SPORTS ANALYTICS
From A To Z
Methods, Definitions & Real-Life Applications
VICTOR HOLMAN
# Table of Contents

About Victor Holman ........................................................................................................................................... 1

About This Book ................................................................................................................................................ 2

Introduction to Analytic Methods ...................................................................................................................... 3

Sports Analytics Maturity Model ......................................................................................................................... 4
  Sports Analytics Maturity Model Phases ........................................................................................................ 4
  Sports Analytics Key Success Areas ............................................................................................................. 5

Allocative and Dynamic Efficiency .................................................................................................................... 7
  Optimal Strategy in Basketball .................................................................................................................. 7

Backwards Selection Regression ......................................................................................................................... 9
  Competition between Sports Hurts TV Ratings: How to Shift League Calendars to Optimize Viewership .... 9

Bayesian Generalized Linear Model ..................................................................................................................... 11
  Shrinkage Estimation of NFL Field Goal Success Probabilities ................................................................. 11

Bayesian Hierarchical Latent Variable Models .................................................................................................. 13
  Eye on the Ball: The Relationship Between Sensorimotor Abilities and On-Field Performance in Professional Baseball .......................................................................................................................... 13

Bayesian Hierarchical Model .............................................................................................................................. 15
  Estimating an NBA Player's Impact on His Team's Chances of Winning ......................................................... 15

Bayesian Model and Simple OLS Regression ....................................................................................................... 17
  Moral Hazard in Long-Term Guaranteed Contracts: Theory and Evidence from the NBA ......................... 17

Bayesian Multinomial Model ............................................................................................................................... 19
  Exploring the Potential of the Plus/Minus in NCAA Women's Volleyball via the Recovery of Court Presence Information ............................................................................................................................................. 19

Binary Logic Model ............................................................................................................................................ 21
  Matchup Models for the Probability of a Ground and a Ground Ball Hit ......................................................... 21

Bipartite Graph Algorithms .................................................................................................................................. 23
  Player Centrality on NBA Teams .............................................................................................................. 23

Bland Altman Plot Analysis ................................................................................................................................. 25
  Bland Altman Plot Analysis in Sports Analytics ........................................................................................ 25

Bonferroni Adjusted Comparisons ...................................................................................................................... 27
  A Closer Look at the Prevalence of Time Rule Violations and the Inter-Point Time in Men's Gland Slam Tennis ................................................................................................................................................. 27

Bookmaker Consensus Model ............................................................................................................................ 29
Searching for the GOAT of Tennis Win Prediction ................................................................. 29
Bootstrap Simulation .................................................................................................................. 31
Improving Fairness in Match Play Golf through Enhanced Handicap Allocation .................. 31
Boruta Algorithm ....................................................................................................................... 33
An Investigation of Three-Point Shooting through an Analysis of NBA Player Tracking Data ........................................................................................................... 33
Brownian Motion Process ........................................................................................................ 35
The Implied Volatility of a Sports Game .................................................................................... 35
Chapman- Kolmogorov Equations ............................................................................................. 37
Evaluating NBA End-of-Game Decision-Making ........................................................................ 37
Clustering .................................................................................................................................. 39
Big 2’s and Big 3’s: Analyzing How a Team’s Best Players Complement Each Other ............... 39
Coaching Assistance Models ...................................................................................................... 41
Coaching Assistance Models as a Method of Sports Analytics ................................................ 41
Recognizing and Analyzing Ball Screen Defense in the NBA .................................................. 42
Competitiveness Metrics ............................................................................................................ 44
The Competitiveness of Games in Professional Sports Leagues ................................................ 44
Computer Vision Techniques ..................................................................................................... 45
Using Computer Vision and Machine Learning to Automatically Classify NFL Game Film and Develop a Player Tracking System ............................................................................. 45
Conditional Probability ............................................................................................................... 47
Applying Conditional Probability to Sports Analytics .............................................................. 47
Conditional Random Field ......................................................................................................... 48
"Quality vs. Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data .................................................................................................................. 48
Conservation of Runs Framework ............................................................................................... 50
openWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball ...................................................................................................................... 50
Convex Hull ................................................................................................................................ 52
Predicting Golf Scores at the Shot Level ..................................................................................... 52
Convolutional Neural Networks .................................................................................................. 54
Mythbusting Set-Pieces in Soccer .............................................................................................. 54
Cross-Validated Root-Mean-Squared Error ................................................................................. 56
Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing ......................... 56
Data Science Methods ................................................................................................................ 58
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Science Methods in Sports Analytics</td>
<td>58</td>
</tr>
<tr>
<td>Data Tables</td>
<td>60</td>
</tr>
<tr>
<td>The Effect of Ball Wear on Ball Aerodynamics: An Investigation Using Hawk-Eye Data</td>
<td>60</td>
</tr>
<tr>
<td>Data Visualization</td>
<td>62</td>
</tr>
<tr>
<td>Data Visualization as a Method of Sports Analytics</td>
<td>62</td>
</tr>
<tr>
<td>Applying Data Visualization in Sports Analytics</td>
<td>63</td>
</tr>
<tr>
<td>Deductive Reasoning and Inductive Reasoning</td>
<td>65</td>
</tr>
<tr>
<td>Deductive Reasoning and Inductive Reasoning as a Method of Sports Analytics</td>
<td>65</td>
</tr>
<tr>
<td>Deep Imitation Learning</td>
<td>67</td>
</tr>
<tr>
<td>Data-Driven Ghosting Using Deep Imitiation Learning</td>
<td>67</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>69</td>
</tr>
<tr>
<td>Deep Learning of Player Trajectory Representations for Team Activity Analysis</td>
<td>69</td>
</tr>
<tr>
<td>Deep Reinforcement Learning</td>
<td>71</td>
</tr>
<tr>
<td>The Advantage of Doubling: A Deep Reinforcement Learning Approach to Studying the Double Team in the NBA</td>
<td>71</td>
</tr>
<tr>
<td>Demand Model</td>
<td>73</td>
</tr>
<tr>
<td>Evaluating the Effectiveness of Dynamic Pricing Strategies on MLB Single-Game Ticket Revenue</td>
<td>73</td>
</tr>
<tr>
<td>Descriptive, Predictive and Prescription Analytics</td>
<td>75</td>
</tr>
<tr>
<td>Three Phases of Sports Analytics</td>
<td>75</td>
</tr>
<tr>
<td>Econometrics</td>
<td>77</td>
</tr>
<tr>
<td>Experience and Winning in the National Basketball Association</td>
<td>77</td>
</tr>
<tr>
<td>Egocentric Ghosting Model</td>
<td>79</td>
</tr>
<tr>
<td>Learning an Egocentric Basketball Ghosting Model using Wearable Cameras and Deep Convolutional Networks</td>
<td>79</td>
</tr>
<tr>
<td>Encounter Detection Algorithm</td>
<td>81</td>
</tr>
<tr>
<td>Esports Analytics through Encounter Detection</td>
<td>81</td>
</tr>
<tr>
<td>Euclidean Distances</td>
<td>83</td>
</tr>
<tr>
<td>Acceleration in the NBA: Towards an Algorithmic Taxonomy of Basketball Plays</td>
<td>83</td>
</tr>
<tr>
<td>Event Tree Models</td>
<td>85</td>
</tr>
<tr>
<td>Accounting for Complementary Skill Sets When Evaluating NBA Players' Values to a Specific Team</td>
<td>85</td>
</tr>
<tr>
<td>Expected Possession Value</td>
<td>85</td>
</tr>
<tr>
<td>POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data</td>
<td>86</td>
</tr>
<tr>
<td>Expected Rewards</td>
<td>88</td>
</tr>
<tr>
<td>How to Play Strategically in Fantasy Sports (and Win)</td>
<td>88</td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Experiments, Events, Probabilities and Odds</td>
<td>90</td>
</tr>
<tr>
<td>Applying Experiments, Events, Probabilities and Odds to Sports Analytics</td>
<td>90</td>
</tr>
<tr>
<td>Extra-Trees Classifiers</td>
<td>91</td>
</tr>
<tr>
<td>Frequencies</td>
<td>93</td>
</tr>
<tr>
<td>No Referee Bias in the NBA: New Evidence with Leagues' Assessment Data</td>
<td>93</td>
</tr>
<tr>
<td>Frequency Distributions</td>
<td>95</td>
</tr>
<tr>
<td>Frequency Distribution as a Method of Sports Analytics</td>
<td>95</td>
</tr>
<tr>
<td>Frequency Distributions as a Method of Sports Analytics</td>
<td>96</td>
</tr>
<tr>
<td>Gaussian Copula Model</td>
<td>98</td>
</tr>
<tr>
<td>Generalized Linear Spatial Regression Model</td>
<td>100</td>
</tr>
<tr>
<td>The Pressing Game: Optimal Defensive Disruption in Soccer</td>
<td>100</td>
</tr>
<tr>
<td>Graph, Slope, Correlation</td>
<td>102</td>
</tr>
<tr>
<td>A Development Model to Guide the Recruiting of Female Shot Putters at the NCAA Division 1 Championship Level</td>
<td>102</td>
</tr>
<tr>
<td>Graphing, Stratification, Correlation, and Causation</td>
<td>104</td>
</tr>
<tr>
<td>Fast Starters and Slow Finishers: A Large-Scale Data Analysis of Pacing at the Beginning and End of the Marathon for Recreational Runners</td>
<td>104</td>
</tr>
<tr>
<td>Grey Relational Analysis</td>
<td>106</td>
</tr>
<tr>
<td>Ranking Regular Seasons in the NBA's Modern Era Using Grey Relational Analysis</td>
<td>106</td>
</tr>
<tr>
<td>Heat Maps</td>
<td>108</td>
</tr>
<tr>
<td>From Sports to Science: Using Basketball Analytics to Broaden the Appeal of Math and Science Among Youth</td>
<td>108</td>
</tr>
<tr>
<td>Heuristic Algorithms</td>
<td>110</td>
</tr>
<tr>
<td>Heuristic Algorithms in Sports Analytics</td>
<td>110</td>
</tr>
<tr>
<td>Hidden Markov Model</td>
<td>112</td>
</tr>
<tr>
<td>Counterpoints: Advanced Defensive Metrics for NBA Basketball</td>
<td>112</td>
</tr>
<tr>
<td>Image Transfer Model</td>
<td>114</td>
</tr>
<tr>
<td>Fans' Responses to the National Basketball Association's (NBA) Pilot Jersey Sponsorship Program: An Experimental Approach</td>
<td>114</td>
</tr>
<tr>
<td>Inertial Sensor Training Load Metrics</td>
<td>116</td>
</tr>
<tr>
<td>Volume and Intensity are Important Training Related Factors in Injury Incidence in American Football Athletes</td>
<td>116</td>
</tr>
</tbody>
</table>
Injury Risk Mitigation System ................................................................. 118
  An Enhanced Metric of Injury Risk Utilizing Artificial Intelligence ............................................. 118
Interviews and Surveys .................................................................................................................. 119
  Decision-Making in the National Basketball Association .............................................................. 119
Kelly Criteria ............................................................................................................................... 121
  Modified Kelly Criteria .............................................................................................................. 121
Kernel Density Estimation ............................................................................................................ 123
  New Metrics for Evaluating Home Plate Umpire Consistency and Accuracy ......................... 123
Knapsack Problem ....................................................................................................................... 125
  Optimal Decision-Making in the NBA Free Agency Market ....................................................... 125
Latent Trajectory Model .............................................................................................................. 127
  A Finite Mixture Latent Trajectory Model for Modeling Ultrarunners’ Behavior in a 24-Hour Race ... 127
Linear Discriminant Analysis ....................................................................................................... 129
  Using Multi-Class Classification Methods to Predict Baseball Pitch Types .............................. 129
  Using PITCHf/x to Model the Dependence of Strikeout Rate on the Predictability of Pitch Sequences .................................................................................................................. 130
Linear Regression ........................................................................................................................ 132
  Drafting Errors and Decision Making Bias in the NBA Draft .................................................... 132
Log Difference ............................................................................................................................. 134
  Gender Differences in Marathon Pacing and Performance Prediction ...................................... 134
Machine Learning ...................................................................................................................... 136
  Machine Learning as a Method of Sports Analytics ............................................................... 136
  An Open-Sourced Optical Tracking and Advanced eSports Analytics Platform for League of Legends 137
  High-Resolution Shot Capture Reveals Systematic Biases and an Improved Method for Shooter Evaluation .................................................................................................................. 138
Markov Chain Decision Processes ............................................................................................... 140
  Replaying the NBA ............................................................................................................... 140
Mathematical Programming ......................................................................................................... 141
  Mathematical Programming in Sports ...................................................................................... 141
Mean and Median ........................................................................................................................ 143
  Measuring Mean and Median in Sports Data ............................................................................. 143
Mixed Integer Programming ....................................................................................................... 144
  An Analytical Approach for Fantasy Football Draft and Lineup Management ..................... 144
Mixed-Integer Linear Programming ............................................................................................. 146
Does Artificial Grass Affect the Competitive Balance in Major League Soccer? ........................................ 177

Alleviating Competitive Imbalance in NFL Schedules: An Integer-Programming Approach ............ 146

Are Daily Fantasy Sports Gambling? .................................................................................................. 147

Modern Portfolio Theory ...................................................................................................................... 149

Three Point Shooting and Efficient Mixed Strategies: A Portfolio Management Approach .......... 149

Monte Carlo Simulation ....................................................................................................................... 151

Misadventures in Monte Carlo ........................................................................................................... 151

Multiple Imputation ............................................................................................................................. 153

An Examination of Statistical Disclosure Issues Related to Publication of Aggregate Statistics in the Presence of a Known Subset of the Dataset using Baseball Hall of Fame Ballots .................................................. 153

Multivariate Statistics .......................................................................................................................... 155

Multivariate Statistics as a Method of Sports Analytics ..................................................................... 155

Negative Binomial Regression Model ................................................................................................. 157

The Nature of Regional Bias in Heisman Voting .................................................................................. 157

Network Analytics ............................................................................................................................... 159

Network Analytics as a Method of Sport Analytics ............................................................................ 159

Network Diffusion Methods ................................................................................................................. 161

A Network Diffusion Ranking Family that Includes the Methods of Markov, Massey, and Colley ...... 161

Network-Style Model ............................................................................................................................ 163

A Method for Using Player Tracking Data in Basketball to Learn Player Skills and Predict team Performance .................................................................................................................................................. 163

Neural Network Algorithm .................................................................................................................... 165

(batter/pitcher) 2vec: Static-Free Talent Modeling With Neural Player Embeddings ...................... 165

Neural Network Algorithm ..................................................................................................................... 167

Classifying NBA Offensive Plays Using Neural Networks ................................................................. 167

