PROGRAM

September 21, 2013

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

Symposium Co-Chairs:

Mark E. Glickman, Department of Health Policy and Management, Boston University School of Public Health

Scott R. Evans, Department of Biostatistics, Harvard University School of Public Health

Director of Speaker Recruitment and Sponsorship:

Jason W. Rosenfeld, Charlotte Bobcats

Sponsors:

- Boston Chapter of the American Statistical Association (`http://www.amstat.org/chapters/boston/`)
- Section on Statistics in Sports of the American Statistical Association (`http://www.amstat.org/sections/sis/`)
- Harvard University Department of Statistics (`http://www.stat.harvard.edu/`)
- Sports Data Hub (`http://www.sportsdatahub.com/`)
- ESPN Stats & Info (`http://espn.go.com/blog/statsinfo/`)
- RStudio (`http://www.rstudio.com/`)
- Revolution Analytics (`http://www.revolutionanalytics.com/`)

Wifi access:
Guest wifi access will be available for conference participants in the Science Center. Open a browser on your laptop or mobile device, and click on "Guest Access" on the welcome screen. You will be asked for your name, phone number, and e-mail address, and then a firewall check will be performed. After completing this process, you will have wifi access.

Acknowledgments: We wish to thank everyone who helped to make NESSIS possible. We would also like to thank Kevin Rader, Sowmya Rao, Tom Lane, Mingfei Li, Eugenie Coakley, Dick Evans, Phil Everson, Robin Lock, Mike Zarren, Iram Farooq, Betsey Cogswell, Dale Rinkel, Maureen Stanton, Jeffrey Myers, and Huichao Chen for their parts in helping with the symposium.

2013 New England Symposium on Statistics in Sports

September 21, 2013

Welcome Address

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

Morning Featured Session: Lecture Hall C

9:30am – 10:15am: Richard Smith, SAMSI and University of North Carolina
"Completing the Results of the 2013 Boston Marathon"

10:15am – 11:00am: Jan Vecer, Frankfurt School of Finance and Management
"Crossing in Soccer has a Strong Negative Impact on Scoring: Evidence from the English Premier League and the German Bundesliga"

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	Lecture Hall D
Dan Cervone/Alex D'Amour, Harvard University *"State of Transition: Estimating Real-Time Expected Possession Value in the NBA with a Spatiotemporal Transition Model and Player Tracking Data"*	Robert H. Carver, Stonehill College *"Wherever the Dickey Winds Up, or is the Wind Up? How Weather Affects the Knuckleball"*
John Ezekowitz, Harvard University *"The Hot Hand: A New Approach to an Old 'Fallacy' "*	Stephanie Kovalchik, RAND Corporation *"Trends in Singles Play Intensity on the ATP Tour"*
Andrew Miller, Harvard University *"Quantifying Offensive Player Types in the NBA with Non-Negative Matrix Factorization"*	Dennis Lock, Iowa State University *"Using Random Forests to Estimate Win Probability before Each Play of an NFL Football Game"*
Philip Everson, Swarthmore College *"Mid-Game Predictions for NBA Basketball"*	Michael A. Pane, Carnegie Mellon University *"Trouble with the Curve: Identifying Clusters of MLB Pitchers using Improved Pitch Classification Techniques"*

Lunch break: Foyer area

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

Afternoon Featured Session: Lecture Hall C

2:00pm – 2:45pm: Luke Bornn, Harvard University, and
Kirk Goldsberry, Harvard University
"XY, A New Look at the NBA"

2:45pm – 3:30pm: Jim Albert, Bowling Green State University
"Assessing Streakiness in Home Run Hitting

Poster Session: Foyer area

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

Panel Discussion: Lecture Hall C

5:00pm – 6:30pm: *"Baseball Analytics for Strategy, Scouting, and Decision-Making"*

Moderator: Andy Andres – Head Coach and Lead Instructor, MIT Science of Baseball Program; and Fenway Park Datacaster/Stringer for mlb.com and MLBAM (Gameday)

Panelist: Vince Gennaro – President of SABR; consultant to MLB teams, 2006-2013

Panelist: Ben Baumer – Visiting Assistant Professor, Smith College, statistical analyst for Baseball Operations with the NY Mets, 2004-2012

Panelist: Eric M. Van – Sabermetric Baseball Operations consultant for the Red Sox, 2005-2009

Post-NESSIS Get-Together

6:30pm – 9:00pm: *Grafton Street Pub & Grill, Harvard Square*
1230 Massachusetts Ave., Cambridge, MA 02138
`http://graftonstreetcambridge.com/`

Oral Presentation Abstracts

ASSESSING STREAKINESS IN HOME RUN HITTING

Jim Albert[†]

Bowling Green State University

[†] E-mail: *albert@bgsu.edu*

The media is fascinated with "ofers" – spacings between successes for baseball hitters. We collect spacings between home runs for all hitters in the last 50 baseball seasons. Consistent and streaky models are defined on the basis of the underlying geometric probabilities and a Bayes factor statistic is developed to compare the two models. By looking at the ensemble of test statistics over all players, we see if the patterns of streakiness differ from what one would predict from a consistent hitting model. We identify home run hitters who have interesting patterns of streaky performance.

