MSAM 2019 – @MWSprtAnalytics
ABSTRACTS – ORAL PRESENTATIONS

Program Purposefulness: Using Analytics in High School Basketball

Kyle Allen, Pine City High School (MN)

Abstract: Kyle Allen has been head boys basketball coach at Pine City, Minnesota for eight years, during which time the Dragons have received statewide and national attention from the likes of the Wall Street Journal, Basketball Talk Pro, KFAN radio, FOX 9, among others. The reason: the program's philosophy and use of data desegregation.

According to an article in MSHSL's John's Journal, "mathematics is a big part of (Allen's) job as head coach of the Dragons boys basketball team. That's because everything is measured and charted: not only the typical things like shooting percentages and rebounds, but also talking. Yes, the Dragon's keep track of talking. And that's just the start of what makes his basketball team unlike any other. The most visible example: they rarely shoot two-point shots other than layups, and focus on firing from outside the three-point line. It's all based on math."

NBA Lineup Analysis on Clustered Player Tendencies

Jonathan Bosch, Syracuse University
Samuel Kalman, Purdue University

Abstract: Recently, basketball has been considered more of a "position-less" game. The majority of NBA players have skillsets and tendencies that cannot be defined by their position on paper. With ten seasons (2009-2018) of NBA player stats that account for skill, opportunity, and tendency, we were able to cluster players together into nine groups that categorized the player's "playing type" by using model-based clustering in R (mclust package). After analyzing the distribution of statistics within each cluster, we were able to name each cluster based on what that cluster tended to do on the basketball court. Realizing that many players are versatile and bring a variety of skills to the game, we used the cluster probabilities (probability that a certain player is in each respective cluster) to divide each player among the nine clusters. After gathering the past ten seasons of NBA five-man lineup data, we transferred our results from the clustered playing types to create "soft lineups". These "soft lineups" contained combinations of the nine player types, based on the cumulative cluster probabilities from the five players in that lineup. Using the soft lineups, we bootstrapped 100 Random Forest models and created 100 predictions of the net rating for each possible "soft lineup" combination (3.1 million lineups). This gave us the ability to create a prediction interval for each possible "soft lineup" combination, and a deep insight into how well a given lineup will perform. Our work offers a more specific way for people to consider player types and positions in the NBA. It also provides insight into what combination of player types yield the most effective basketball performance. This can

be beneficial to NBA front offices with acquiring talent, as well as coaches making in game lineup decisions.

Estimating (Four) Factor Values in the NBA: A Seemingly Unrelated Regression Analysis

Jonathan Bosch and Dax Speakman, Syracuse University

Abstract: 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.

Does General Admission Alcohol Availability Affect College Football Attendance and Revenues?

Stacey Brook, DePaul University

Abstract: The percentage of universities selling alcohol to legally-aged general admission spectators at Football Bowl Subdivision stadiums has increased over 300% from 2007 to 2017. Previous research examines the effect of alcohol availability using either season or game attendance but not revenues. The literature is extended by analyzing football program revenues with data from the majority of public athletic programs using the NCAA Membership Financial Reporting System covering the 2004 to 2016 seasons. Using instrumental variables techniques, general admission alcohol availability has no statistically significant effect on aggregate regular season home attendance. Using quantile regression, FBS stadium general admission alcohol availability has no statistically significant impact on ticket sales, contribution revenue or generated revenues; increases concession revenue for programs at the 75th and 50th quantiles of the conditional distribution; and has a negative impact on total operating revenues at all but the lowest quantile of the conditional distribution.