Non-Parametric Correlations ............................................................................................................... 169

Non-Parametric Correlations in Sports Analytics ............................................................................... 169

Normal Likelihood Sampling Model .................................................................................................... 171

Position Importance in NCAA Football .............................................................................................. 171

Normality Tests ................................................................................................................................... 173

Tests of Normality as a Method of Sports Analytics .......................................................................... 173

Observation Chi-Square ....................................................................................................................... 175

Side-Out Success and Ways that Points Are Obtained in Women's College Volleyball .................. 175

Ordered Logit Model ............................................................................................................................ 177

Does Artificial Grass Affect the Competitive Balance in Major League Soccer? .............................. 177
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinary Least Squares Regression</td>
<td>179</td>
</tr>
<tr>
<td>Heterogeneity and Team Performance: Evaluating the Effect of Cultural Diversity in the World's Top Soccer League</td>
<td>179</td>
</tr>
<tr>
<td>PageRank and Elo</td>
<td>181</td>
</tr>
<tr>
<td>Ranking and Prediction of Collegiate Wrestling</td>
<td>181</td>
</tr>
<tr>
<td>Pairwise Function</td>
<td>183</td>
</tr>
<tr>
<td>Modeling the NCAA Basketball Tournament Selection Process Using a Decision Tree</td>
<td>183</td>
</tr>
<tr>
<td>Panel Fixed Effects Model</td>
<td>185</td>
</tr>
<tr>
<td>When Do Soccer Players Peak?</td>
<td>185</td>
</tr>
<tr>
<td>Play Sketching</td>
<td>187</td>
</tr>
<tr>
<td>In play sketching, coaches can sketch plays and instantly see how their opponent is likely to respond.</td>
<td>187</td>
</tr>
<tr>
<td>Bhostgusters: Real-time Interactive Play Sketching with Synthesized NBA Defenses</td>
<td>187</td>
</tr>
<tr>
<td>Player Injury Forecasting</td>
<td>189</td>
</tr>
<tr>
<td>Player Injury Forecasting as a Method of Sports Analytics</td>
<td>189</td>
</tr>
<tr>
<td>Poisson Factorization</td>
<td>191</td>
</tr>
<tr>
<td>A Generative Model for Predicting Outcomes in College Basketball</td>
<td>191</td>
</tr>
<tr>
<td>Populations and Samples</td>
<td>193</td>
</tr>
<tr>
<td>Populations and Samples in Empirical Research</td>
<td>193</td>
</tr>
<tr>
<td>Prediction Model</td>
<td>195</td>
</tr>
<tr>
<td>Going Inside the Inner Game: Predicting the Emotions of Professional Tennis Players from Match Broadcasts</td>
<td>195</td>
</tr>
<tr>
<td>Predictive Model Weights</td>
<td>197</td>
</tr>
<tr>
<td>Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights</td>
<td>197</td>
</tr>
<tr>
<td>Price of Anarchy</td>
<td>198</td>
</tr>
<tr>
<td>The Price of Anarchy in Basketball</td>
<td>198</td>
</tr>
<tr>
<td>Principal Component Analysis</td>
<td>200</td>
</tr>
<tr>
<td>A Scalable Framework for NBA Player and Team Comparisons Using Player Tracking Data</td>
<td>200</td>
</tr>
<tr>
<td>Probability</td>
<td>202</td>
</tr>
<tr>
<td>Measuring Excitement in Sport</td>
<td>202</td>
</tr>
<tr>
<td>Applying the Rules of Probability to Sports</td>
<td>203</td>
</tr>
<tr>
<td>Probabilistic Graphical Model</td>
<td>204</td>
</tr>
<tr>
<td>Graphical Model for Basketball Match Simulation</td>
<td>204</td>
</tr>
<tr>
<td>Probabilistic Physics-Based Model</td>
<td>206</td>
</tr>
</tbody>
</table>
Support Vector Machine ................................................................................................................. 268
A Team Recommendation System and Outcome Prediction for the Game of Cricket .................. 268
Survival Analysis Model .................................................................................................................. 270
New Applications of Survival Analysis Modeling: Examining Intercollegiate Athletic Donor Relationship Dissolution .................................................................................................................. 270
Team Strategy Model Framework .................................................................................................. 272
Invasion Team Sports: Strategy and Match Modeling ................................................................. 272
Text and Sentiment Analysis .......................................................................................................... 273
Text and Sentiment Analysis as a Method of Sports Analytics ...................................................... 273
Third-Order Polynomial Regression .............................................................................................. 275
Assessing the Offensive Productivity of NHL Players using In-Game Win Probabilities .......... 275
Three-Dimensional Markov Model ............................................................................................... 277
An Analysis of Curling Using a Three-Dimensional Markov Model .......................................... 277
Transformations .............................................................................................................................. 279
Using Transformations to Improve Measure of Team and Player Performance ....................... 279
Uplift and Persuasion Modeling ..................................................................................................... 280
Uplift and Persuasion Modeling as a Method of Sports Analytics .............................................. 280
Variables ......................................................................................................................................... 281
A Starting Point for Analyzing Basketball Statistics ........................................................................ 281
Variation .......................................................................................................................................... 283
Measuring the Variation in Sports Data .......................................................................................... 283
Comparing Between-Team and Within-Team Variation ............................................................. 283
Weighted Least Squares Regression Model .................................................................................... 285
An Improved Adjusted Plus-Minus Statistic for NHL Players ....................................................... 285
Weighted Plus/Minus ....................................................................................................................... 287
A Weighted Plus/Minus Metric for Individual Soccer Player Performance .................................. 287
Weighted Voronoi Diagram ............................................................................................................ 288
NBA Court Realty ............................................................................................................................ 288
Wilcoxon Signed Ranks and Kruskal Wallis H Tests .................................................................... 290
Wilcoxon Signed Ranks Tests and Kruskal Wallis H Tests as a Method of Sports Analytics ........ 290
Win Probability .............................................................................................................................. 292
The Problem with Win Probability ................................................................................................. 292
Win Probability Framework ............................................................................................................ 294
Estimating an NBA Player’s Impact on His Team’s Chances of Winning ................................................................. 294
Win Probability Matrix ................................................................................................................................................ 296
Using Cumulative Win Probabilities to Predict NCAA Basketball Performance ...................................................... 296
About Victor Holman

Victor Holman is a sports analytics and performance expert, management consultant and author. His thought leadership in business performance, sports performance and process improvement has been demonstrated in several books, frameworks and products he has published over the last 15 years. Victor has served as a performance expert, data and enterprise systems project manager and management consultant to over 100 organizations, including Fortune 500 companies and Federal Agencies. He’s been featured in major television, radio and print media outlets for his ability to help small businesses outperform their competition by applying strategies, frameworks and management tools that work for large, successful businesses. And Victor is applying these same concepts to sports management and sports analytics.

Victor believes that like small businesses, less talented sports programs can level the playing field by applying sports analytics methods that transcends their analytics programs beyond management decision making and into quantitative results delivered by the players. He also believes that to make a team function at optimal performance, it’s critical that game data is gathered, analyzed and translated into value that can be easily measured and delivered by its players. Victor helps teams figure out how to apply their data and existing analytics systems to help players measure value, increase IQ and create synergy using his Agile Sports Framework.

Victor authored the Agile Sports Guide, which outlines the step by step process for getting players to execute the coach’s strategy through an award system that measures player’s ability to adapt their games to the team strategy and measures their contribution to wins.

Victor developed the Agile Sports Analytics application, which provides the tools and drives the processes of the Agile Sports Framework.

Victor was recently recognized by Feedspot as the #2 sports analytics blog on the internet. His blog provides the latest strategies to help coaches, analysts and management improve the IQ, team dynamics & execution of their players. His blog also provides insight into all of the various statistical methods applied in sports analytics along with the key innovators and researchers who continue to push the sports analytics envelope. His blog posts are the foundation of this book.

Victor developed the Sports Analytics Maturity Model, which outlines five phases, seven key success areas, 26 best practices and 130 key processes for sports programs looking to gain a competitive edge across all aspects of the program.

And Victor has developed business and sports performance videos which have accumulated over 1 million views.

To learn more about how Victor can help your team leverage analytics to gain a competitive advantage, contact him at victor@agilesportsanalytics.com or (888) 861-8733.
About This Book

This book was created for non-mathematicians who love sports, are interested in sports analytics, but do not have the time, or desire to learn complex algorithms or processes within the various statistical methods. This book strives to provide a range of sports analytics methods in a manner that the average reader can understand their concepts and more importantly, how the concepts have been applied and leveraged in the sports world by some of today’s leading innovators in sports analytics. While the reader won’t walk away from this book understanding how to execute or program these sports analytics methods, they will walk away with countless new ideas for how analytics can be leveraged to gather new insight into data.

Sports analytics can be confusing, even for technical people. As you will find out in this book, different terms sometimes have the same meaning, for example the dependent variable, the response, the effect and the outcome are the same. Also there are numerous choices to make when planning to apply a single analytic method, such as which distribution to use, which test, which graph, which variation of the method, which control chart, etc. And each unique situation determines all of these factors. And to make it even more complex, experts sometimes disagree on which variation of the method is most appropriate. And to add to confusion, many sports analytics research involves applying multiple sports analytics methods to provide a complete analysis.

This book is unique in that it identifies numerous sports analytics methods, it’s alphabetically arranged, it outlines primary and secondary sports analytics methods applied in various research papers. Reading each of the research articles, and the sports analytics methods applied when performing the research will help the reader understand which statistical tool to use to approach their analytic challenges.
Introduction to Analytic Methods

Analytic methods are comprised of any method that takes data and organizes it in order to understand the meaning behind the information and make decisions based on that meaning. These methods look at data, throw away the garbage, and look for relationships among the remaining data even when it appears to be random. Sports outcomes do not follow a set of rules. You cannot guarantee who will win a game or which player will be the star of the day. There is always an element of uncertainty and analytic methods help put these uncertainties into perspective and look for patterns.

Analytic methods rely heavily on data analysis. Consequently, statistics is the base of analytics. Statistics range from the simple high school stats of mean, median, and mode to the higher-level stats of linear and polynomial regressions. Understanding the basics behind the statistics allows fans to gain a greater understanding and appreciation for the game they enjoy so much. When the analysts start talking during the game, fans will not get lost in the numbers, but be able to see past the numbers to the meaning behind them.

Statistics have two important roles within analytics. The first is their use in extracting relevant information and determining relationships between often seemingly unrelated numbers. This is done with the use of statistical models. Statistical models take all of the information and summarize it for our use. It is not necessary to sit and manually go through hundreds, if not thousands, of pieces of information looking for relationships. The statistical models are able to do this for us. It is important to remember that the conclusions the models make are not a guarantee. They discard the random events in order to develop the conclusions, but random events will continue to occur, affecting the outcomes.

A second role of statistics relates to the idea of uncertainty. Statistical models tell us what will happen if all things stay the same. However, things never do stay the same, so there is an element of randomness and error within the conclusions. Part of the model's role is giving some idea of the level of uncertainty involved. The conclusions are not guaranteed but they are still useful in determining what steps should be taken next to lead to the highest probability of winning.

Analytic methods have changed the face of sport over the years. Fans look at win/loss ratios and the statistics of their favorite players. Analysts' jobs are dependent on analytics; their job is taking the data from any aspect of sport and analyzing it in such a way that it allows them to draw conclusions. Which team is more likely to win the championship? Which teams should draft which players? Many sports teams now hire their own analysts. They want guidance in what strategies they should employ for their team to be the most effective. Which players are working at their peak performance and which are not? Are training methods adequate or do they need to some modification? What are the strengths and weakness of their opponents? What strategies give them the best chance of beating their opponents? For the answers, they look to the statistics.
Sports Analytics Maturity Model

With sports data increasing at a staggering pace, teams looking for a competitive advantage are realizing the benefits that business intelligence (BI), data discovery, and advanced analytics provide. Whether your team has evolved its analytics strategies to yield real-time actionable data for making quick decisions, or it still tracks performance using spreadsheets or simple BI reporting tools and dashboards, data analysis plays an essential role these days in performance and winning games. Victor Holman developed the Sports Analytics Maturity Model to help teams determine their sports analytics maturity when compared with other teams in their competitive level.

Sports Analytics Maturity Model Phases

*Discovery Phase* – Operational level reporting. Statistical or analytic tools have been purchased to meet an immediate need, but team lacks analytics expertise.

*Foundation Phase* – Advance reporting. Full commitment to analytics for player performance optimization and reaching team goals.

*Competitive Phase* – Strategic reporting. Committed to aligning analytics to other departments in order to reach organizational goals.
Predictive Phase – Predictive analytics reporting. Analytics consistently yield financial and performance gains across departments, and focus is on optimization and innovation.

Innovative Phase – Prescriptive analytics reporting. Machine intelligence. Transformational data. Data driven decision making. Focus is on prediction, optimization and innovation across the organization.

Sports Analytics Key Success Areas

People
Mature sports teams have players who understand analytics and how they can create value to help team reach goals. Their analytics staff are skilled data architects and analysts who are able to communicate analytics effectively throughout the organization. They understand their supporters and thrive to enhance the fan experience.

Technology
Mature teams use relevant tools and systems to enable high-quality data collection, intelligent reports, useful visualizations, and advanced analysis such as statistical modeling, predictive analytics and machine learning.

Data
Mature teams collect and manage data from multiple sources, both internal and external to the team, including unstructured data, geospatial data, etc. They document and apply data management and ownership policies, and establish analytics governance teams. And they use a range of technologies and techniques including data warehouses, Hadoop, and analytic appliances, data mining and statistical analysis, which form an analytics ecosystem.

Analytics
Mature teams strive to gain analytical insights through automation, and optimize decision making through analytics. They produce transformational data and focus on prediction, optimization and innovation across the organization.

Strategy
Mature teams define very clear team, player and organizational objectives that are measured by structured Key Performance Indicators (KPIs) designed to quantify success or failure. They integrate multiple data sources and actively leverage analytics for transformational change.
Profitability

Mature teams leverage analytics to maximize revenue operations and recurring revenues. They have insights on products/services that yield the highest profit margins. They can adjust faster to changing economic conditions and tailor their products and services for a diverse fan and sponsor base.

Process

Mature teams leverage formal frameworks to drive performance excellence both on the court and off. Their process includes regular inspection, swift adaptation when necessary and continuous improvements.
Allocative and Dynamic Efficiency

Allocative efficiency occurs when resources are so allocated that it is not possible to make the team better in one area without negatively impacting the team in another area.

Dynamic efficiency balances a short term approach to team improvement with a long term approach.

Optimal Strategy in Basketball

This is a review of the basketball strategy research applying allocative efficiency and dynamic efficiency techniques conducted by Brian Skinner and Matthew Goldman.