XY, A NEW LOOK AT THE NBA

Luke Bornn[†1], Dan Cervone[1], Alex D'Amour[1], Alex Franks[1], Kirk Goldsberry[2], Ryan Grossman[1], Andrew Miller[1]

Harvard University[1], ESPN[2]

[†] E-mail: *bornn@stat.harvard.edu*

In this talk, I will explore the state of the art in the analysis and modeling of player tracking data in the NBA. In the past, player tracking data has been used primarily for visualization, such as understanding the spatial distribution of a player's shooting characteristics, or to extract summary statistics, such as the distance traveled by a player in a given game. In this talk, I will present how the XY Hoops research group at Harvard is using advanced statistics and machine learning tools to answer previously unanswerable questions about the NBA. Examples include "How should teams configure their defensive matchups to minimize a player's effectiveness?", "Who are the best decision makers in the NBA?", and "Who was responsible for the most points against in the NBA last season?"

WHEREVER DICKEY WINDS UP, OR IS THE WIND UP?: HOW WEATHER AFFECTS THE KNUCKLEBALL

Robert H. Carver†

Stonehill College, Easton MA

† E-mail: *rcarver@stonehill.edu*

When R.A. Dickey was with the NY Mets he published an autobiography entitled "Wherever I Wind Up: My Quest for Truth, Authenticity, and the Perfect Knuckleball." The knuckleball is one of the rarest pitches in the repertoire of major league pitchers. It has the potential to confound batters (as well as catchers and umpires) with its unpredictable movement, which results from the absence of rotation on the ball and the interplay of the air turbulence and pressure differentials on the stitches and smooth surface of the baseball. The pitch is notoriously difficult to control, but when effective its slow speed and wide arc leaves batters extremely frustrated. Using data from Pitch FX and the National Weather Service, this line of research has examined the impacts of wind, barometric pressure and humidity on the path of knucklers thrown over selected seasons by both Dickey and veteran Tim Wakefield of the Boston Red Sox. The availability of detailed data about pitch movement in live games enables this line of research.

STATE OF TRANSITION: ESTIMATING REAL-TIME EXPECTED POSSESSION VALUE IN THE NBA WITH A SPATIOTEMPORAL TRANSITION MODEL AND PLAYER TRACKING DATA

Dan Cervone†, and Alex D'Amour

Statistics Department, Harvard University, Cambridge, MA, USA

† E-mail: *dcervone@fas.harvard.edu*

Despite many recent innovations in basketball analytics, we have trouble quantifying the value of events that do not directly lead to points or turnovers. For example, how can we evaluate a point guard's decision to pass to the center instead of pulling up for an uncontested 3-pointer from the top of the arc? Ideally, one would compare these choices in terms of their expected possession value (EPV), or the number of points that the team could expect to score by the end of the possession given the player's decision. Until recently, data sets of the detail necessary to effectively estimate EPV have been unavailable.

The SportVU player tracking technology, which records the location of all ten players 25 times per second, provides us a massive spatiotemporal data set that makes estimation of

EPV possible. We consider a probabilistic model for the evolution of a possession given the exact spatial configuration of the players. The possession is represented as a sequence of states from a finite state space where transition probabilities depend on the spatial history of the players and ball. By estimating these transition probabilities, we can derive the real-time EPV's of a possession. Using this framework, we can draw conclusions at the level of the possession (At which moment did this become a high-value possession?), the player (In general, how good is this player's decision-making?), and the team (How does this offensive system generate points?).

MID-GAME PREDICTIONS FOR NBA BASKETBALL

Phil Everson[†1] and Jimmy Charite[2]

Swarthmore College, Swarthmore, PA, USA[1], *Columbia University, New York, NY, USA*[2]

[†] E-mail: *everson@swarthmore.edu*

The point totals for the home team and the road team in an NBA game follow a distribution that is approximately bivariate Normal. If the home team is favored by d points and the over-under is given as t points, then the mean points for the home and road teams are very nearly $(t+d)/2$ and $(t-d)/2$. For several complete NBA seasons we have point spread and over-under numbers for each game, as well as the time remaining and the point totals at the time of each score change in each game. Using these data we estimate how the mean and covariance matrix of the remaining points change throughout a game. This information allows us to make predictions about the final outcome at any time during a game.

THE HOT HAND: A NEW APPROACH TO AN OLD "FALLACY"

John Ezekowitz[†], Carolyn Stein, Andrew Bocskocsky

Harvard University, Cambridge, MA, USA

[†] E-mail: *john.ezekowitz@gmail.com*

The vast literature on the Hot Hand Fallacy in basketball rests on an assumption that shot selection is independent of player-perceived hot or coldness. In this paper, we challenge this assumption using a novel dataset of over 83,000 shots from the 2012-2013 National Basketball Association (NBA) season, combined with optical tracking data of both the players and the ball. We create a comprehensive model of shot difficulty using relevant initial shot conditions, and use it to show that players who have exceeded their expectation over recent shots shoot from significantly further away, face tighter defense, are more likely to take their team's next shot, and take more difficult shots. We then turn to the hot hand itself and present two

empirical strategies, including a novel technique to quantify shot trajectories, that find small yet significant hot hand effects. We argue that the hot hand appears once it is understood in a relative rather than an absolute sense. Our estimates of the hot hand effect range from 1.4 to 3 percentage points in increased likelihood of making a shot.

