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.