The end goal for any basketball game is to win the game. In order to maximize the chance of winning, teams need to evaluate the effectiveness of their strategies in influencing the outcome of the game.

One decision a team must face is when should shots be taken and which player should be taking the shots. When a player is faced with a shooting possibility, they must decide whether they should take the shot or pass the ball to a teammate. Which option will have the greater probability of ending up with the team scoring points? This model looks at the number of times a particular play is used during a game and how many points, on average, are earned by the play. If one particular play is used too often during a game its effectiveness decreases so coaches need to understand how to maximize the potential of each strategy by determining how often to use it within the course of a game.

Another factor in basketball strategy deals with time. A team's decision-making process is constrained by the shot clock. As the shot clock winds down a team often becomes more desperate to score, often taking lower quality shots, which reduces the expected outcome of the possession. Players must take into account the amount of time left on the shot clock when making decisions. Is it better to shoot now or pass the ball and wait for a better opportunity?

A third factor is the riskiness of a strategy. The ultimate goal is not to score points but to win the game. Before taking a shot, a player must look at the probability scoring. Are they likely to make points with the shot or not? Underdog teams can maximize their potential by pursuing risky strategies while stronger teams maximize their potential with a more conservative strategy. It is impossible to know exactly which plays a team will be utilizing ahead of time, however analysts and coaches can look at past games played by the opponents as an indicator of future performance and adapt their own strategy accordingly.

The greatest difficulty in determining optimal strategy is estimating the efficiency of different offensive tactics. This is combined with the fact that each offense play's efficiency is also affected by the defensive strategies of the opponent.

Determining a team's optimal strategy is based on maximizing their expected number of points and minimizing the opponent's expected number of points. Coaches can look at several possibilities to help determine optimal strategy. Some choices are to increase or decrease the number of 3-point shots taken, manipulating the time clock, or calling timeouts.
As advancements continue to be made in the field of basketball statistics, tools will be developed to help coaches and analysts more accurately determine the optimal strategy for individual teams within the league.

Analytics methods used in this research: allocative efficiency, dynamic efficiency
Backwards Selection Regression

Backwards selection regression is a statistical regression method of entry where all of the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation. Different from forward selection where predictors are added one at a time beginning with the predictor with the highest correlation with the dependent variable.

Source - https://www.statisticssolutions.com/selecion-process-for-multiple-regression/

Competition between Sports Hurts TV Ratings: How to Shift League Calendars to Optimize Viewership

This is a review of the sports TV ratings research conducted by Jim Pagels, applying backwards selection regression.

Sporting events attract large television audiences and sports programming rights fees have grown exponentially. In the fall of 2013, 34 out of the 35 highest-rated programs in the United States were NFL games. It is also important to remember that teams rely on the income received from TV rights.

While live events attract large audiences, when two are televised at the same time they split the viewership. The question is why the overlap between sports exists if the ratings and income are crucial to teams. This overlap can be resolved as the period from mid-June to the end of August is empty of professional sports, with the exception of the MLB, which garners the lowest television rankings among the four major sports.

What would happen if the NHL or NBA moved their playoffs so that they were not in direct competition with other? What if the World Series was played a time when it was not competing with the NFL?

TV ratings data was collected for nationally broadcast games between 2000-2014, amounting to 1,824 MLB, NBA, NHL, and NFL games. Games are considered to be overlapping if their start times are within 1.5 hours of each other. Overlaps between leagues and within leagues are all included.

A multiple linear regression was used to analyze the data by isolating the adjusted effects each variable has on total viewership. A backwards selection regression is then used to remove variables that are not significant indicators of ratings.

The results indicate that the NHL is hurt the most by the overlap, especially when competing against the NBA playoffs. The NHL also loses ratings when overlapping with the NBA and MLB regular seasons. Despite the widely held belief that MLB regular season and playoffs are hurt by overlap with the NFL, the statistics do not agree, actually indicating no impact at all.

To avoid this issue of losing ratings due to overlap the leagues could coordinate their schedules. One possible solution would be for the NHL to move its playoffs from April-June to July-August. A second solution is the MLA could start its season earlier to avoid overlap with college football. The NFL starting its season two months later in November after the MLB playoffs have finished is another solution.
While the data indicates that competition hurts TV ratings, it is highly likely that it overestimates viewership gains that would be garnered from eliminating overlap. It assumes that if a league moves its season to the summer months a larger viewership would be attained. However, people are typically busy with outdoor pursuits in the summer and would not likely watch televised sports like they do during the winter.

Analysts could use this information to look at the television schedule and ratings for games to see if there is a way to improve the schedule that would allow a larger viewership for each of the four major leagues in the United States.

Analytics methods used in this research: Multiple Linear Regression, Backwards Selection Regression
Bayesian Generalized Linear Model

Generalized Linear Model is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Shrinkage Estimation of NFL Field Goal Success Probabilities

This is a review of the Bayesian generalized linear model research conducted by Jason A. Osborne and Richard A. Levine.

Field goal kickers with the National Football League are released by teams at a relatively high rate. When kickers miss field goals viewed by management as costing the team the game, they are often released. This indicates that management believes that kickers have different levels of abilities; however, there is limited data supporting this assumption.

Kicks are assumed to be independent and the dependence of success probability on possible factors is modeled using generalized linear models. The possible factors investigated are the distance of the attempts and the identity of the kicker.

The dataset includes every field goal attempted, including those that resulted in a penalty. Information from the 1998-2014 seasons is overviewed, weeks within season, games within week, and plays within game. The final dataset consists of 17,104 field goal attempts made by more than 111 kickers. The outcome of the kick is classified as good or not good, with not good including missing, blocks, and fumbled holds.

Analysis shows that the complementary log-log (CLL) link exhibits a better fit than probit and logistic functions.

Estimates of an individual kicker’s success probability have large variances unless there a large data set for that kicker. Estimates with large variances can be reduced via shrinkage estimation, or shrinking it towards another estimator which has a lower variance. A Bayesian generalized linear model with a CLL link using empirical Bayes estimation can naturally shrink the estimates of probability success towards central values.

The estimation methodology can estimate the probability of success for an individual kicker at a given distance and is able to rank NFL kickers. Another application is determining whether different stadiums have an effect on the probability of making a field goal. The research indicates that yes, some stadiums are stingier in giving up field goals. Also of interest is the ability of each kicker to kick long-distance field goals. The outcomes provide a method for ranking kickers in regard to this ability. Kickers are ranked based on short yardage field goals, intermediate, and long-distance field goals.
The data analyzed covers a period of 16 years and indicates that kicker ability appears to be changing with improved accuracy and experience. Consequently, the probability of success in long-distance field goal attempts is increasing.

Coaches and management can use this information during contract negotiations to aide in determining appropriate salaries. It would also help when scouting prospects and looking at making trades. Rather than making what appear to hasty trades after a kicker misses an important field goal, managers could analyze whether another kicker has a higher ranking and is therefore more likely to have better success than the team's current kicker.

It must be noted that this method has a weakness in that the kicker is held accountable for all missed field goals. However, at times, the kicker is only partially at fault and at other times, the fault lies entirely with someone else. In the future, this model could be modified to account for these issues, creating more accurate rankings of the kickers.

Analytics methods used in this research: Shrinkage Estimation, Bayesian generalized linear model
Bayesian Hierarchical Latent Variable Models

A Bayesian Hierarchical model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies through a directed acyclic graph. Bayesian Hierarchical models subset themselves by containing three or more levels of random variables or use latent variables. One level uses within-unit analysis and another level for cross-unit analysis. Within-unit model describes individual respondents over time. Across-unit analysis is used to describe the diversity, or heterogeneity, of the units.

Source - https://static1.squarespace.com/static/53dc912de4b05be42a8acc61/t/53ddc610e4b0e8a580e87e6e/1407043988020/BayesRnD72514.pdf

A latent variable model is a statistical model that relates a set of observable variables (so-called manifest variables) to a set of latent variables. It is assumed that the responses on the indicators or manifest variables are the result of an individual's position on the latent variable(s), and that the manifest variables have nothing in common after controlling for the latent variable (local independence).


Eye on the Ball: The Relationship Between Sensorimotor Abilities and On-Field Performance in Professional Baseball

This is a review of the Bayesian hierarchical latent variable models conducted by Kyle Burris and Greg Appelbaum.

Many experts consider hitting a pitched baseball to be the hardest thing to do in sport. The athlete must determine the type of pitch and its trajectory, decide to swing the bat in coordination with the path of the ball - all done within milliseconds.

In this study, Sensory Station assessments from 252 professional baseball players were compared to game statistics to study the relationship between sensorimotor skills and baseball production. The assessments consisted of nine computerized sensorimotor tasks, each designed to evaluate a specific aspect of the player's visual-motor abilities. The nine tasks measured the player's ability to: make out small details on distant objects, detect the contrast between an object and its background, quickly and accurately measure the relative distance of an object, shift their attention and recognize peripheral targets, determine the number of targets both near and far way, remember and recreate visual patterns, respond to rapidly changing targets, and respond to "go" or "stop" stimuli, and finally the ability to quickly respond to simple visual stimulus.

Baseball production is measured using five different statistics. On-base percentage (OBP) measures how frequently a batter reaches base per plate appearance. Times on base includes hits, walks and hit-by-pitches but do not include the times a player reaches base due to defensive errors. Players need to be able to consistently hit the ball or draw a walk in order to have a high OBP. Walk percentage (BB%) is the
percentage of a batter’s plate appearances that result in the player being walked. Players with a high walk percentage are able to accurately determine whether a pitch is a ball or a strike. Strikeout percentage (K%) measures the number of times a batter strikes out. Slugging percentage (SLG) measures the power of the hitter...recognizing the fact that not all hits are equally valuable.

Analyzing the data showed that sensorimotor abilities are related to on-base, walk and strikeout percentages, but not to slugging percentage. These relationships are self-evident. Players with high on-base and walk percentages combined with low strikeout percentages must be able to determine the location of the pitch in the strike zone and decide whether to swing or not. Slugging percentage would deal more with strength, rather than sensorimotor abilities. Players with a high walk percentage typically scored high on tests relating to hand and eye coordination as well as response times. Players who have a low strikeout percentage tend to score high on the tests relating to measuring the relative distance of an object, recognizing peripheral targets, and responding to rapidly changing targets.

These relationships would indicate that sensorimotor screenings would be highly beneficial for player scouting. Screenings would help differentiate those players who are more likely to excel in the professional arena.

Analytics methods used in this research: Bayesian hierarchical latent variable models, on-base percentage, walk rate, strikeout rate, slugging percentage, correlation
Bayesian Hierarchical Model

Estimating an NBA Player's Impact on His Team's Chances of Winning

This is a review of the NBA research conducted by Sameer K. Deshpande and Abraham Wyner applying a Bayesian Hierarchical Model.

Some catchers in the MLB are known for their ability to frame pitches, catching them in such a way that the chance of it being called a strike is increased. Framing may have a large impact on a team's success, yet has been largely overlooked over the years. The question is to what extent does a catcher's framing skills affect the umpire's call compared to other factors like pitch location, count, and the pitcher himself.

Bayesian hierarchical model is developed to estimate each umpire's probability of calling a strike while adjusting for the pitch participants, pitch location, and contextual information such as the count. Then the effect each catcher has on an umpire's chance of calling a strike can be estimated. These estimated effects can be translated into average runs saved during a season.

Determining pitch location is done with a smoothed estimate of the historical log-odds of a called strike as an implicit parameterization of pitch location rather than specifying an explicit parameterization in terms of the horizontal and vertical components.

Plotting the spatial distribution of pitches clearly demonstrates that the spatial distribution of pitches varies a great deal with different combinations of batter and pitcher handedness. Consequently, a separate smoothed estimate of the historical log-odds of a called strike is created for each combination of batter and pitcher handedness.

General additive models for each combination of pitcher and batter handedness are created. The strike zone for right-handed pitchers closely resembles the one outlined in the rulebook. However, the strike zone moves several inches to the left for left-handed pitchers, especially noticeable between right-handed batters and left-handed pitchers.

The Bayesian hierarchical model is created through a succession of increasingly complex models. The model assumes that the pitch location is the only predictor of an umpire's ball/strike decision but also allows for umpire-to-umpire heterogeneity. The second model incorporates both catcher and count effects which are assumed to be constant across umpires. Constant pitcher and batter effects are added in to the third model. The fourth model included umpire-specific count and effects and the final model m

The predicted strike zones for the upcoming umpire, batter, pitcher and catcher combination is useful to coaches in helping their players understand what to expect regarding the strike zone. This will allow players to make wiser decisions regarding when not to swing the bat based on the expected strike zone. The framing ability of catchers can be compared. Catchers with strong framing capabilities have a positive impact on their team and thus should be valued more highly than other catchers. This is important information when heading into contract negotiations.
Analytics methods used in this research: Bayesian Hierarchical Model, Smoothed Estimate of the Historical Log-Odds, Generalized Additive Model
Bayesian Model and Simple OLS Regression

Bayesian statistics is a theory based on the Bayesian interpretation of probability where probability expresses a degree of belief in an event, which can change as new information is gathered, rather than a fixed value based upon frequency or propensity.

Simple OLS Regression is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variables.


Moral Hazard in Long-Term Guaranteed Contracts: Theory and Evidence from the NBA

This is a review of the NBA contracts research conducted by Arup Sen and J. Bradford Rice, applying Bayesian statistics and Simple OLS Regression.

In the world of sports, it always seems that players play better in the last year of their contract than they do during the earlier years. Teams then seem to offer these players huge contracts in order to retain them, only to see their performance fall as they start the new contract, a situation that is especially frustrating for the fans of the game. Unlike many other sports, basketball contracts are of a simple, fixed-wage design and rarely include incentive clauses. Is this pattern of increased effort in the last year of a player's contract a true phenomenon or simply a figment of the fans' imaginations?

Fixed-effects estimates indicate that players' efficiency score in the final year of a multi-year deal is significantly higher than in the prior year.

The interaction between the player and team is modeled as a 3-period principal-agent game. The principal is risk neutral and the agent is risk averse. During the 3-period game, the player chooses what effort to expel in each period even though the salary for that period was previously negotiated. The outcome is achieved at the end of every period and depends on the player's inherent ability, the level of effort exerted and a random component. The player, as well as the team, has some belief regarding the player's ability and this is updated in a Bayesian fashion by the team at the conclusion of every period. As the player is earning a guaranteed wage for the period, the only incentive to put in the effort is pride and the ability to affect future wages.