COMPLETING THE RESULTS OF THE 2013 BOSTON MARATHON

Dorit Hammerling[1], Matthew Cefalu[2], Jessi Cisewski[3], Francesca Dominici[2], Amy Grady[4], Giovanni Parmigiani[2,6], Charles Paulson[5], Richard Smith[†1,4]

Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina[1], Department of Biostatistics, Harvard School of Public Health[2], Department of Statistics, Carnegie Mellon University[3], Department of Statistics and Operations Research, University of North Carolina at Chapel Hill[4], Puffinware LLC[5], Dana Farber Cancer Institute, Boston[6]

[†] E-mail: *rls@email.unc.edu*

The 2013 Boston marathon was disrupted by two bombs placed near the finish line, that resulted in three deaths and several hundred injuries. In addition, nearly 6,000 runners were unable to finish the race. We were approached by the Boston Athletic Association (BAA), organizers of the Boston marathon, and asked to recommend a procedure for establishing projected finish times for the affected runners. With assistance from the BAA, we created a dataset consisting of all the runners in the 2013 race who reached the halfway point but failed to finish, supplemented by runners from the 2010 and 2011 Boston marathons. The data consist of "split times" from each of the 5 km sections of the course, as well as the final 2.2 km (40 km to the finish). The statistical objective is to predict the missing split times for the runners who failed to finish in 2013.

In this paper, we set this problem in the context of past research on the so-called matrix completion problem, examples of which include imputing missing data in DNA microarray experiments, and the Netflix prize problem. We propose five prediction methods and create a validation dataset to measure their performance by mean squared error and a number of other measures. The best method used local regression based on a k-nearest-neighbors algorithm (knn method), though several other methods produced results of similar quality. Finally we show how the results were used to create projected times for this year's runners and discuss some features of the resulting projections. We present the whole project as an example of reproducible research, in that we are able to make the full data and all the algorithms we have used publicly available, which may facilitate future research extending the methods or proposing completely different approaches.

TRENDS IN SINGLES PLAY INTENSITY ON THE ATP TOUR

Stephanie A. Kovalchik[†]

RAND Corporation, Santa Monica, CA, USA

[†] E-mail: *s.a.kovalchik@gmail.com*

The Association of Tennis Professionals' (ATP) Tour has undergone some dramatic changes since its founding in 1990. Throughout its twenty-three-year history, tournament schedule, prize money, and media coverage have each continued to expand, and the internationalization of the sport was officially recognized in 2009 with the establishment of the ATP World Tour. The changing structure of the Tour has coincided with equally dramatic changes to the game itself. Tennis is more physical than ever. Between the beginning of the Open Era and the present, the dominate playing strategy has gone from serve-and-volley to marathon battles at the baseline that require the highest level of athleticism and endurance from today's tennis champions. To better understand how the demands of professional tennis have changed during the history of the ATP Tour, I examine yearly trends in matches played and match duration among the ATP's top 100 singles players between 1991 and 2012. Using regression analyses and a database of all ATP matches played among a cohort of over 450 male tennis players, I investigate how the intensity of play–measured by match length and game duration–has changed over time and how these trends have varied by surface.

USING RANDOM FORESTS TO ESTIMATE WIN PROBABILITY BEFORE EACH PLAY OF AN NFL FOOTBALL GAME

Dennis Lock[†], Dan Nettleton

Iowa State University

[†] E-mail: *Dennis.f.lock@gmail.com*

Before any play of a National Football League (NFL) game, the probability that a given team will win depends on many in-game variables (such as time remaining, yards to go for a first down, field position and current score) as well as many pre-game variables (such as home team, team win-loss record and opponent win-loss record). We use a random forest method to combine in-game and pre-game variables to estimate Win Probability (WP) before any play of an NFL game. When a subset of NFL play-by-play data for the 12 seasons from 2001 to 2012 is used as a training dataset, our method provides WP estimates that accurately predict game outcomes, especially in the later stages of games. In addition to being intrinsically interesting in real time observers of an NFL football game, our WP estimates can provide useful evaluations of plays and play calls.

QUANTIFYING OFFENSIVE PLAYER TYPES IN THE NBA WITH NON-NEGATIVE MATRIX FACTORIZATION

Andrew Miller†, Luke Bornn, Ryan Adams

Harvard University, Cambridge, MA

† E-mail: *acm@seas.harvard.edu*

We analyze the underlying spatial structure that governs shot selection among NBA players. Using non-negative matrix factorization (NMF), an unsupervised dimensionality reduction algorithm, we uncover a low-dimensional structure that summarizes the shooting habits of NBA players.

Using SportVu data, we model each player's shot selection as a log-Gaussian Cox process on the court and then apply NMF to the inferred intensity surfaces. NMF encodes our assumption that the matrix of all players' intensity surfaces is low rank. The resulting basis surfaces clearly match common intuitions regarding offensive player types: some bases correspond to corner three point shooters, right handed post players, midrange jump shooters, etc. Furthermore, a weight vector is learned for each player, providing a low-dimensional and interpretable vector of positive values that summarizes an offensive player type as a non-negative linear combination of the basis surfaces. This unsupervised technique not only reaffirms common intuitions about player types, but also estimates how much of each type each player is.