Estimating and Predicting Team Quality in MLB: Comparing Pythagorean Expected Wins and Alternative Models from Contest Theory

Justin Ehrlich and Shane Sanders, Syracuse University
Shankar Ghimire, Western Illinois University

Abstract: The Pythagorean Expected Wins Model was developed by Bill James (1980) to estimate a baseball team's expected wins (as distinct from the team's actual wins) over the course of a season. As such, the model can be used to assess how lucky or unfortunate a team was over the course of a season (actual wins – expected wins). From a managerial perspective, such information is valuable in that it is important to understand how reproducible a given result may be in the next time period. In contest-theoretic (game-theoretic) parlance, James' original model represents a (restricted) Tullock contest success function (CSF). We transform, estimate, and compare James' original model and two alternative models from contest theory—the serial and difference-form CSFs—using MLB team win data (2003-2015). The serial CSF estimator dramatically improves wins estimation (reduces root mean squared error) compared to James' original model, an optimized version of James' model, or an optimized difference-form model. We conclude that the serial CSF model of wins estimation substantially improves estimates of team quality, on average. The work provides a real-world test of alternative contest forms.

NFL Revenue Sharing and the Median Voter Theorem

Justin Ehrlich and Shane Sanders, Syracuse University
Shankar Ghimire, Western Illinois University

Abstract: "We're 26 Republicans who vote like socialists (on the issue of revenue sharing)." — late Baltimore Ravens owner Art Modell

Revenue sharing is ubiquitous among North American professional sports leagues. Under pool revenue sharing, above-average revenue teams of a league effectively transfer revenues to below-average revenue teams. Herein, we consider the National Football League (NFL), a League that pools and equally shares national revenues among member teams. The NFL has shared approximately 61 percent of all revenues during the 2000s. We examine—from a median voter theorem perspective—why revenue sharing has been (historically) chosen—via a series of majority-rule votes—by NFL owners. Taking advantage of financial disclosure requirements for the Green Bay Packers, as well as NFL team revenue data from Forbes.com, we disaggregate NFL team-season revenue values into local (unshared) revenue and national (pooled and equally-shared) revenue. We then measure the non-parametric skewness of the NFL (unpooled) team revenue distribution and find consistently significant, positive skewness for each sampled season (i.e., from 2001 through 2016). Mean team revenue is consistently greater than median team revenue such that the median (un-pooled revenue) team benefits from a mean-reverting pooled revenue sharing scheme. Given this result, the median voter theorem predicts that the (pivotal) median voting owner will consistently vote for a revenue sharing proposal (against the absence of revenue

sharing). Distribution of revenues—namely, the existence of outlying large market NFL teams—appears to consistently explain the historical popularity of NFL revenue sharing. A voting game is considered and the equilibrium is solved to illustrate the median voter theorem.

Do Offensive Players Generate More Revenue Than Defensive Players in Major League Baseball?

Justin Ehrlich, Syracuse University
Joel Potter, University of North Georgia

Abstract: Previous research has found evidence that fans prefer offense and that teams reward offensive production more than defensive production. However, Ehrlich, Potter, and Sanders (2019) demonstrate that game attendees do not value one type of play category over another. This talk extends Ehrlich et al (2019) by shifting the focus to revenue, which allows us to analyze the entire fan base of each team instead of only those that attended the games, e.g., broadcasting deals, merchandise sales, etc. Data and analysis from the 2010-18 baseball seasons will be discussed and explored. By decomposing Wins Above Replacement (WAR) into offensive (oWAR), defensive (dWAR), and pitching (pWAR), this talk demonstrates that dollars being spent are not sensitive to one category over another. Since revenue does not reward one WAR category over another, but salaries do, then the MLB labor market is likely inefficient. These findings imply that a revenue maximizing team could take advantage of this inefficient labor market by paying for defensive production at the same level as offensive production.

Elam Ending: Past, Present, and Future

Nick Elam, Ball State University

Abstract: What began as an independent research study in 2007 has now come to life at TBT: The Basketball Tournament, and continues to be adopted by more basketball leagues and events. This presentation revisits the origins and evolution of the hybrid duration format (as the concept was called before being renamed the Elam Ending by TBT), which calls for the final portion of each game to be played without a game clock as a way to preserve a more natural style of play through the end of every game. The presentation addresses research methods used to investigate the necessity and soundness of the concept, approaches used over the years to promote and spark discussion about the concept, the effectiveness of the Elam Ending in its current form and setting, and explores possibilities for the future of the Elam Ending.