Contracts that cover the first 2 periods of a player's life are analyzed. Results indicate that players signing such a contract will increase their effort level over the life span of the contract. While one period contracts are more beneficial for the team, the agent's risk aversion steps into play. The players value security and will make some wage concessions per period in order to sign a longer guaranteed contract. This concession allows the team to compensate for any loss of effort. Finally, more experienced players demonstrate a smaller variety of effort levels over the course of their contract than less experienced...
players do. This is a result of the fact that the ability of more experienced players is better known by the team, which can result in negative consequences if the player is putting out less than optimal effort.

Testing the hypothesis that player effort and performance will improve as the expiry date of the current contract nears is tested using a simple OLS regression of performance on the years remaining in the contract. The dataset includes information on 657 NBA players including their contract terms, annual performance across several dimensions, information on team performance and physical characteristics. Effort decreases the most right after a new contract has been signed.

This information enables analysts and coaches to determine just what the decrease in effort after the beginning of a contract actually is in order to determine terms for future contracts. Coaches, being aware of this phenomenon, can better strategically plan the team for the year, knowing which players are nearing the end of their contract and therefore will put in maximum effort versus those just starting a new contract. Players nearing the end of the contract will be more effective players for the team, assuming a similar level of ability.

Are teams unaware of this phenomenon? No, of course not, teams anticipate the drop of effort level for a player and contract terms are decided with this in mind.

Analytics methods used in this research: Bayesian Model, Simple OLS Regression
Bayesian Multinomial Model

Bayesian multinomial logistic regression is a statistical method applied to model unordered categorical variables. The dependent variable may be in the format of either character strings or integer values. The model is estimated via a random walk Metropolis algorithm or a slice sampler. See for the maximum-likelihood estimation of this model.

Source - http://docs.zeligproject.org/articles/zelig_mlogitbayes.html

Exploring the Potential of the Plus/Minus in NCAA Women's Volleyball via the Recovery of Court Presence Information

This is a review of the NCAA volleyball research conducted by Zachary Hass and Bruce A. Craig applying Bayesian multinomial models.

A stat keeper records information during NCAA women's volleyball games for compilation into summary stats. However, not every action a player makes is recorded and not every recorded action is converted into a traditional statistic. This study was put together because of this issue in order to create a plus-minus metric to help analyze player's actions.

One basic problem of applying a plus-minus statistic to volleyball regards the position of Libero. The Libero is not bound by the same rules as the other players. They have more freedom to move in and out of the game, replacing players in different situations and their movement is not recorded. Due to this, it is difficult to tell which players are on the court at each point in the game.

It is necessary to make an inference regarding whether or not the Libero is on the court and which player they are replacing. This is calculated knowing which seven players might be on the court and the service order. Beginning with the starting rotation, which players are on the court is determined by reviewing the game play by play, which contains some hints as to who is on the court. However, many plays do not contain complete data. In order to determine the missing information a 4988 x 16 matrix is constructed where each row stands for an individual play and the columns include characteristics of the play.

Determining which three players are in the back row at every point of the game is extremely difficult. Therefore, a Bayesian multinomial model is used to predict which player is not on the court. The resulting information allows the construction of a plus-minus rating.

To check the accuracy of the model's predictions the results are compared to actual film coverage of two games. The plus-minus ratings were calculated for the players involved in the two games using the model-based approach along with three other approaches for comparison purposes. The first alternative splits the six spots on the court evenly among the seven players. The second alternative adds in tracking player rotation while the third is a more complete estimate but does not model court presence. The Bayesian model was by far the superior approach.
This study has some limitations, as it is restricted to studying only one team, which suffered relatively few injuries throughout season, thus lessening the variation possible in another year or by another team.

A reliable plus/minus statistic allows coaches and analysts to compare players using the same scale. This comparison could be useful when determining awards at the end of the season or when choosing players for the national team. Professional teams could utilize the statistic when looking at possible additions to their teams.

Analytics methods used in this research: Matrix, Bayesian Multinomial Model, Plus/Minus Rating
Binary Logic Model

In statistics, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1".

Matchup Models for the Probability of a Ground and a Ground Ball Hit

This is a review of the MLB research conducted by Glenn Healey, applying a binary logic model. Approximately 32% of batter/pitcher matchups in the MLB in 2014 resulted in a ground ball. The ground ball has a much lower expected value than the average run value, which makes it beneficial for the pitcher and his team. The probability of a pitch resulting in a ground ball is dependent on both the pitcher and the batter.

A binary logit model is useful in characterizing the probability of a result in a binary experiment as a function of a set of variables so this model is utilized in the research. Specifically the log5 model is used. In this instance, the application incorporates the league, batter, and pitcher ground ball rates.

The next step is to model the probability of a ground ball. To do this it is necessary to determine a set of descriptors for batters and pitchers which can used as variables in the model. Batter and pitcher strikeout and ground rate will be used in this instance. Data is taken from all plate appearances from 2003 to 2014 with the exception of bunts, intentional walks and when the batter is a pitcher. A logistic regression model is applied to the data.

The next step is modeling the probability of a ground ball hit. Again, bunts, intentional walks, and pitchers as batters are excluded from the data. Results show that as a group, right-handed batters have a larger league batting average on ground balls, which is logical as right-handed batters hit more ground balls to the left side of the field, which, in turn, requires a longer throw to first base. The speed and direction of batted balls also affect the probability of a ground ball. Another factor is the opponent's defense, as infielders with a stronger throwing arm will turn more ground balls into outs. A logistic regression is again applied to this data to recover a logit model.

A relatively small sample can be used to determine the probability of a ground ball using pitcher and batter descriptors as variables. The resulting model not only determines batter and pitcher ground ball rates, but also looks at the interaction between batter and pitcher strikeout rates. This creates a method for determining whether a batter or pitcher is favored in regards to both ground balls and strikeouts. Further, the probability that a ground ball becomes a hit is dependent on the pitcher's ground ball and strikeout rates.

Being able to predict the distribution of outcomes for a matchup between a batter and a pitcher would help coaches put together their line up for the game. Having the ability to compare individual batters and pitchers to determine which is favored in terms of ground balls and strikeouts provides coaches with additional insight into which batters are best suited to face a pitcher.
The model could be refined in the future to include batted ball speeds and launch angles, which would provide additional details.

Analytics methods used in this research: Binary Logic Model, Log5 Model,
Bipartite Graph Algorithms

In the mathematical field of graph theory, a bipartite graph (or bigraph) is a graph whose vertices can be divided into two disjoint and independent sets \( U \) and \( V \) such that every edge connects a vertex in \( U \) to one in \( V \). Vertex sets \( U \) and \( V \) are usually called the parts of the graph. Equivalently, a bipartite graph is a graph that does not contain any odd-length cycles.


Player Centrality on NBA Teams

This is a review of the NBA research using bipartite graph algorithms conducted by Sohum Misra.

Basketball is ever growing in its popularity. Teams look for new techniques to help them gain advantages over their competitors. One technique increasing in its use is advanced statistics. These statistics help teams determine the intangible value of an individual player.

Graph theory is one technique that can be used. A bipartite graph was constructed using data from five NBA seasons - from the 2012-13 season through to the 2016-17 season. It included 812 players and 36045 line-ups. A bipartite graph is one that takes two graphs and combines them into one. They are typically used to determine relationships between two different classes of objects. In this case, the graph looked at players and five-man units that play together. Six components were used to construct the graph including assist to turnover ratio, field-goal percentage, offensive rebound percentage, steals to possessions ratio, opponent field goal miss percentage, and defensive rebound percentage.

It was determined that this model had an inherent flaw. Players who played for the same team over the five years were ranked higher than players who played for multiple teams. Some of the league’s top ranked players were not listed in the top rankings as they had played on more than one team. A player's importance to one team is not necessarily the same as their importance on another team as their effectiveness is related to the other players on the team.

In order to deal with this a second graph was constructed. In this graph, the average tenure on a team was calculated. All players who played with a team for less than the average had their weightings boosted and all players who played for a team for more than the average had their weightings decreased. The level of the boost or decrease was varied per player depending on how big the difference was between their length of time with the team and the average tenure. This helps put all players on a level playing field when looking at their importance to a team. However, the results were only marginally better than those of the first model.

Consequently, it was decided to model each team on a separate network. This resulted in a more accurate ranking of the players, but the results were still mediocre (57 to 67% accurate). The better defenders and offenders on a team were ranked higher than their teammates. However, comparing
players across teams was not effective. Players playing on a balanced team would have a lower ranking overall while players on a less balanced team would have inflated rankings.

These findings outline the idea that analysts and coaches need to be cautious when looking at making trades. They need to understand that a player’s effectiveness on one team will not necessarily correspond to that player’s possible effectiveness on a new team. However, coaches can use this model to look at how their players rank against each other on their team in order to determine which players are being used most effectively.
Bland Altman Plot Analysis

Bland Altman Plot Analysis in Sports Analytics

This chapter touches on Bland-Altman plot analysis and its application in the world of sports. Bland-Altman plot analysis is a graphical method that involves plotting the differences of two techniques against their mean or average. The differences can also be plotted against one of the two methods if the method has been recognized to be a reference or gold-standard method.

Bland and Altman, in their experiments, discovered two similar methods that can be used for assessing agreement without assuming that the differences between the recorded test and retest scores are normally distributed.

This method uses a plot, known as the Bland and Altman plot and it is constructed as a simple scatter diagram on an XY graph. The Y-axis bears the difference between the test and retest scores (test - retest) while the means of the test and retest scores ((test + retest)/2) are plotted on the X-axis. The non-parametric methods require that the performance analyst calculate the values outside which a certain proportion of the observations fall.

The values can be obtained by determining the differences between the test and retest scores, and then recording the range of values that is left after a percentage (2.5%) of the ‘sample’ data is removed from each end of the frequency distribution.

Performance assessment using the Bland-Altman method identifies critical events (called performance indicators) in individual or team sports. Such events are thought to be a key determining factor to success in those particular sports. Many of these performance indicators are discrete, categorical events (counts or frequencies), such as the number of winners and errors in tennis or badminton, number of shots on goal in soccer etc. Others may be ratios of variables that often represent the characteristics elements of efficiency of performance.

The Bland-Altman method requires that analysts treat individual sport performances as individual variables, and this is usually done by reporting the reliability coefficients for each of the performance indicators of interest. This process is important especially where the performance indicator used is in a relatively rare event in a given sport (e.g. the number of shots on goal in soccer).

Bland-Altman method also recommends that sports analysts use a minimum sample size of $n = 50$ during studies that are aimed at assessing agreement with the parametric 95% limits of agreement method (that is, those predicated on normal distribution theory). Also, generating larger sports samples aids sports analysts performing ‘test-retest’ reliability studies on the data.

Conversely, when the performance indicator is a relatively common event, such as the number of passes made in a game’s activity, the analyst can perform simple techniques like dividing a single match into two-minute periods and entering the number of passes performed in each time period into the analysis system. This way, a sample of at least 40 time periods for a rugby match (80 minutes duration) or 45 time periods in a soccer match can be generated, and the reliability of the data entry (between the test and the retest) accessed.
In conclusion, Bland Altman plot helps sports analysts access individual player performance as well as the reliability of sports data entry.
Bonferroni Adjusted Comparisons

The Bonferroni correction is an adjustment made to P values when several dependent or independent statistical tests are being performed simultaneously on a single data set. To perform a Bonferroni correction, divide the critical P value (α) by the number of comparisons being made. For example, if 10 hypotheses are being tested, the new critical P value would be α/10. The statistical power of the study is then calculated based on this modified P value.

The Bonferroni correction is used to reduce the chances of obtaining false-positive results (type I errors) when multiple pair wise tests are performed on a single set of data. Put simply, the probability of identifying at least one significant result due to chance increases as more hypotheses are tested.

Source - https://www.aaos.org/AAOSNow/2012/Apr/research/research7/?ssopc=1

A Closer Look at the Prevalence of Time Rule Violations and the Inter-Point Time in Men's Gland Slam Tennis

This is a review of the tennis analytics research conducted by Otto Kolbinger, Simon Grossmann, and Martin Lames, applying descriptive statistics, regression analysis, and Bonferroni adjusted comparisons.

Paragraph 29a of the official International Tennis Federation's Rules of Tennis states that 20 seconds is the time limit allowed players between points. It appears to all watching a game that this rule is often broken by the players but not enforced by the umpires. This study looks at the rule violations and the umpire's response. The influence of various factors regarding the time between points are also examined. Those factors include physiological, illustrated by the duration of the previous rally; physiological, illustrated by the number of played sets and service games; and tactical, illustrated by the current scoring streak and importance of the point.

Data from 21 matches at the 2016 Australian Open Men's Singles tournament including 6231 rallies were analyzed. Data included serving player, receiving player, current score, whether it was a first or second serve, the number of strokes in the rally, the winning of the rally and the time between points. Any unusual events were also recorded like a player changing equipment, umpire overruling the line judge, umpire demanding that the audience be quite, other player appeals, and warnings. The analysis was limited to first serves and excluded any points that included any of the aforementioned unusual events.

First, the frequency of rule violations is determined using descriptive statistics and then two regression models will be used to determine which factors influence the duration of interruptions between points. The average time between points was higher than the limit of 20 seconds. This occurred 2034 times and only two were penalized by the umpire.

Bonferroni adjusted comparisons regarding the duration between points demonstrated 11 significant differences, especially between early and late games of a set with the mean time increasing later in the game. The average time was higher if the opponent won the last point than if the server won the last
point. Time between tiebreaks was significantly higher than time between regular plays. The duration of the previous rally also had a significant impact on the time taken between points.

The results indicate that players will use extra time between points in order to recover, to disrupt the rhythm of their opponent, and to improve their focus for important point rallies.

The two warnings issued by umpires were not for the times that exceeded the limit the most, indicating that not only is the rule not judged as written, it is not judged fairly.

Analysts can use this to compare players regarding the average time they take between points. Players could be ranked according to the time and then analysts could look for any correlations between time and player ability or any other factors.

Either the rulebook needs to be revised or umpires need stricter guidelines regarding when to call time violations or some combination of both in order to ensure that all players receive the same treatment.

Analytics methods used in this research: Descriptive Statistics, Regression Model, Bonferroni
**Bookmaker Consensus Model**

Bookmakers odds are an easily available source of ‘prospective’ information that is often employed for forecasting the outcome of sports events. In order to investigate the statistical properties of bookmakers odds from a variety of bookmakers for a number of different potential outcomes of a sports event, a class of mixed-effects models is explored, providing information about both consensus and (dis)agreement across bookmakers. Bookmaker model selection yields a simple and intuitive model with team-specific means for capturing consensus and team-specific standard deviations reflecting agreement across bookmakers. The resulting consensus forecast performs well in practice, exhibiting high correlation with the actual tournament outcome. Furthermore, the agreement across the bookmakers can be shown to be strongly correlated with the predicted consensus and can thus be incorporated in a more parsimonious model for agreement while preserving the same consensus fit. >

Source

https://www.researchgate.net/publication/228667032_Bookmaker_Consensus_and_Agreement_for_the_UEFA_Champions_League_200809

**Searching for the GOAT of Tennis Win Prediction**

This is a review of the tennis prediction research conducted by Stephanie Ann Kovalchik, applying bookmakers consensus model.