Through synthesis with spatial shooting ability, the method gives understanding of the effectiveness of defensive strategies in pressuring an offensive player into low-outcome positions. This provides a novel way to measure and compare defensive schemes and players. We also show that this technique can be extended to other facets of the game, such as passes that lead to shots. This allows us to both visualize and quantify less salient aspects of the game.

TROUBLE WITH THE CURVE: IDENTIFYING CLUSTERS OF MLB PITCHERS USING IMPROVED PITCH CLASSIFICATION TECHNIQUES

Michael A. Pane†, Samuel L. Ventura, Rebecca C. Steorts, and Andrew C. Thomas

Carnegie Mellon University, Pittsburgh, PA, USA

† E-mail: *mpane@andrew.cmu.edu*

The PITCHf/x database, which records the location, velocity, and trajectory of every pitch thrown in Major League Baseball (MLB), has allowed the statistical analysis of MLB to flourish since its introduction in late 2006. Pitches in PITCHf/x are currently classified

into pitch types (e.g. "fastball" or "curveball") with MLB's proprietary training data and neural networks clustering and classification algorithm. Since pitch type training data can be expensive, unreliable, and difficult to obtain, we take an unsupervised approach for labeling pitch types. We use model-based clustering with a multivariate Gaussian mixture model and an adjusted Bayesian Information Criterion to identify clusters of pitches. We then label the clusters with pitch types using a heuristic pitch classification algorithm. We verify our results via cross validation, validation by prediction strength, and through visual inspection. Using features of the resulting pitch type clusters for each pitcher, we can cluster MLB pitchers into groups of similar pitchers using basic clustering techniques, such as k-means. Our method builds a strong foundation towards addressing many open MLB research questions, such as predicting pitcher injury and identifying favorable batter-pitcher matchups.

CROSSING IN SOCCER HAS A STRONG NEGATIVE IMPACT ON SCORING: EVIDENCE FROM THE ENGLISH PREMIER LEAGUE AND THE GERMAN BUNDESLIGA

Jan Vecer[†]

Frankfurt School of Finance and Management

[†] E-mail: *j.vecer@fs.de*

Crossing in soccer plays a significant role in scoring, about 23% of all goals scored in the recent seasons of the English Premier League are the result of crosses. Set play crosses (after free kicks or corners) represent about 8% of all goals and open play crosses represent about 15% of all goals. However, crossing from an open play is hugely inefficient, only 1 open cross out of 92 leads to a goal on average. Statistical evidence indeed confirms that games with smaller number of open crosses tend to lead to more goals. When we estimate the impact of open crossing on scoring of the individual teams using multilevel Poisson regression model, we conclude that the net effect of crossing is negative for all teams. An average team is expected to score additional 0.656 goals per game if it reduced open crossing. The quality of the team is the major explanatory factor on the number of such missed scoring opportunities, stronger teams during attack miss more goal opportunities when crossing than weaker teams in general. Teams such as Arsenal, Chelsea, Liverpool, Manchester City or Tottenham have a potential of scoring an extra goal per match if they reduced open crossing. A reversed picture is seen in the defense analysis, more goal opportunities are missed when crossing against weak teams than crossing against strong teams in general. Interestingly, the actual conversion of open crosses to goals plays only a minor role for explaining the impact of open crossing on goals.

Poster Presentation Abstracts

GOLF HANDICAP SCORES MODELED VIA DISTRIBUTION OF AVERAGES OF MOVING ORDER STATISTICS

Sonia Bandha[†1], Thomas Spencer III[2], David J. Horntrop[1]

Department of Mathematical Sciences, New Jersey Institute of Technology[1], School of Management, Walden University[2]

[†] E-mail: *sb373@njit.edu*

Previous studies in golf have analyzed and compared the performances of individual golfers based on their handicaps and the nature of a tournament. The purpose of USGA Handicap System is to allow golfers of varying skill levels to compete fairly. The goal of this research is to study the effectiveness of the current handicapping system. To accomplish this, a golf handicap index is viewed as a moving average of moving order statistics. Simulation is used to obtain the handicap scores and their corresponding average handicap indices in order to observe how changes and trends in the scores affect the average handicap indices.

OPENWAR: AN OPEN SOURCE SYSTEM FOR OVERALL PLAYER PERFORMANCE IN MAJOR LEAGUE BASEBALL

Ben Baumer[†1], Gregory Matthews[2], Shane Jensen[3]

Smith College[1], University of Massachusetts[2], and the Wharton School of the University of Pennsylvania[3]

[†] E-mail: *bbaumer@smith.edu*

Within baseball analytics, there is substantial interest in comprehensive statistics intended to capture overall player performance. One such measure is Wins Above Replacement (WAR), which aggregates the contributions of a player in each facet of the game: hitting, pitching, base-running and fielding. However, current versions of WAR depend upon proprietary data, ad hoc methodology, and opaque calculations. We propose a competitive aggregate measure, openWAR, using public data, more rigorous methodology, and transparent calculations. We discuss a principled baseline compared to the nebulous concept of a "replacement" player. Finally, we use simulation-based techniques to provide variance estimates for our openWAR measure.