Predicting Players' Performance in the National Football League

Ashlyn Hartman, Millikin University

Abstract: The National Football League (NFL) teams each year make a large investment in players they draft. The factors that determine the players' performance were studied by using data such as college statistics, NFL combine data, and NFL draft results. Performance in the NFL is defined as lifetime earnings in this study. We created a random forest to determine the most important attributes in predicting the

players' performance. The study presents the results of my investigation of finding if NFL combine has a relationship to the performance of the player in the NFL.

Nonparametric Prediction Intervals in Sports

Chance Johnstone, Iowa State University

Abstract: We deliver a novel nonparametric prediction interval methodology based on the ranking method introduced in Harville 1977 (otherwise recognized as the Massey Method). Typical prediction intervals are based on estimates of the mean and variance of an observation and include assumptions on the normality, zero expectation, and constant variance of error terms. We make only the latter two assumptions. With this, we build an increasing collection of modified residuals coming from weekly, out-of-sample predictions that approximates the true predictive error distribution for outcomes of a competition. The quantiles of this collection, combined with weekly predictions, gives a (1- α) prediction interval for the margin-of-victory of a specific game. We apply this methodology to National Football League (NFL), Women's National Basketball Association (WNBA), and English Premier League (EPL) data. We also provide simulations to better explore the asymptotic properties of this prediction interval methodology across an infinite season, and extend it to determine both win probabilities and spread probabilities.

More Than Just Butts in Seats

Madison Koch, Minnesota Vikings (NFL)

Abstract: It is no surprise a team plays better in front of a full stadium of fans, but, when this is your team's normal home game setup, how do you give the players and coaches an even bigger home field advantage? For the Minnesota Vikings, it is about using data analytics to know fans as more than just the butts in the seats – stadium, couch, or barstool. Come explore how the Vikings get the most home field advantage possible by leveraging data to create unique 365-day fan experiences.

Development of Predictive Process for Season Outcomes in a Local Collegiate Baseball League

Jessica Kraker, University of Wisconsin - Eau Claire

Abstract: We gathered and analyzed statistics and cumulative records of the collegiate summer Northwoods Baseball League to explore whether and how various new metrics (originally developed for Major League Baseball) might translate to a new context. Based on information about the team dynamics over the season, we discuss amendments to the original model (simple Bradley-Terry predictions), as well as options for integrating individual player statistics. Functions are developed in R to integrate historically available player statistics with new measures, in order to better assess player utility.

Analyzing Equity in High School Cross Country Competition

Matt Kretchmar, Denison University

Abstract: We study how the enrollment size of a high school affects its ability to compete fairly in cross country distance running. We build a statistical model of high school runners and use Monte Carlo simulations to explore the potential advantages that larger schools have over smaller programs. Attendees will develop an understanding and appreciation of how enrollment size impacts competitive equity in schools; understand how Monte Carlo simulations work and can be applied to sports analytics; and experience an example of using data collection to build a statistical model used in simulation.

Performance is Not a Strategy: Winning off the Field

David Longstreet, FanThreeSixty

Abstract: Technology and digital media have fundamentally transformed the way fans watch and attend sports events. From streaming content, team apps and digital tickets to eCommerce and in-venue Wi-Fi, fans today are more digitally connected than ever before. And while this has led to a deeper connection between fans and their favorite teams, it has also created an explosion of valuable fan information. This data, ranging from demographic and behavioral to social data, gives teams a more progressive and accurate way to understand their fans.

Delivering a personalized fan experience requires a clear data strategy and taking the necessary foundational steps.

Basketball Data Analytics Battle 2.0

Rick Spellerberg, Iowa Center for Interdisciplinary Training
Vicki Hamdorf, North Cedar High School (IA)

Abstract: The Basketball Data Analytics Battle is a grades 6-12 competition that introduces students and teachers to the world of data analytics. Using qualifying team data for the 2013-2019 NCAA basketball tournaments, students create an algorithm that predicts the ten teams that have the best chance of winning the 2020 tournament. This talk will introduce the audience to the competition and provide comparative outcome information from the 2018 and 2019 competitions. Included will be observations from participating students and teachers that came from survey results.