Where there are sports, there are people predicting the outcome. Trying to figure out who will win fascinates people of all ages all around the world. This includes fans of the sport of tennis.

Models that predict outcomes of singles matches in professional tennis are plentiful and the purpose of this research is to study 11 of those models for their accuracy and discriminatory power. Testing the models was accomplished using the dataset from the 2014 Association of Tennis Professionals Tour. Also investigated are the differences between performance on different surfaces and different tournament levels. The aim is to identify the major components of win ability in professional tennis as well as to determine how current models can be improved.

The eleven models fall into three categories: regression, point-based and paired comparison models. Regression models directly model the winner of a match and typically are based on the probit family and include player rankings as a predictor. Point-based models determine the probability of winning on serve and from there compute a prediction for the match outcome using an algebraic formula, which assumes that points are independent and evenly distributed. Paired comparison models weigh the possibilities of different options. A bookmaker consensus model was included as a standard of reference for predictive performance.

To maintain a fair comparison all models were analyzed using one-year’ worth of data, including 2395 ATP singles matches played during the 2014 season. Models were evaluated based on: prediction accuracy, calibration, log-loss, and discrimination. Prediction accuracy is the percentage of correct
predictions. Calibration looks at the expected wins across a number of matches. Log-loss measures whether the player with the higher ranking won and if the corresponding prediction was correct. Discrimination is determined by measuring the mean prediction for matches won by higher-ranked players minus the mean prediction for the matches they lost.

In terms of accuracy, regression and point-based models performed similarly, with the regression models that included player ranking being the most accurate. Adding in additional predictors did not improve the accuracy of the models. The paired comparison models showed improved accuracy as the amount of data increased. Overall, the bookmaker consensus model was the most accurate in its predictions.

In terms of calibration, several models indicated a bias in that the models tended to underestimate the higher-ranked player's likelihood of winning, predicting more upsets than actually occurred during the year analyzed.

The bookmaker consensus model also generated the lowest log-loss of all models, which indicates that it made fewer overconfident predictions. The point-based models predicted the greatest amount of upsets.

Again, the bookmaker consensus model demonstrated the best discriminatory ability. Regression-based models were the worst in this area.

Accuracy differences were found as related to player rank, tournament level, and type of surface. The greatest differences were found between matches which included a top 30 player, and matches that did not. All models were significantly more accurate regarding their predictions of matches involving a top 30 player. Predictions for Grand Slam matches were also more accurate than other tournaments. Finally, predictions for grass and hard-court tournaments were more accurate than those for tournaments using clay courts.

The only model that came close to the bookmaker consensus model in accuracy was Elo, which is the only model that considers a player's past wins and losses. Recent performances are weighted more heavily as well as wins against stronger opponents.

This information provides useful data for analysts who are always trying to generate the best predictions. It can also point analysts to the opportunity to create an even better model using the strengths of the various models.

Analytics methods used in this research: Regression Model, Point-based Model, Paired Comparison Model, Bookmaker Consensus Model
Bootstrap Simulation

In statistics, bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods. Generally, it falls in the broader class of resampling methods.


Improving Fairness in Match Play Golf through Enhanced Handicap Allocation

This is a review of the bootstrap simulation and distribution graph research conducted by Timothy Chan, David Madras, and Martin Puterman.

There are two categories of golf competitions. In a stroke play competition the player with the fewest strokes wins the game. In match play, the player who wins the greatest number of holes wins the game. As each golfer is unique, having a unique set experience and expertise, a handicap system has been devised in order to put players at a more even level with each other, in order to create fairer competitions.

A handicap indicates how well a person plays compared to par when they are playing their best. At the end of a competition that uses handicaps, the players' handicap is subtracted from their score and the person with the lowest score at that point wins the match. In a match play game the process is more complicated in that the strokes a player 'gives' their opponent depends not only upon their own handicap but also on the ranking of each hole.

This research looks at whether the current handicap system allows for a truly fair game or if another approach could be more effective at providing each player with the same probability of winning the match. Data was collected from four casual stroke play tournaments held at the Shaughnessy Golf and Country Club in Vancouver, Canada including the player's handicap and their scores on each of the 18 holes.

The data was graphed which indicated that the players’ handicaps and performance were equivalent. The graph also indicates that the net score for those with middle to high handicaps were more varied than the net score for those with a low handicap.

In order to calculate the over-all win percentages a bootstrap simulation approach was used. The 73,512 matches were sampled with replacement a total of 10,000 times. Following this, the win percentage was determined at each handicap differential using the same method.

Theoretically, the current handicap system should give each player a 50% chance of winning. This simulation calculated that the better player actually won 53% of the time indicating a bias in the current system towards the player with the lower handicap.
Three possible changes to the system are proposed: create three new hole rankings, change the holes to which the handicap differential is applied, and vary the number of extra strokes given to the weaker player. Giving an additional 0.5 strokes to the weaker player created the fairest outcome of all the scenarios.

This information provides analysts with an additional tool in determining likely outcomes of different matches, as the current system appears to favor the better player. Golf courses and tournaments could use this information to develop a handicap system that is fair to all players, allowing for an optimal level of competition.

Analytics methods used in this research: Distribution Graph, Bootstrap Simulation
Boruta Algorithm

Boruta is an all relevant feature selection wrapper algorithm, capable of working with any classification method that output variable importance measure (VIM); by default, Boruta uses Random Forest. The method performs a top-down search for relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies, and progressively eliminating irrelevant features to stabilize that test.

Source - https://www.rdocumentation.org/packages/Boruta/versions/6.0.0/topics/Boruta

An Investigation of Three-Point Shooting through an Analysis of NBA Player Tracking Data

This is a review of the thesis by Bradley A. Sliz applying the Boruta Algorithm to analyze three point shooting.

Player tracking data in basketball allows analysts to refine more advanced statistics than ever before. These advanced statistics help teams determine how different strategies affect the outcome of the game. One strategy that can be analyzed is three-point shooting.

A Boruta algorithm is used to measure the importance of the different variables, as it is a more current and advanced method.

In the Boruta algorithm, three-point shooting is analyzed from two different perspectives. The first looks at how team strategies and actions affect the success of three-point shots. The data is input into the model to determine what impact different offensive plays have on the outcome of a three-point shot.

The second perspective looks at the impact the individual player has on the success. It looks at the expected number of three-point shots for a player versus the actual number of shots they attempt during a game. This assists in determining whether players are reacting to the game in an expected or unexpected way. It also defines which players take three-point shots more often. The number of successes for each player is also taken into account.

Analysis of the team results shows that teamwork is a more accurate predictor of success than the play of individual players. The distance between the shooter and the nearest defender is the strongest predictor of shot success. While the talent of the player has some effect on the outcome, other aspects such as the movement of the ball and the timing of the shot actually have a greater affect.

Analysis of the player results allows comparison between players who take three-point shots more often and less often than the average and the effectiveness of each player.

Analysts could use this information in scouting future talent. It would help in analyzing which players might be a better fit for their team and the optimal way to utilize each player's talent.

Coaches can use this information to determine which of their players is taking three-points less effectively and have them focus on alternative strategies. Conversely, they can identify players who are
under-utilizing the three-point shot who need to be encouraged to do so on a more regular basis in order to enhance the overall effectiveness of the team. Team strategies can be adjusted to take advantage of the effectiveness and ineffectiveness of the different players. Coaches can also look at the three-point shooting tendencies of their opponents. Strategies are then developed to attempt to force those less effective players to take more three-point shots. At the same time, coaches can determine ways to prevent the more effective players from having the chance to shoot a three-pointer.
Brownian Motion Process

A standard Brownian (or a standard Wiener process) is a stochastic process \( \{W_t\}_{t \geq 0} \) (that is, a family of random variables \( W_t \), indexed by nonnegative real numbers \( t \), defined on a common probability space \( (\Omega, F, P) \)) with the following properties:

1. \( W_0 = 0 \).
2. With probability 1, the function \( t \rightarrow W_t \) is continuous in \( t \).
3. The process \( \{W_t\}_{t \geq 0} \) has stationary, independent increments.
4. The increment \( W_{t+s} - W_s \) has the \( \text{NORMAL}(0, t) \) distribution.

Source - https://en.wikipedia.org/wiki/Wiener_process

The Implied Volatility of a Sports Game

This is a review of the sports research conducted by Nicholas G. Polson and Hal S. Stern applying a distributional model and Brownian motion process.

Why do we like to watch sporting events? One main reason is that we can never be sure of the outcome of the event until the end. There is always a chance that the underdog will win and the favorite will lose. Uncertainty and sports go together hand in hand.

Uncertainty of a game's outcome is assessed using the betting point spread and the probability of one team winning implied by the betting odds. The volatility of the outcome is measured by looking at the development of the game score, beginning with a distributional model for the evolution of the outcome in a sports game. The distribution of the lead of one team over the other is specified as a Brownian motion process.

A method of updating implied volatility throughout the course of the game is created using real-time changes in bettor's assessments. With on-line betting there is an almost continuous information trail available to assess the implied expectation of the probability of one team winning at any point of the game. Uncertainty of the outcome is measured as the variation associated with the final score of the game. Implied volatility is defined as the outcome for the entire game or for the remaining part of the game. The market-implied volatility is determined by bettors' or analysts' assessments of the game outcome. The expected margin of victory is the point spread while the probability that a team wins is based on money-line odds. Finally, the implied volatility is a market-based assessment of the level of uncertainty seen in the difference between the scores of the two teams involved in the game.

Betting and prediction markets have shown that teams who are experiencing a long losing streak and consequently are considered extreme underdogs tend to be underpriced by the market. Volatility could also be measured using index betting spreads in which bettors provide assessments of over-under lines. These lines are used to make wagers regarding the total number of points scored in a game.
The point spread and money-line odds do not have a direct correlation, as it is very possible that two games involving heavily favored teams have the same money-line odds but the point spread is very different. It is possible this is due to the idea that the market has an expectation that the volatility is much higher for the game with larger point spreads.

Looking at point spreads and money-line odds provides analysts with another measure on which to compare games. Why are some games considered to be so much more volatile than other games, even when both games have a highly favored team considered to be far superior than their opponent? This research focuses on applying the model to football games. Analysts could then apply this model to other sports to see if the results are consistent across sports or not.

Analytics methods used in this research: Distributional Model, Brownian Motion Process
Chapman- Kolmogorov Equations

In mathematics, specifically in the theory of Markovian stochastic processes in probability theory, the Chapman–Kolmogorov equation is an identity relating the joint probability distributions of different sets of coordinates on a stochastic process.


Evaluating NBA End-of-Game Decision-Making

This is a review of the NBA decision making research conducted by Patrick McFarlane, applying Chapman- Kolmogorov Equations.

In the game of basketball, a team's strategy towards the end of the game often differs from that employed earlier in the game. Teams that are trailing may intentionally foul in order to get the ball back quickly while teams that are leading by three points may intentionally foul rather than giving the opposing team the possibility of scoring a three-point shot.

An end-of-game tactics metric (ETM) was developed to look at the decisions made within the last three minutes of a game, specifically decision related to shooting a two-point field goal, a three-point field goal and intentionally fouling. ETM determines the difference between the win probability of the optimal tactic and the actual tactic used by a team.

Team statistics are used to determine the most advantageous tactic for a possession. The information provided is used to determine the win probability as well as the probability of success for the various offensive and defensive tactics. These form various chains of post-possession win probabilities from which to choose. Nonlinear functions, based on score differential and time remaining, are features of the model.

This win probability model is based on time remaining, score differential, possession, and point spread and is created with a logistic regression. The model looks at all possible scenarios, or states, to determine the win probability for both teams.

The win probability model is used as the basis for the end-of-game tactics metric as each team has a win probability and decisions to make each time they gain possession of the ball. The decisions analyzed in this study are shooting a two-point or three-point field goal for the offensive team and for the defense the decision regarding whether or not to intentionally foul. ETM is useful as it is able to calculate the best decision and compare the team's actual decision with the optimal one. Following the decision, Chapman-Kolmogorov equations calculate the new win probability as the sum of the probability of all the possible outcomes of the decision multiplied by the winning probability after those outcomes.

To determine aggregate ETM results play-by-play data from the 2015-2016 NBA season were analyzed. The winning percentage in close games, those ending with less than a five point differential or in overtime, versus the average difference between the ETM of a team and their opponent are plotted and analyzed with a line of best fit created through a least squares process.
The resulting negative correlation coefficient shows that a higher win probability in close games is related to a more negative ETM difference. The data demonstrates a wide variability, indicating that decision making in the last minutes of a game depends on a wide variety of factors.

Results also indicate that the decision of the leading team to intentionally foul, forcing two free throw attempts rather than allowing for the possibility of a three-pointer becomes optimal with nine seconds or less remaining in the game. Earlier it is not the optimal tactic. Intentionally fouling when losing is optimal only for a very small score differential.

Coaches and analysts can use this to determine the amount of win probability a team will give up if they chose a less than optimal strategy. This provides the knowledge to better determine if the less than optimal tactic such as an intentional foul provides the best probability of winning the game.

ETM could be expanded in the future to consider specific players or plays rather than the aggregate.

Analytics methods used in this research: End-of-Game Tactics Metric, Logistic Regression, Chapman-Kolmogorov Equations, and Least Squares
Clustering

Clustering, in the context of databases, refers to the ability of several servers or instances to connect to a single database. An instance is the collection of memory and processes that interacts with a database, which is the set of physical files that actually store data.

Source - https://www.techopedia.com/definition/17/clustering-databases

Big 2's and Big 3's: Analyzing How a Team's Best Players Complement Each Other

This is a review of the clustering and regression research conducted by Robert Ayer.

Why do some teams appear to underperform while others seem to over perform? Looking at a team with several star players leads to high expectations, yet the team does not live up to these expectations. Contrarily, teams that lack star power are able to play and win beyond what any stat indicates they should be able to do. One explanation for this is how well the players fit together and complement each other.

This research in clustering and regression examines this phenomenon in order to analyze various types of combinations and which combinations lead to more wins. The research assumes that success in the NBA is primarily dependent on three factors: the talent of the players, the skills of the coaches, and how well the team players fit together. Player talent and coaching skill are easy to analyze. Teams with more player talent and great coaching should win more games. As this is not always the case the team fit factor must be incorporated.

Players from across the league are grouped together based on their similarity to each other. Variables that are looked at include offensive and defensive rebounds per game; assists, steals, blocks, and turnovers per game; personal fouls per game; field goal attempts and successes per game; free throw attempts and successes; and 3-pointers attempted and made per game. Players may appear in more than one group but ideally would not appear in more than two.