CLOSING THE GAP: INFERRING DEFENSIVE INTENT FROM NBA PLAYER TRACKING DATA

Alexander Franks†

Harvard University

† E-mail: *afranks@fas.harvard.edu*

The use of SportVU tracking technology in the NBA, which records player locations at 25 frames per second, has the potential to revolutionize basketball analytics. In particular, this data should enable us to quantify player and team defense in a way that has previously not been possible. However, answering questions about defensive competence requires an understanding of intent. Using the SportVU data, we develop an unsupervised model for identifying "who's guarding whom" during the course of an NBA possession. We treat each defender as an agent attempting to optimize their defensive position at every moment in time. We model a defender's position using a Gaussian hidden Markov model with five latent states corresponding to the offensive player being guarding. We identify the "optimal" position for a defender as a linear combinations of spatial coordinates (e.g. ball and offender position). The model also allows for the inclusion of player specific covariates and information about the interaction between defenders. Thus, we are not only able to infer who a defender is guarding but how they defend them. The model can be used to to compare the effectiveness of different individual defenders in different ways. For instance, with this model, we can estimate the number of points a defender is personally responsible for giving up ("points against").

ESTIMATING GOAL PROBABILITIES IN THE NHL

Calla Glavin, Brian Macdonald†, Nicholas Clark

United States Military Academy, West Point, NY USA

† E-mail: *bmac@jhu.edu*

Using hierarchical Bayesian techniques, we estimate the probability that an NHL shot will be a goal based on several details of the shot. Predictors include those which have been used in previous studies, like shot location, and whether or not the shot was a rebound shot. We also use the goalies and the shooters involved in each shot as predictors, as well as the glove hand of the goalie, the handedness of the shooter, and the fatigue of the players involved. The results of the model can be used to analyze goalies, shooters, team offenses, and team defenses. For example, the coefficient for a goalie can be used to form an adjusted save percentage statistic for that goalie which accounts for the quality of shots that he faces. An added benefit is that adjusted save percentage is pulled towards the league average for goalies

that have faced relatively few shots. The coefficients for shooters can be used in a similar way to develop a statistic that quantifies finishing ability, or a player's ability to capitalize on the shots that he takes. Year-to-year correlations of these statistics, as well as the ability of these statistics to predict future performance, are also examined.

PREDICTING HEISMAN TROPHY VOTING: A BAYESIAN ANALYSIS

Daniel Heard[†]

Duke University

[†] E-mail: *dph11@stat.duke.edu*

The annual voting for the Heisman trophy is a highly publicized event. Many theories exist regarding the patterns of voters and certain biases that exits. Here, we look at a generalized linear mixed effects model to examine the effects of certain offensive player attributes in the votes they receive. The data used in model fitting included Heisman trophy voting results from 2000-2010, and predictions were made for the 2011 and 2012 seasons. The data illuminate certain regional, conference and positional effects contributing to candidates receiving higher vote totals.

AN UNCONSTRAINED MODEL FOR COVARIANCE STRUCTURE FOR MAJOR LEAGUE BASEBALL BATTER'S SALARY WITH THE WEIGHTED OFFENSIVE AVERAGE

Chulmin Kim[†]

University of West Georgia, Carrollton, GA, USA

[†] E-mail: *ckim@westga.edu*

The positive-definiteness requirement for the covariance matrix may impose complicated nonlinear constraints on the parameters. Kim (2012) proposes an unconstrained parameterization for the covariance structure for the multivariate longitudinal data, and then to model its parameters parsimoniously. Kim (2013) also introduced the weighted offensive average (WOA) as a variation of on base plus slugging (OPS) which explains not only a batter's hitting performance but also his non-hitting performance to generate runs for his team such as stolen bases, walks, and etc. We adopt Kim's unconstrained model for the covariance structure for Major League Baseball batter's salary with the Weighted Offensive Average.

DESIGNATED PERFORMANCE? PLAYER PERFORMANCE AND THE EFFICACY OF THE DESIGNATED PLAYER RULE IN MAJOR LEAGUE SOCCER

Jun Woo Kim[†1], Justin Lovich[2], Seung Hoon Jeong[3]

The State University of New York at Brockport, NY, USA[1], The Florida State University, FL, USA[2], Kyung Hee University, Yongin, South Korea[3]

[†] E-mail: *jk07e@my.fsu.edu*

In 2007, Major League Soccer (MLS) adopted Designated Player Rule, commonly referred to as the "David Beckham Rule", as part of collectively bargained salary regulations. Previous research in the area of labor markets in the MLS has focused on the relationship between player salary and on-field performance (Lee & Harris, 2012) and the impact of superstar status on player wages (Kuethe & Motamed, 2010). Such emphasis on players. salaries and various individual and performance factors (i.e., wage determination) leaves unaddressed the on-field efficacy of the MLS-designated player rule and performance attributes. Therefore, this study attempts to illuminate whether the Designated Players in MLS perform better than non-Designated Players and to clear the fog on the determinants of player performance. Data include five seasons (2007-2011) with 2095 observations. After conducting a series of robust regression analyses, we found that Designated Players play more minutes, tally more assists, and score more goals than non-Designated Players. Importantly, we identified and affirmed key determinants of player performance, including playing positions, number of games played and started, number of fouls committed and suffered, number of offside infractions committed, number of yellow cards received, number of career years, and country of origin, while controlling for time-constant dummy variables (i.e., year of the season played) and players' teams. This study extends previous work and fills a prominent gap in the literature by providing indirect evidence of the on-field efficacy of the MLS-designated player rule.

