT model can be used to compute an xT value (i.e., the probability of the team scoring during the next few events) for each field location.

A crucial shortcoming of existing xT models is that they exclude the possibility of turnovers, whereby the analysis is focused entirely on scoring goals. Moreover, these models are limited to actions taking place during only a few subsequent events and the part of the field from which a goal can be expected to be scored during those events.

We present an improved xT model which accommodates turnovers and subsequent ball movement as well as the negative xT posed by the opponent coming into the possession of the ball. Our model also makes it possible to consider all the events up to the next shot (towards either of the goals), thereby covering the entire field. We illustrate the model using real data from Stats Perform.

DEVELOPING MULTIVARIATE EXTREME VALUE METHODS TO EXPLORE NFL PROSPECT DATA

Russell, Brook T.† (1); Newman, Sydney (1); Hogan, Paul (2)

(1) Clemson University School of Mathematical and Statistical Sciences, Clemson, SC, USA; (2) Clemson University Football Strength and Conditioning, Clemson, SC, USA

† E-mail: *brookr@clemson.edu*

Recently, the use of extreme value methods has increased in the field of sports analytics, allowing analysts to model the far upper tail of response distributions in a wide variety of sports. However, much of this work has relied on univariate extreme value theory. In this talk, we discuss two novel approaches for analyzing NFL prospect data that have been developed within the framework provided by multivariate extreme value theory.

In order to better understand the current battery of NFL Combine tests, we first develop a method that aims to characterize their multivariate dependence structure for both typical and elite-level NFL prospects. Through analysis of two pairwise dependence matrices, one characterizing dependence in the center of the distribution and the other characterizing dependence in the far upper tail of the distribution, we find that several events have relatively high levels of association and conclude that fewer Combine events may be sufficient. Understanding NFL player success based on collegiate and NFL Combine metrics is highly challenging.

We next develop a method that attempts to identify an optimal function of NFL prospect predictor variables (for skill position players) that yields the highest possible degree of asymptotic dependence with a response variable. Informally, two variables are asymptotically dependent if the probability that both are at extreme levels simultaneously is non-zero. We consider responses that measure NFL success, and find that this approach faces challenges that are similar to those in previous approaches, but also has the potential to offer different insights.

AN INVERSE OPTIMIZATION ANALYSIS OF THE FOURTH DOWN DECISION IN FOOTBALL

Sandholtz, Nathan† (1); Wu, Lucas (2); Puterman, Martin (3); Chan, Timothy CY (4)

(1) Brigham Young University, Provo, UT, USA; (2) Simon Fraser University, Burnaby, BC, Canada; (3) University of British Columbia, Vancouver, BC, Canada; (4) University of Toronto, Toronto, ON, Canada

† E-mail: *nsandholtz@stat.byu.edu*

The fourth down decision in football has been primarily studied as an optimization problem: using win probability as the objective function, analysts estimate optimal decisions for every fourth down situation. Prescriptions from these models have been publicly available for decades, informed by increasingly sophisticated win probability models in recent years. Despite this availability, NFL coaches' observed fourth down decisions have remained distant from the analysts' recommendations. Inverse optimization provides a mathematical framework to make sense of the gap between coaches' decisions and analysts' prescriptions. Leveraging this paradigm, we assume that the coaches' observed decisions are optimal but that the risk preferences governing their decisions are unknown. Our goal is to infer these latent risk preferences such that the resulting optimization model yields their observed decisions as optimal (or minimally suboptimal).

To this end, we model a football game as a Markov decision process where the observed fourth down policy (estimated from NFL play-by-play data) is assumed to be optimal. In order to render the observed policy as optimal, we parameterize the optimality criterion of the MDP rather than assuming risk-neutrality (i.e. the expectation), which is typically the default. Using the quantile function as a flexible risk measure, we determine the quantile over future rewards which renders the coaches' observed decisions as minimally sub-optimal. We find that coaches universally exhibit conservative risk preferences, but that they have different risk preferences depending on various features of the fourth down situation (field position, score differential, etc.).

THE STATISTICS OF SPIN IN SOCCER

Weaver, Jackson† (1); Shaw, Laurie (2)

(1) Harvard University, Cambridge, MA, USA; (2) City Football Group, Manchester, UK

† E-mail: *Jacksonweaver@college.harvard.edu*

Data science and sports have long shared a connection, and with the introduction of tracking data, new branches of research are now possible. Soccer being a continuous, fluid sport means that tracking data is vital to analyzing the game. Previous research has focused on the movements and

trajectories of players, but using physics we can model and examine the trajectories of the soccer ball. Combining the high frequencies of observations with the physical modeling of ball dynamics, we can start to measure and analyze how players use spin. This paper develops this model and creates a fitting procedure to infer key physical parameters from large batches of corner kicks, such as spin rate and initial velocity, to find exactly how the ball was kicked. Corner kicks are chosen as they represent extended, clean trajectories. For the first time, we demonstrate that ball spin rates can be inferred from the spatiotemporal data, even taking into account the confounding effects of the wind. Once fit, the general distribution for corners across many games were analyzed to discover general trends, giving a new look at how players use spin in soccer.