Talent tends to be more subjective so how do you quantify it? In this approach each player was evaluated based on their efficiency during the season. This was done by adding up points, rebounds, assists, steals, blocks, field goals made, free throws made and turnovers and then subtracting from that total field goals attempted and free throws attempted. The number of minutes played was used to weight these efficiencies. The outcome then allows us to rank the players in terms of talent which then provides an estimate of the total talent on a team.

Some conclusions from this clustering and regression research were what you would expect but others were more unexpected. First, how players are combined together has a significant impact on the number of wins. Players belonging to the same group do not work well together. They are bringing the same talents to the table and therefore do not create a well-rounded, effective pairing. Talented players from different groups tend to mesh better together and be more effective. One unexpected finding is
that teams with two high-scoring guards do appear to have a good fit as they exceed the expected outcome.

Analysts could take the efficiency stats of the various players and work out combinations that would provide the ultimate "fantasy" team. Teams could use this research to determine which available players would best complement the players already on the team, forming the best fit and, thus, hopefully the best chance of winning games.

Analytics methods used in this research: Clustering and regression
Coaching Assistance Models

Coaching assistance models are methods and/or tools which allow coaches to adjust analytic algorithms in order to bridge machine learning and the coaching staff’s expert judgement.

Coaching Assistance Models as a Method of Sports Analytics

This chapter focuses on coaching assistance models, a sports analytics method that aids coaches and analysts in determining the best course of action for a team, especially in basketball.

The coaching assistance model involves the use of ball screens (commonly known as pick and rolls). In ball screens, an offensive player helps the ball-handler elude his/her defender and make a shot or pass by getting in between the ball handler and defender and creating space for the ball handler. When the offensive player sets a screen for a teammate, the player here is referred to as the screener.

This model uses the SportsVU position data system that helps identifies ball screens, the defensive pattern used and the possession outcome of the screen.

For making ball screens, four players are involved- the ball handler, the screener, the ball handler defender, and the screener defender.

In using this model, defensive strategies can fall into any of these categories:

- Over: the ball handler defender is between the ball handler and the screener.
- Switch: when the screener defender switches place with the ball handler defender.
- Under: when the ball handler defender player is not with the ball handler and the screener.
- Trap/double-teaming – when both defenders double-team (blocks) the ball handler.

The team defensive strategy to use is chosen based on the offensive skill of the ball handler as well as his shooting ability. SportsVU position data models allow coaches and sports analysts to analyze how individual players score against the different defenses and thus, help to determine the best defensive strategy to use against individual players.

For example, if the SportsVU position data show that a particular player/ball handler is particularly adroit at scoring against the switch defense, the team could suggest employing another strategy to improve their winning chances.

Other advantages of this model are:

- Teams can identify areas of poor defensive performance. This can improve their chances of winning in other subsequent games.
- The performance of a player during a game can be accessed as it detects what strategy works bests for a ball handler or screener. This further guides in developing strategies to improve his gaming skill.
- It enables analysts determine the best possible defensive pairs and their most effective form of defense. This can help them optimize their players in for better performance.
In conclusion, the primary aim of performing sports analytics is to help teams identify their area of weakness and strength so as determine the best course of action thereby, improving their sporting performance. With the use of this model, basketball teams can have a better understanding of player’s strategies as well as what players will best fit in a competition against their opponents.

Recognizing and Analyzing Ball Screen Defense in the NBA

This is a review of the sports analytics coaching assistance research conducted by Avery McIntyre, Joel Brooks, John Guttag and Jenna Wiens.

The ball screen (also called a pick and roll) is an offensive tactic that is regularly used in the NBA. To defend against it requires that the defense stop the ball handler from making the shot and denying the screener the opportunity to score. Failing to do this often results in the offense scoring points.

A step has been taken towards building a tool designed to help analyze the way ball screens are defended and determine relationships between defensive strategies and possession outcomes. This system takes in SportVU position data and identifies ball screens, the defensive scheme used, and the offensive outcome in terms of points per possession. For each ball screen four players are identified: the ball handler, the ball handler defender, the screener, and the screener defender. Defensive strategies are separated into four categories: (1) Over - when the ball handler defender stays between the ball handler and screener, (2) Under - when the player defending the ball handler does not stay between him and the person screening, (3) Switch - when the player defending against the ball handler switches places with the player defending the screener, and (4) Trap - when both defenders double team the ball handler.

This tool was then used to review four seasons of data in which over 270 thousand screens were identified. Information was collected regarding the defensive strategy, how many points were scored and which players were involved. As a result it is possible to determine how effective each of the defense strategies are in actual game action.

It was noted that the distribution of the four defensive strategies remained fairly constant over the four seasons. The over defense was consistently used far more often than the other three and the trap defense was used the least. Not all teams followed the same pattern as some teams were able to use the trap defense a greater percentage of the time with good success.

Teams typically decide which defensive strategy to use based on the ball handler’s offensive skill and shooting ability. Analysts are able to look at how individual players score against the different defenses. This will aide teams in deciding what type of defense to use against individual players. If the ball handler is particularly adept at scoring against the over defense then another strategy should be employed. If a ball handler has difficulty scoring against the over defense then this should continue to be used in that situation. Defensive strategies can be individualized to optimize the defense’s ability to prevent the offense from scoring.
Analysts can look at how well defensive pairings are able to execute the different defensive strategies. Analyzing the data suggests that players are not equally effective at all strategies. This would aide teams in determining how to best match up their defensive pairs with their most effective form of defense. Looking at how players do when paired up with different players allow teams to determine how to optimize their players in different situations.

This tool can be used by teams to quickly and easily assess a player's performance during a game and use that information as a guide for developing strategies to improve that player's skills. Teams can also use this tool to identify areas of poor defensive performance and focus on improving these for future games.

Analytics methods used in this research: coaching assistance tool, machine learning, distributions, learned regression
Competitiveness Metrics

The Competitiveness of Games in Professional Sports Leagues

This is a review of the professional sports league competitiveness research conducted by Craig E. Wills.

Competition is why we have sports teams and what draws fans to the games. Games that are not competitive result in fans leaving the stadium early or changing channels on their televisions. Non-competitive games also change the tactics used by the teams. Losing teams will use riskier tactics in order to try to get back in the game, while the winning team will often play more conservatively, and may put in younger, less experienced players to give them experience and rest their star players.

However, what does it mean for a game to be competitive? What factors determine if a game is competitive? Do those factors differ across different sports? Is competitiveness a universal idea across sports?

This study examines competitiveness of games across six different professional sports leagues, five North American sports leagues in baseball, basketball, football, hockey, and soccer plus the English professional soccer league. This selection of leagues offer a wide variety in the sports played, and also allows a same sport comparison between the two soccer leagues.

In order to analyze each sport, scoring events in each game played in one or more seasons for each league were documented, including which team scored, how many points they scored, and at what point in the game they scored.

Competitiveness is explored through several different measures including if a team is leading, if a team is leading by more than a certain number of points, if a team is in the lead and wins the game, if a team is in the lead and never relinquishes that lead, if the team already leading goes on to win, and the probability that a team leading at any given point in the game will remain ahead until the end of the game.

Using these metrics, games in each league were analyzed regarding competitiveness. The results indicate that the MLB league has more games considered to be less competitive. The two soccer leagues tend to be the most competitive, mainly due to the fact that large portions of the game are played with the teams tied. When a team does score, the game becomes much less competitive with a low likelihood that the lead will be relinquished. The NBA is least competitive in four of the metrics. The NFL and NHL fall somewhere in the middle.

The theory of the home team having the advantage is also examined. NFL teams spend the greatest amount of time ahead while playing at home while NHL teams spend the least time. Home team advantage does not apply to MLS or MLB teams.

Television executives could use this information when scheduling sporting events, especially when those events overlap.

Analytics methods used in this research: Multiple Metrics
Computer Vision Techniques

Computer vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from digital images or videos. From the perspective of sports analytics, it seeks to automate tasks that the human visual system can do.

Computer vision tasks include methods for acquiring, processing, analyzing and understanding digital images, and extraction of high-dimensional data from games in order to produce numerical or symbolic information, e.g., in the forms of decisions.


Using Computer Vision and Machine Learning to Automatically Classify NFL Game Film and Develop a Player Tracking System

This is a review of the computer vision and machine learning research conducted by Omar Ajmeri and Ali Shah.

NFL coaches are always looking for weaknesses in their opponents. They spend countless hours going through game film, searching for ways to increase their chances of winning. The process is tedious and often results in many errors. Scouting your own team is very straightforward but scouting all of your opponents is a very time-consuming task.

The classification of NFL "All-22" film has been automated from start (offensive formation labeling) to finish (video player tracking coordinates throughout the life of a play).

A wealth of data can be generated from the resulting system that can help us see coaching tendencies. Analysis of the most common formation, Singleback Ace Pair Slot (two tight ends on the right side of the line, two wide receivers to the left of the line, with one running back) generated the following information: (1) equally likely to pass or run out of this formation; (2) 65% of runs were to the right side, equally split between right guard and right end, with few runs up the middle; (3) 81% of passes were on short routes; and (4) passes gained an average of 4.6 yards per play, while runs averaged just 2.5 yards per play. This is just the beginning of what is possible, giving defensive coordinators and players the ability to create a stronger game plan against their opponents.

Another factor that can be analyzed is player tracking. Looking at average speed and acceleration throughout a game can help coaches gain a better understanding of player fatigue which would allow them to develop game plans that would take this into account.

It is also possible to gain a clearer understanding of how effective players are at running different patterns. Coaches will be able to compare how fast different players are able to run the same type of pattern, allowing for a side-by-side comparison. This would be useful when looking at possible trades, free-agent signings and contract negotiations.
This classification process has the potential to save coaches hours of time that could be spent on more productive activities. At the same time it will give coaches a greater insight into their opponents, allowing them to design a more effective game plan. It will positively affect game planning, scouting, and provide better evaluation of individual players and coaches. The ability to analyze masses of player location data in a short period of time will change how football coaches scout and analyze players and opposing coaches through the league.

This classification process could be extended to replicate the analysis for defensive formations and player analysis. With an analysis of both offensive and defensive formations coaches could examine the relationship between offensive and defensive play calling. The process could also be expanded to include kickoff and punt coverage formations.

Analytics methods used in this research: computer vision techniques, weightings
Conditional Probability

Probability is defined as a measure of how often a particular event will take place if the experiment occurs repeatedly. Probability ranges from zero to one. Zero indicates that the event is impossible, and one indicates that the event definitely will occur. The higher the probability the more likely it is that the event will transpire.

Applying Conditional Probability to Sports Analytics

Probability relates to all past experiments, or occurrences, of that sequence. However, it is often useful to narrow down the situation to include only those with specific criteria. Instead of simply looking at the probability of a baseball team winning a game, you may want to know the probability of the New York Yankees winning when Aaron Judge hits a home run. In this case, you would use conditional probability. Conditional probability is a measure of the probability of an event happening given that another event has occurred. Conditional probability allows us to add extra conditions to the scenario we want to explore.

Conditional probabilities are very helpful as they allow us to include additional information or assumptions into the calculation. They provide information pertinent to a more specific event, rather than simply a general event. However, you must be careful when determining which information to use or the conditional probability may show a relationship between two events that are completely independent, meaning they have no effect on each other. You might look at the probability of the Seattle Mariners winning a game when Edwin Diaz is in the bullpen. The probability will be inaccurate, as having Edwin Diaz in the bullpen does not mean that he necessarily pitched in the game and if he did not pitch he would have no effect on whether or not the Mariners won.

Analysts can use conditional probability in a myriad of ways. They can look at how any player or event can affect the outcome of a game or the likelihood of a player scoring in a given situation. This analysis gives them the capability to rank the effectiveness each player has in a given situation. This information can be used when looking at possible trades or draftees. Which of the available players would have the greatest likelihood of scoring for their new team? This information helps determine which players would be the best choices to improve the weak spots in their team's roster.

Coaches can use conditional probability to help determine their line-up for any specific game. A baseball coach can use conditional probability to determine the likelihood of each his players scoring a run when facing the scheduled pitcher. It also helps in deciding when to put a designated hitter or designated runner into the game in order to increase their probability of scoring in a given scenario and improving the team's chances of winning the game.
Conditional Random Field

Conditional random fields (CRFs) are a class of statistical modeling method often applied in pattern recognition and machine learning and used for structured prediction. CRFs fall into the sequence modeling family. Whereas a discrete classifier predicts a label for a single sample without considering "neighboring" samples, a CRF can take context into account; e.g., the linear chain CRF (which is popular in natural language processing) predicts sequences of labels for sequences of input samples.


"Quality vs. Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data

This is a review of the soccer shot prediction research conducted by Patrick Lucey, Alina Bialkowski, Mathew Montfort, Peter Carr, and Iain Matthews.

Within any soccer game, the team with the most shots on goal is often not the winner. Why is this? Obviously, not every shot has the same probability of resulting in a goal. Using player-tracking data, this research looks at how to quantify the value of a shot.

Fine-grained player tracking data is used to determine several features: how close the shot is to the goal, the distance from the defender, the number of defenders between the shot and goal as well as the position of the other attackers on the field.

Match context is also taken into account with shots organized into six different context groups, namely open play, counter attack, corners, penalties, free kicks and set pieces. Grouping shots this way and then looking at which contexts lead to the greatest number of goals tells us that teams have a greater probability of scoring while on a counter attack than during a normal possession. A normal possession is more likely to result in a goal than a penalty kick. Corners have the lowest probability among all six contexts of resulting in a goal.

Player and ball tracking data is used from one season of games from a professional soccer league in Prozone. This resulted in 9732 shots from which spatiotemporal patterns were analyzed. This information was inputted into a Conditional Random Field in order to estimate the probability of a team scoring from a given shot. Teams are analyzed in regards to quantity and quality of shots using team tendencies rather than individual characteristics.

The position of the defenders has a major impact on the offense in terms of the decision regarding whether or not to attempt a shot as well the actual execution of the shot. In order to determine the defenders' orientation towards the ball, it is first necessary to determine if there are any defenders located between the shot and the goal. If any are present, the Euclidean distance between shot and defender is determined. If there are no defenders within the area, the distance is given a negative value. All match contexts show that the probability of scoring increases when there is no defender between the shot and the goal.
Another factor to look at is the number of defenders in this area. A standard point-in-polygon calculation is utilized to determine which players are within this area.

Defensive structure is determined by assigning roles to each player, done by finding the permutation of the raw location points, which minimizes the distance, and then utilizing a Hungarian algorithm to ultimately assign a role to each player. From this, four further features are created: the distance between the defensive line, the distance between the back-line and the midfield-line, the number of defensive role changes and the number of attackers in front of the defensive center.