PERFORMANCE ANALYSIS OF BATSMEN AND BOWLERS IN CRICKET

Ananda Manage[†1], Steve Scariano[1], Danush Wijekularathna[2]

Sam Houston State University[1], Texas Tech University[2]

[†] E-mail: *manage@shsu.edu*

Cricket is a team sport played between two teams. Usually each team consists of eleven players. Performance analysis of cricket players is always an intricate task due to the correlated nature of the variables used to quantify contributions to the team. Lack of transparency of

current methods, probably due to commercial confidentiality, creates a necessity for new and lucent evaluative methods. Here we present a brief review of existing methods. And also we present a simple, yet straightforward, method for analyzing the performance of cricket players.

SURVIVAL ANALYSIS OF THE MEN'S 100 METER DASH RECORD

Reza Noubary[†]

Bloomsburg University

[†] E-mail: *rnoubary@bloomu.edu*

In 2012 London Olympic seven out of eight finalists in men's 100 meter dash crossed the finish line in under 10 seconds. This and recent performances of sprinters such as Bolt and Blake have raised the bar in this event and has made experts wonder, not whether but when a new record will be set and how much it will lower the present one. Seeking for an answer, many prominent researchers have tried to model the improvements of records over time with a goal of forecasting the future records. This article describes our attempt based on analyzing the historical progression of the records.

HOW THE WEST WILL BE WON: USING MONTE CARLO SIMULATIONS TO ESTIMATE THE EFFECTS OF NHL REALIGNMENT

Stephen Pettigrew[†]

Harvard University, Cambridge, MA, USA

[†] E-mail: *pettigrew@fas.harvard.edu*

The NHL has realigned its conferences and divisions, and starting in the 2013-2014 season the Eastern Conference will feature 16 teams and the Western Conference will feature 14. Because there are 8 playoff spots available in both conferences, 57% of teams in the West will make playoffs, compared to just 50% of teams in East. As a result we should expect that, on average, the last team to make the playoffs in the West will have a worse record than the last playoff team in the East. I call the difference in points earned by the 8th seed in each conference the "conference gap." My purpose in this paper is to figure out how big we should expect this gap to be under the new alignment. Using tens of thousands of Monte Carlo simulated seasons, I demonstrate that the conference gap will be about 2 or 3 points, meaning that Eastern Conference teams hoping to make the playoffs will have to win 1 to 2

games more than Western Conference playoff-hopefuls. I also show that almost 40% of the time, the 9th place team in the Eastern Conference would have made the playoffs if they had only been located further west. My findings have tremendous implications on ensuring that the NHL operates under rules that are fair and equitable for all teams.

MODELLING UMPIRE MISCLASSIFICATION OF BALLS AND STRIKES USING PITCH FX DATA.

Justin Post[†], Jason Osborne

North Carolina State University, Raleigh, NC, USA

[†] E-mail: *jbpost2@ncsu.edu*

In major league baseball, a tremendous amount of new information about pitch location, trajectory, velocity and movement has become available with the development of "PITCHf/x" technology. The precise determination of pitch location, along with other video replays, have heated up the discussion of expanding the use of instant replay to improve umpiring during games. One suggestion has even been to give managers the ability to dispute a fixed number of calls, similar to the "challenge flags" afforded head coaches in the National Football League. We analyze PITCHf/x data to investigate the effects of a number of factors on pitch misclassification rates, including pitch type, velocity and umpire fatigue. Some discussion of techniques to acquire these data from public sources using R and perl is included.

EVALUATING PLAYER AND TEAM PERFORMANCE IN AUSTRALIAN RULES FOOTBALL

John Sannini[†]

Temple University Philadelphia, PA

[†] E-mail: *jsannini@yahoo.com*

Data for three Australian Rules Football games consisting of information for all game possessions and GPS player positioning was analyzed to determine the best statistical indicators of team and individual player success.

The number of shots on goal amassed by a team is the best single indicator of the number of points said team will score in a game. In a similar vein a given teams chances of winning a game are strongly tied to their ability to attempt more shots on goal than the opponent.

Team possessions directly resulting from a turnover forced in the mid-forward and forward areas (i.e. the opponents half of the field) end in a goal being scored roughly 22% of the time; all other possessions end in a goal being scored 6% of the time.

Due to the relatively small 12.5% of possession time non-starting players are used in Australian Rules Football and the small quantity of bench players, more precise player evaluations may be made based solely upon the points scored for and against their team while they are in the game. Because of the tendency for players to occupy more than 1 of the 5 position groups (i.e. D, F, HF, M, W) on the field during a game, this plus/minus metric can further be broken down by time spent at each position in order to identify the depth of talent at each position and aid in putting individual players into field assignments where they are more likely to succeed.