Offensive features incorporated into the model include whether it was a long pass, cross, dribbling and taking on players, or pressing. Also included is the position of the player who passes the ball to the shooter, the pace of the players, and how the team on offense moves relative to the defense.

From all of this information the expected goal value of each shot is calculated using logistic regression.

This information can be used to compare and contrast characteristics between teams. The efficiency of the various offences and defenses can be calculated and also be used for comparison purposes. Coaches can further analyze the efficiency of their team within different matches, providing information regarding strengths and weaknesses of their team, as well as their opponents. This information would help build stronger strategies to incorporate while playing the various opponents. Finally, analysts can look at the results and determine whether the winning team won based on talent or luck.

Training players to evaluate the expected value of a shot would ultimately lead to a greater shot efficiency. However, at the same time, defenders could use the same information to improve their defensive strategy.

Analytics methods used in this research: Conditional Random Field, Euclidean Distance, Standard Point-in-Polygon Calculation, Permutation, Hungarian Algorithm, Logistic Regression
Conservation of Runs Framework

A conservation of runs framework is based on the idea that for every run value gained by the offense a corresponding run value is lost by the defense. Offensive run values are adjusted for several factors beyond the control of the players including the ballpark and handedness of the pitcher and batter. Base runners are only given credit for advancing beyond what would be expected given their starting position, number of outs, and the hitting event. Hitting performance is evaluated relative to the expected hitting performance based on all players at the same fielding position. The defensive run value must be divided between the pitcher and other fielders involved. The division is determined by how difficult the batted ball was to field. The entire value is assigned to the pitcher if the plate appearance does not result in a play such a strikeout or home run. Any value not assigned to the pitcher is divided among the fielders who were potentially responsible for the ball.


openWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball

This is a review of the major league baseball research conducted by Benjamin S. Baumer, Shane T. Jensen, and Gregory J. Matthews.

The purpose of sports is to win. Analysts use statistical methods to study key areas of interest. One challenge is quantifying the contributions of individual players to their team and the number of games the team wins. Major League Baseball uses wins above replacement, or WAR, in this determination.

Wins above replacement is comprehensive and easy to understand but has two problems, namely a lack of uncertainty estimation and a lack of reproducibility. In order to combat these issues openWAR is developed, which is based on a conservation of runs framework with uncertainty values estimated by resampling methods. Central to the model is the idea that the positive and negative consequences of all runs scored must be allocated across four aspects of baseball performance: batting, base running, fielding, and pitching.

In openWAR, the win above replacement is defined as the sum of all contributions from each of the four aspects, compared to a hypothetical replacement level player after controlling for factors like ballpark, handedness, and position. OpenWAR is not a statistic to be used for forecasting but instead is a retrospective measure of player performance. The credit or blame for hits on balls in play is shared between the pitcher and fielders. The extent to which each is responsible is determined by the location of the batted ball.

A conservation of runs framework is based on the idea that for every run value gained by the offense a corresponding run value is lost by the defense. Offensive run values are adjusted for several factors beyond the control of the players including the ballpark and handedness of the pitcher and batter. Base runners are only given credit for advancing beyond what would be expected given their starting
position, number of outs, and the hitting event. Hitting performance is evaluated relative to the expected hitting performance based on all players at the same fielding position. The defensive run value must be divided between the pitcher and other fielders involved. The division is determined by how difficult the batted ball was to field. The entire value is assigned to the pitcher if the plate appearance does not result in a play such a strikeout or home run. Any value not assigned to the pitcher is divided among the fielders who were potentially responsible for the ball.

For each plate appearance, the run values for each base runner, fielder, hitter, and pitcher are calculated. The overall run value for a player is determined by adding those run values across all plate appearances involving that player as a hitter, pitcher, base runner, or fielder.

Team management is able to accurately evaluate players in regards to their contribution to the team when looking at offering contracts or making trades. Players in different positions can be ranked according to their openWAR score. This would facilitate comparison between high and low ranking players, allowing coaches to determine weaknesses within their players that can be improved upon in order to increase team effectiveness.

Instead of putting all the blame for a loss on the pitcher or all the credit for the win on the batter this statistic allows all players to take their part of the blame or the credit, holding them more accountable for their actions.

Analytics methods used in this research: WAR, openWAR, Conservation of Runs Framework, Resampling
Convex Hull

In mathematics, the convex hull or convex envelope or convex closure of a set X of points in the Euclidean plane or in a Euclidean space (or, more generally, in an affine space over the reals) is the smallest convex set that contains X. For instance, when X is a bounded subset of the plane, the convex hull may be visualized as the shape enclosed by a rubber band stretched around X.


Predicting Golf Scores at the Shot Level

This is a review of the golf research conducted by Christian Drappi and Lance Co Ting Keh.

To date the majority of statistics generated for golf focus on evaluating, quantifying, and comparing golfers' performances.

A model is developed to predict golfer scores on a shot-by-shot basis. This prediction is based on the current state of the shot including shot number, distance from the hole, and lie of the ball as well as features of the golfer's skills, course, hole and wind speed.

Data was collected covering all shots taken on the PGA tour from 2009 to 2015 for use a training set and 2016 data was used as a test set.

The first step was to decide which classes of features are needed for the analysis. For the purpose of this model, four features were chosen: golfer, hole, course, and game-state features.

When building the golfer features past data is weighted in order to measure current skill levels and normalizing skill based on the difficulty of the course. The skills produced are scoring average, driving distance, driving spray, smoothed strokes gained via long approach, short game and putting. Smooth strokes gained is computed by determining where each player gained or lost strokes over the course of a round with an added penalty function for strokes drained or shots that hole out from a long distance. Then a golfer's performance is normalized based on the difficulty of the course.

Hole features used include green size, sand area, normalized penalty strokes and fairway width. These are created by manipulating the spatial data. The size of the green is estimated by computing a convex hull. A K-nearest neighbor statistic is utilized to deal with sand bunker characteristics.

Course features include difficulty, fairway height, rough height, stimp, wind speed, fairway, and green firmness. Difficulty was computed by looking at the residual when the actual scoring average is regressed with the scoring average feature. The game-state features include the distance from the hole and lie. Golfer skill was trained with a single hidden layer neural set, a more comprehensive neural net was built for training the golf skill, hole, and game state features. Shot-by-shot probabilities were forecast using a softmax regression, random forests, and a single-layer neural network.
This model can be used by golfers to help them determine their best strategy during a golf game that leads to the greatest probability of making a shot and winning the game. Course managers can apply this model in determining any needed changes that could create an improved course. Analysts in determining which golfer is making optimal choices, or in determining changes a golfer should make in order to maximize their potential, can employ this model.

Analytics methods used in this research: Weighting, Convex Hull, K-nearest Neighbor, Single Hidden Layer Neural Set, Softmax Regression, and Random Forests
Convolutional Neural Networks

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery. CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.


Mythbusting Set-Pieces in Soccer

This is a review of the convolutional neural networks research conducted by Paul Power, Jennifer Hobbs, Hector Ruiz, Xinyu Wei, and Patrick Lucey.

In the English Premier League, the discrepancy between the larger market and smaller market teams grows consistently wider. One strategy that the small market team could use to help decrease this disparity is increasing the effectiveness of their team's set pieces. Set pieces in soccer are any play that begins with the ball at a standstill following a stoppage in play due to the ball going out of bounds or a foul call. Set pieces include free kicks and corner kicks. Free kicks are those kicks a team can take without their opponent interfering. Corner kicks happen when the ball goes out of bounds over the goal line, without a goal being scored, and the last player touching the ball is a member of the defending team.

A Convolutional Neural Network (CNN) is used to study the effectiveness of a team's set pieces, as well as their expected goal value. The network analyzed three seasons of English Premier League data consisting of tracking data of the players and ball, as well labels for each event. The dataset included over 12,000 corner kicks and 3,600 free kicks.

A set-piece grammar model was created to define the sequencing of events, both simple and complex. It looks at how the set piece begins, how it ends, and what events occur in between the two. Also analyzed is the effectiveness of the kick as well as the defending teams' ability to defend against the kicks.

The set piece must get past the first defender or it will never be successful. The way the ball is kicked also influences whether or not a shot or goal is made. The data indicates that while in-swinging kicks have a lower chance of creating a shot, the chance of scoring actually doubles.

The tactics of the defending team are also analyzed, specifically looking at whether they are playing a man-to-man or zone defense. It also looks at where the players of the defending team are positioned on the field. The model accurately predicts the type of defense that an opposing team will utilize in any given situation at a success rate greater than 80%.

Analysts are able to determine how closely the style of the opposition resembles other teams they have faced in the past. They are also able to determine the approach the opposing team typically take when
defending against a corner kick. They can also look at whether a team is consistent in their playing style or if their type play varies from game to game, depending on whom they are playing against.

Coaches are able to use the resulting information to determine the defensive skills of their opponent and the likelihood of how they will defend against a set piece. This can be used to develop a strategy to maximize the team's probability of scoring when facing these opponents.

Analytics methods used in this research: convolutional neural network
Cross-Validated Root-Mean-Squared Error

Root Mean Squared Error Calibration (RMSEC) is derived from an average of how well (or how close) the data points are to a calibration line using all data points (samples). Root Mean Squared Error Cross Validation (RMSECV) is an internal method to 'test' your model (or calibration model) to make sure it isn't badly skewed by a data point or if the model is over fitted, or if there are any outliers. RMSECV is determined by the validation method chosen. The most popular being a 'leave one out' method. Here a data point is removed when building a 'new' calibration model with a new RMSEC, then the data point is put back, but a different data point is removed and new calibration model is made with a new RMSEC . The two RMSECs can be averaged giving us the beginnings of RMSECV. This process is repeated until every point has a chance at being left out.

Source - https://www.researchgate.net/post/PLS_for_regression_RMSEC_vs_RMSECV

Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing

This is a review of the NBA research conducted by Joseph Sill applying cross-validated root-mean squared error.

Adjusted Plus Minus (APM) is a technique used for evaluating players. Adjusted Plus Minus starts with the idea that what is the most important is the player's contribution to the team's margin of victory. The +/- stat is the difference between the points the player's team scored and the points the opponents scored while the player was on the floor. Using the +/- stat to evaluate a player has an inherent problem in that it does not take into account who the player played with and played against.

APM is a method which uses regression to take into account the teammates and opponents while a player was on the floor. The regression process produced a rating that indicates the effect the player would be expected to have on a team's margin of victory if he were to replace an average NBA player in the team’s lineup.

There are questions as to the validity of the APM. The results can seem misleading, especially when the analysis is run on data from a single season. One problem is multicollinearity, which relates to situations where pairs of players are very frequently or very rarely on the floor at the same time which affects the statistics regarding the individual player. Another problem is overfitting. Overfitting happens when the model describes random errors or noise instead of the underlying relationship. Overfitting happens when a model fits all the training data, including plays and situations that are flukes and unlikely to ever occur again. This problem decreases the model’s ability to predict future performance. For example, a player has one of those unexplainable, extraordinary games and scores points way beyond their usual average. Using this to predict that player’s scoring percentage over the next few games is unrealistic. In order for APM to be more accurate it must be able to discount these outlying events.

Does APM have the ability to predict what will happen in games which are not included in the dataset inputted into the APM model? In order to answer this question a cross-validation (CV) technique is used. The game is divided into fragments (based on who was on the floor vs. whom and for how long). The
APM model is then used to generate a prediction about the outcome of each game fragment within the game. These game fragment predictions are added together to get a prediction about the overall margin of victory during the game. This predicted margin of victory is then compared to the actual margin of victory. CV can also be used to determine the best settings for guidelines such as the minutes cut-off (for the reference player) and the emphasis placed on the prior year.

Applying the cross-validation technique to the APM results in improved accuracy. This improved accuracy can help teams better evaluate how individual players contribute to the team.

Analytics methods used in this research: adjusting plus/minus, overfitting, regression, cross-validation root-mean-squared error, ridge regression, regularization, out-of-sample testing
Data Science Methods

Data Science Methods in Sports Analytics

This chapter briefly discussed data science methods and the role they play in sports.

The use of data science in sports is growing. We have seen some examples in the soccer, basketball, the Olympic Games, national and international competitions and the list goes on. This is because data science and sports analytics are opening new data points and project insights.

*What is Data Science?*

Data Science is a broad term in which scientific methods, mathematics, statistics and other tools are applied to datasets to extract knowledge and insights. Essentially, data science makes use of multifaceted tools to deal with Big Data and get useful information from it.

Analytical sports data vary depending on the use case. Many teams use social network trackers to detect signs of fan dissatisfaction, in order to stay in touch with fan needs and habits.

*Data Science in the NBA*

In the NBA, player tracking tools are used to track where players pass and shoot the ball, how they behave in offensive and defensive situations and the probability that they will succeed or fail. The teams that process their data most effectively tend to win. It is no accident that the Golden State Warriors, who integrate daily into their daily decision making, are one of the most successful teams in history.

*Improvement of the fan experience*

This kind of public relations approach can allow teams to get through controversies and keep their fans happy. For example, Bayern Munich of the Bundesliga in Germany has a permanent agreement with SAP to handle global efforts to monitor the management of fans. The next time you complain online about a big trade or a bond issue for a new stadium, the team may be listening a little closer to what you might expect.

*How the data science is being used: Predictive Analytics*

Predictive analytics is highly valued for teams that collect a lot of data, as it gives them more information to take into account when making their decisions, such as player acquisitions and draft selections. Smart organizations understand that they should look at every situation from as many angles as possible, especially when making important decisions. This way new metrics and data points can be consistently generated, which allows teams to obtain as much information and insight into their opponents and team habits.

Science and data technology are now playing an increasingly important role in player perception. ESPN, for example, is one of many groups that create their own statistical models designed to identify which players have the best chances of succeeding at the next level, and/or which players are the best fit for a team.
Data Science really began to expand its profile and exploration in basketball after the introduction of Synergy Sports Technology. Synergy Sports Technology is a service that collects videos and statistics and compiles it into a database that is easily accessible and easy to use.

Through the tools developed by Synergy Sports Technology, you can track statistics down to the minute and then immediately access the relevant play and assess the player and/or team’s actions. This means that scouts are easily able to read the efficiency with which a player drives with his or her left hand, and then see clips of them executing that play.

*Predictions of players*

On the other hand, ESPN has developed a model that aims to project the possibilities of a player to become a star of the sport.

The ESPN model tries to project the Plus-Minus Statistics of a player to predict player value when on the court. There are many other groups that make similar models, and all of them add to the discussion of data science and it’s potential.

Ultimately, the science of data and sports analytics has spread and the exploration process for NBA teams is one of the many key examples. Increasingly, teams rely on these methods to help their scouting departments obtain more information about individual perspectives.