"SET ONE" OR "SET TWO"?: WINNING STRATEGY IN A SQUASH COURT

Eiki Satake[†1], Benjamin Atchinson[2], Nichlous Sedlock[2], Rob Page[2]

Emerson College, Boston, MA[1]; Framingham State University, Framingham, MA[2]

[†] E-mail: *eiki_satake@emerson.edu*

Under the current International Rules of Squash, a player can only earn a point if the player is also the server. If the server loses a rally, she loses the serve and the score remains unchanged. The ultimate goal of each player is to be the first to reach nine points, thus being declared the winner. The only exception occurs when a score is tied at eight. In this case, the player who reaches eight points first (always the receiver) must choose whether to play to nine points (called "Set One") or ten points (called "Set Two") to decide the winner.

Consider two players, called A and B. Suppose that the score is tied at eight, with player A to serve. The question of interest to player B is whether or not it is advantageous to call for "Set One" or "Set Two," given the various skill levels of each player.

The authors established a Bayesian Probability model and programmed a statistical simulation, when two players' skill levels are varied (prior probabilities) in order to calculate the probabilities of winning the game (posterior probabilities) under "Set One" and "Set Two" conditions, and determine the accuracy of the model, i.e., how well the model predicts the actual outcomes of the observed games.

AGE EFFECTS IN THE MLB DRAFT

Justin Sims, Vittorio Addona[†]

Macalester College

[†] E-mail: *addona@macalester.edu*

Major League Baseball (MLB) franchises expend an abundance of resources on scouting in preparation for the June Amateur Draft. In addition to the classic "tools" assessed, another

factor considered is age: younger players may get selected over older players of equal ability because of anticipated development, whereas college players may get selected over high school players due to a shortened latency before reaching the majors. Additionally, Little League rules in effect until 2006 operated on an August 1–July 31 year, meaning that, in their youth, players born on August 1 were the eldest relative to their cohort. We examine the performance of players selected in the June Draft from 1987-2011. We find that for high school draftees, both relative age and absolute age have a significant negative effect on the number of games played in MLB and on wins above a replacement player (WAR). Absolute age also significantly reduces the odds that a drafted player reaches the majors, but relative age does not. For college draftees, we find that older, but not relatively older, players appear in fewer MLB games. Neither age measure has a significant effect on WAR or on the odds of reaching the majors for college players. Had the draft market operated efficiently, neither relative age nor age on draft day would have captured additional variation in performance after controlling for draft position and other factors. We conclude that teams have undervalued both absolutely, and relatively, younger high school players in the draft.

PLAYERS, PASSIVITY, AND PENALTIES: A STUDY OF AGGRESSION IN THE NFL

Kevin Snyder[†1], Michael Lopez[2]

Southern New Hampshire University, Manchester, NH[1], *Brown University, Providence, RI*[2]

[†] E-mail: *kevin.m.snyder@gmail.com*

While previous work has identified scenarios where football coaches are too conservative (punting on 4th down, kicking FGs, etc.), few, if any, have considered player aggression (Kovash & Levitt, 2009; Rosen & Wilson, 2007; Alamar, 2010). Further, previous research in hockey suggests that aggressiveness is heightened early in games due to the ability to recover from a penalty (Jewell, 2009), in addition to enhancing increasing the team's chance of winning (Widmeyer & Birch, 1984). Based on this literature and our data set, we suggest that football players exhibit unnecessarily conservative levels of aggression early in games. We sample all NFL plays from 2002 to 2012, isolating the occurrence of either offensive holding or defensive pass interference calls. Even after accounting for game and play specific variables, including team characteristics, type of play, and the game's score, we find the likelihood of both penalty types follows a quadratic trend, low at the beginning and ends of the game, but high in the middle. In particular, an extremely low penalty rate early in the game suggests either a passive approach on behalf of the players or a purposeful burn-in period from the officials, where infractions are called with more leniencies. We surmise it is more plausible that players pace themselves and forego opportunities for aggressive play.

IS SMALL BALL A NEW WAY OF WINNING NBA GAMES?

Masaru Teramoto[†1], Chad L Cross[2]

Drexel University, Philadelphia, PA, USA[1], Crossroads Wellness, LLC, Las Vegas, NV, USA[2]

[†] E-mail: *Masaru.Teramoto@drexel.edu*

Traditionally, size is considered a crucial factor in basketball. Recently, however, much debate has been made over playing with small lineups or "small ball" in the National Basketball Association (NBA). The aim of this study is to examine whether playing small ball is associated with better outcome of the NBA game analyzing data of the 2011-12 NBA season.

Each of the 30 teams was ranked according to the average height of players weighted by each player's minutes played. Variables defined include the weighted means of height for: 1) all players, 2) front court players, and 3) back court players. In that season, teams ranked in the bottom half for the average height of all players won 48.9% of the regular season games. Teams with smaller lineups tended to give up more points per possession, whereas height did not appear to influence offensive production. In the playoffs, playing small ball was associated with higher winning percentage. In particular, teams with smaller lineups in the front court positions performed better on defense (rank correlation = 0.462). Three of the four teams that advanced to the Conference Finals (Miami Heat, Oklahoma City Thunder, and Boston Celtics) were ranked in the bottom half of the 16 playoff teams for the average height of all players.

The success of these teams may further promote the usage of small ball especially in the playoffs. More analysis is needed to truly determine the effectiveness of playing small ball on winning NBA games.

EWING THEORY AND HAWTHORNE EFFECTS IN MAJOR LEAGUE BASEBALL

A.C. Thomas[†]

Carnegie Mellon University

[†] E-mail: *acthomas@stat.cmu.edu*

"A star player who departs a team makes the team better by their absence." This is the essence of Ewing Theory, popularized by ESPN's Bill Simmons and coined by his associate Dave Cirelli, when they observed that teams with Patrick Ewing often performed better without him playing than with him. While the media perception of the team in question is what drives the legend of Ewing Theory, the notion that a team's non-star players are

motivated to perform better without their star to lead them is very much of interest. Since baseball is a sport where interactions between players are less obvious, we use data from Retrosheet to show that there are noticeable effects on team performance when star players are injured and when they return.

We find a small but statistically significant increase in the rest of the team's offensive output in games following the injury, consistent with the Ewing effect. In the games after the return of the injured star, however, we also find an increase in team offense, rather than the hypothesized decrease. We submit instead that the additional offensive output is more consistent with the Hawthorne effect – the additional focus put on the team at the injury and during the return incentivizes the team to perform better due to the attention they receive.

STATISTICAL ANALYSIS OF CONCUSSION HISTORY AND EXECUTIVE FUNCTION

Yorghos Tripodis[†], Philip Montenigro, Robert A. Stern

Boston University, Boston, MA, USA

[†] E-mail: *yorghos@bu.edu*

Concussion is a brain injury defined as a complex pathophysiological process affecting the brain, induced by biomechanical forces. The Center for Disease Control estimates between 1.6 and 3.8 million sports-related concussive injuries occur annually in the United States. In our current study we used the LEGEND (Longitudinal Examination to Gather Evidence of Neurodegenerative Disease) dataset. We included 162 participants that played football at any level but have not participated in any other contact sport. Participants were asked to spontaneously estimate the number of concussions they had sustained throughout their lives. After providing a number, interviewers read a formal definition of concussion that included all the accepted symptoms. Participants were then asked to re-estimate the number of concussions. They were also asked to estimate the number of concussions after which they lost consciousness as well as the total number of major concussions (with or without loss of consciousness). We assessed the relationship between concussion history using each definition of concussion separately and self-report measures of executive function (BRIEF-A). We estimated this relationship using independent linear regression models, and mixed effect models to allow for correlated outcomes from the same participant. We also corrected for the presence of heteroscedasticity. We show that the choice of statistical method affects significantly conclusions. Although all definitions of concussion were significantly related to the outcomes using independent linear regression, only estimate based on the provided concussion definition is significantly related to the outcomes when appropriate adjustments for multiple measurements and heteroscedasticity are made.

GOING FOR IT... WHAT FACTORS CONTRIBUTE TO A FOURTH DOWN CONVERSION?

Yuting Wang, Tracy L Morris†

University of Central Oklahoma, Edmond, OK, USA

† E-mail: *tmorris2@uco.edu*

It's fourth down and two yards to go. The ball is on your own 40 yard line. Do you go for the first down or punt? Coaches have to consider so many factors when making this decision, and must do so in a matter of seconds. Consequently, coaches tend to make the conservative decision to punt. In this research, data was collected from www.cfbstats.com concerning all fourth downs during the 2011 college football season. Only fourth downs for which the ball was either passed or rushed were included in the data set. Logistic regression was used to construct a model for estimating the probability of converting a fourth down. Variables considered for inclusion in the model were playing surface, team (home or away), site (team or neutral), opponent (BCS or non-BCS), period number, score, yards to go, and spot.

BASEBALL SCOUTING REPORTS VIA A MARKED POINT PROCESS FOR PITCH TYPES

Andrew Wilcox†, Elizabeth Mannshardt

North Carolina State University, Raleigh, NC, USA

† E-mail: *agwilco2@ncsu.edu*

The implementation of the PITCHf/x system in major league ballparks has allowed for the collection of rich spatial data sets regarding pitch location. In this paper we use PITCHf/x data collected on Clayton Kershaw during the 2012 regular season to explore the use of marked point processes to model pitch location. Through this modeling we show that each pitch type follows its own independent point process where the location of a pitch has a significant relationship with the pitch type. For the specific case of Clayton Kershaw we find that fastballs are thrown most frequently but with a small dependence on location, while changeups are not thrown often but are always on the outer half of the strike zone. We also explore how this point process changes when Kershaw is pitching with two strikes on a batter. This information is valuable to Major League hitters and as such marked point processes could further be used in the creation of scouting reports for hitters.

Directions to ***Grafton Street Pub & Grill***

Walking directions: Exit the Harvard Science Center by the doors near Lecture Halls C and D, and go through the gates into Harvard Yard. Walk to the path on the left of the library (large building with tall columns), and exit Harvard Yard onto Mass Ave. Cross the street, and walk to the left one short block. Grafton Street Pub and Grill will be on your right.

Grafton Street Pub and Grill
1230 Massachusetts Ave.
Cambridge, MA 02138
(617) 497-0400