But it is not only the organizations themselves that are entering this field, since the media are integrating more and more metrics and advanced analysis in their coverage project, which in turn influences the participation of the fans with respect to preliminary perspectives.
Data Tables

Data tables are any display of information in tabular form, with rows and/or columns named. A table stored in, or derived from, a database. Data is complex, and all data is different. Accordingly, Data Tables has a wealth of options which can be used to configure how it will obtain the data to display in the table, and how it processes that data.

Source - https://datatables.net/manual/data/

The Effect of Ball Wear on Ball Aerodynamics: An Investigation Using Hawk-Eye Data

This is a review of the Hawk-Eye Electronic Line-Calling System Data research conducted by Simon Choppin, Simon Albrtecht, James Spurr, and Jamie Capel-Davies.

As technology has improved, the ability to analyze the game of tennis has been enhanced. One such technology is the Hawk-Eye Electronic Line-calling System. This system automatically tracks the ball and players during a game, producing such data as ball velocity, bounce locations, and flight trajectories of the ball. This system also gives players the ability to challenge line-calling decisions. The data produced by the system is used to generate statistics during a game but is also used for scientific research. The data has allowed sports engineers to evaluate the effectiveness of the equipment being used by the players.

One important piece of equipment in a tennis game is the ball itself. The data produced by the Hawk-Eye Electronic Line-Calling System has been used to research the effect wear has on a ball’s aerodynamics and how it affects the play.

A data set was created from 364 matches played during the 2012 to 2017 Davis and Fed Cup tournaments. The data included 71,019 total points. Three data tables were created. The first was a match data table, which included information such as the players involved, the tournament, and the date the match was played. The next table was the point table, which included information about the flow of the game. This included who was serving, the current score, the winner of the point, etc. The third table was the shot data table, which included information regarding the physical characteristics of a shot such as the impact the racket had on the ball, how the ball passed the net, and the impact the ball had on the court.

Following this, the type of ball used in each point was determined. A ball was considered new during the first two games it was used. A ball was considered used during the last two games it was used. On average, a ball was hit with a racket 39 times before being exchanged for another ball.

To determine the aerodynamics of the balls only first serves were looked at. The deceleration of the ball between being hit with the racket and making contact with the court was calculated. It was determined that serve speeds were slightlying higher with new balls resulting in higher impact speeds on the court and less time in the air. New balls slow down less than used balls as the felt of a used ball is actually raised up rather than flattened.
Further study is needed to look at whether serving with a new ball has a psychological impact on the player and how different types of rackets change the impact they have on the ball and the resulting velocities.

As this study is refined, tournaments can use the resulting information to determine the optimal time to introduce new balls into a match.
Data Visualization

Data Visualization as a Method of Sports Analytics

Data visualization, like other sports analytics methods involves the use of data and advanced statistics to measure team performance and gain a competitive sports advantage. Data visualization is a visual communication that presents data in graphical and pictorial form.

With data visualization, large data and records can be quickly and conveniently analyzed without having to go through the data piece by piece.

Data presented in this manner are in the form of charts, graphs, maps and tables. Today, I will discuss how data visualization is used to improve team performance.

Data Visualization in Sports

Sir Clive Woodward, the past coach of the English rugby team, at the Sports Performance and Tech Summit of 2014, spoke about how the knowledge of data visualization helped his team in winning the 2003 Rugby world cup final, held in Australia. The team players were reported to have analyzed their various performances through the use of data visualization and worked on their weaknesses prior any competition; thus, their predictable success in 2003.

Data visualization methods usually require the use of software and tools, such as Microsoft Excel, PowerPoint, and Tableau.

These tools can collect data from several databases with numerous records and create a pictorial image of them. They are generally referred to as data visualization software.

Data visualization can be applied to sports in the following ways:

It helps sport teams identify areas that need attention or improvement

The analytic software commonly used here is Prozone. ProzoneTM links several complex and unrelated data together as a unit so it can be easily understood by its users. It also aids teams and coaches to gain better understanding of their opponents’ strategy as well as suggests most suitable operations and actions that can be adopted for better results. It does this by highlighting players that would best fit within a team tactical system. This software can also be used in football, baseball, and basketball analytics.

Data Visualization is Time-Saving

Because data visualization provides for flexibility in converting one chart to another, teams can modify their data in a short or no period of time at all. It also saves time required for reading long reports and data.

It is important for spot diagnoses and problem solving
For instance, a fall in team’s performance can be easily detectable on the bar chart when, without having to go through the rigors of searching through several data and records. Data visualization programs also provides other visualizations like action rates, individual player movement, passing diagrams etc.

*It provides an interface to interact with data*

Interactive data visualizations can enable teams to manipulate their data records in previous games and competitions in order to uncover factors such as reasons for their deficiencies in a competition, etc. This can greatly improve teams’ performance.

In conclusion, the data revolution in sports and gaming has improved the way coaches, teams and fans interact with events on the field. Sport teams will perform better if the method of data visualization is applied in their analytics.

**Applying Data Visualization in Sports Analytics**

Sports analytics is a rapidly evolving technique that involves the use of data and advanced statistics to measure team’s performance and gain advantage in the competitive sports field.

Several methods had been adopted over the years to improve the results of sports analytics. Methods like classical statistics, simulation, frequency distribution and data visualization have been proven to enhance performance in sports and this presentation aims to shed light on how the application of data visualization in sports analytics can be used to optimize performance in sports.

Data visualization is a modern visual communication that presents data in graphical and pictorial form. It is a quick and convenient way of analyzing large records and data. Data visualization makes it easier for users to understand their large data in a simple formal. Data presented in this manner are in the form charts, graphs, maps and tables.

As technology continues to transform our world, data collection has become more crucial in sport analytics and can be conducted with relative ease, thus, leading to the development of advanced statistics and data optimization.

The use of data visualization cuts across various sectors, as, research, advertising and marketing, banking, business setups, factories, health care, sports etc. Hence emphasizing its importance in enhancing productivity.

Data visualization usually requires the use of software and tools, such as Microsoft Excel, PowerPoint and Spreadsheet etc. These tools come with database software which collects data from the database containing numerous records and creates a pictorial image of them.

The benefits of data visualization in sports analytics include;

*Helping teams Identify areas that need attention or Improvement*

Visual representation of data statistics of team players helps to aggregate performance data into easily ways data can be interpreted and understood.
A popular performance analytics software used by many teams is Prozone, this software links complex and unrelated software and brings data together in such a way that can easily be understood by its users. This software allows teams and coaches to gain better understanding of opponents' strategy and also suggest most suitable operations and actions by highlighting players that would best fit within a team tactical system. It also measure individual player metrics and load to prevent players injuries.

Time-saving

Data visualization saves time required for reading long reports. Also, making changes to the charts and graphs is much easier as the data visualization software provides flexibility to convert one chart to another and make changes to specific data which needs to be modified.

Allows team to immediately spot problems and draw conclusions from them

For instance, it can be easily detectable on the bar chart when there is a fall in the team's performance without having to search through several data and records. This has a way of spurring the team to quick action for better performance. It also provides visualizations like action rates, individual player movement passing diagrams etc. Thus enabling team members to perform quick analysis of their strengths and weaknesses.

Provides an interface to interact with data

Interactive data visualizations can enable teams manipulate the data to uncover other factors such as the reasons for their deficiencies in a competition. This creates a better attitude for use of analytics.

A notable example of how data visualization has helped (he world of sports can be traced to 2003 when the English rugby team won the World cup in Australia.

Sir Olive Woodward, at the Sports Performance and Tech summit gave an account of how he utilized the knowledge of data visualization in helping his team win the world cup. The team players were reported to have analyzed their performances through the use of data visualization and worked on their weaknesses prior any competition: and this led to their predictable success in 2003.

In conclusion, the use of data in sports and games has improved the way coaches, teams and fans interact with events on the Held, the use of data visualization in sports analytics has proven to be one of the best methods in performing sports analytics.

Data visualization not only assists teams in determining the best course of action, it also enhances the performance and efficiency of team players. This method of sports analytics should be incorporated even in all sporting teams to boost performance and increase creativity of teams and individuals.
Deductive Reasoning and Inductive Reasoning

Deductive Reasoning and Inductive Reasoning as a Method of Sports Analytics

This chapter focuses on another technique of sport analytics which looks quite simple, but is very effective in enhancing the performance of a player or a team—and that is Deductive and Inductive Reasoning.

Coaches often use this method to reach a particular conclusion—inductive reasoning is used to reach a generalized conclusion from a specific instance while deductive reasoning is used to reach a specific conclusion from a generalized instance about a player or the team.

For deductive and inductive reasoning in sports analytics, concepts and variables of interest are first identified and hypothetical situations are then developed between the variables. One example is the study of O’Donoghue and Ingram in 2001 regarding tennis strategy at Grand Slam tournaments. This is research, they formed three concepts around gender, court surface and strategy. It was around these three concepts that hypothesis were developed using the links and associations between the concepts.

The various methods that are used in such situations include intuition, experience, authoritative sources, and reasoning. But here, we focus on reasoning.

Reasoning can either be deductive or inductive and coaches and sport analysts employ this reasoning technique to determine the potential links between the concepts they have formed. Let’s look at how the above concepts work.

Gender

Looking at the concepts formed in the study of O’Donoghue and Ingram, it can be said that if there were differences between style of tennis played by males and females, the revelation of this knowledge can be used by a coach to enhance the style of a new player to suit the style of their gender.

Court Surface

It has been noticed that in grand slam tournaments, some players performed relatively better on certain court surfaces and performed less successfully on other surfaces. This validated the consideration of court surface as a good concept, which reasoning techniques can be applied in tennis analytics.

Strategy

Another concept that was formed through this study was based on the strategy players adopted during their tennis matches. Typically, strategies are planned out before the match or even before the tournament. So for a coach to increase the performance of his or her player during a tournament, they might be able to leverage deductive or inductive reasoning to ascertain the best strategy that will help their player win the game.
Having come up with various concepts such as explained above, coaches, sport analysts and even commentators can rely on their reasoning technique to foretell the chances of one player winning against another.

**Deductive Reasoning**

In performance analysis, after various concepts have been formed, then different variables can be generated, which can be used around each of these concepts through deductive reasoning. So deductive reasoning is where a hypothesis is formed and tested through a systematic observation from the concepts and variables.

What is peculiar about deductive reasoning is that speculations are formed about general variables observed about a concept and these speculations become the hypothesis which are tested in order to arrive at a specific conclusion.

For example, the concept of gender in a tennis match mentioned above can generate variables such as: male tennis players are physically stronger than the female players. This is a generalized theory, which can result into a specific conclusion that it is most likely that a male tennis player will perform better on ground strokes than a female player due to strength. Or at least the stronger female players will perform better on ground strokes than weaker female players.

**Inductive Reasoning**

In inductive reasoning, however, observations derived from real life events are used to generate a general hypothesis, which is further tested. For instance, a tennis analyst who has observed particular matches in tournament finals over a period of time can use inductive reasoning to provide a case study about how a particular concept, such as gender or strategy can determine who will win another final.

In conclusion, deductive and inductive reasoning is a simple logical technique used in sport analytics to come up with either a specific or generalized conclusion about a player. This helps coaches better assess the players, according to their abilities.
Deep Imitation Learning

Imitation learning techniques aim to mimic human behavior in a given task. An agent (a learning machine) is trained to perform a task from demonstrations by learning a mapping between observations and actions. The idea of teaching by imitation has been around for many years; however, the field is gaining attention recently due to advances in computing and sensing as well as rising demand for intelligent applications. The paradigm of learning by imitation is gaining popularity because it facilitates teaching complex tasks with minimal expert knowledge of the tasks.

Data-Driven Ghosting Using Deep Imitation Learning

This is a review of the deep imitation learning research conducted by Hoang M. Le, Peter Carr, Yisong Yue, and Patrick Lucey.

Statistics typically compare players and teams to the league average. However, this can be very limiting. A coach may want to compare their team against an opposing team rather than the league average. Or they might want to determine what characteristics allows an opponent to dominate in a certain aspect of the game in order to adopt that style for themselves or develop a strategy to defend against it. A data-driven ghosting technique can be useful in these situations.

A network employing deep imitation learning can be used. This type of network allows ghosted players to anticipate the movements of their teammates as well as the moves the opposition will make. Imitation Learning is also known as learning from demonstrations. Machines learn good policies and skills by observing expert behavior. Deep learning involves learning complex layers of information hidden within data. The network learns from player tracking and event data from 100 games played in a professional soccer league. Additional information was inputted into the system including the main role of each player, coordinates of each player and the ball, as well the distance and angle of each player towards the ball and goal.

Analysts can use this network to look at how different teams stack up against each other. It can help determine which team is expected to win, especially in cases where the two teams are evenly matched.

This network can now be used to look at the consequences that would emerge from hypothetical situations. This provides coaches with a safe way to examine different types of plays without putting their players in harm's way. Coaches can see how their team compares to the average team in order to help determine their specific strengths and weaknesses.

The ghosting scenario can be changed so that instead of reflecting an average team a specific team is reflected instead. Specific aspects of the teams, such as defensive behavior, can be compared. This gives coaches the ability to look at how other teams would handle different situations or how they might react to the plays they plan to use in their next match. From this they formulate a game plan to counter the decisions the other team is expected to make. Comparing characteristics between teams can help decipher what makes their opponent tick allowing coaches to develop effective play strategies.
Analytics methods used in this research: data-driven ghosting, deep imitation learning, variance, expected value
Deep Learning

Deep Learning is an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning is a subset of machine learning in Artificial Intelligence (AI) that has networks capable of learning unsupervised from data that is unstructured or unlabeled.


Deep Learning of Player Trajectory Representations for Team Activity Analysis

This is a review of the spatio-temporal patterns and deep learning research conducted by Nazanin Mehrasa, Yatao Zhong, Frederick Tung, Luke Bornn, and Greg Mori.

Player interactions are a key component of team sports. Every player has a unique trajectory style. Putting these players together as a team creates a unique style for that team. These unique styles would allow this model to recognize individual players simply from their trajectory style. The same would apply to recognizing teams.

Experiments are done on two different sets of data. The first set uses player positions and appearances gleaned from broadcast video footage of NHL hockey games. The second set includes trajectory or route features only. The routes players use are determined by using a tracking system that records player positions during NBA basketball games.

NHL players are tracked automatically in broadcast videos using SPORTLOGiQ Inc. The data taken from this includes the positions of the players on the ice as the game progresses. At the same time, it is also possible to determine the events taking place during the game. Six types of events are defined: pass, dump out, dump in, shot, carry and puck protection. Positional and event data is collected from 8 games. The goal is to be able to predict which event is taking place based solely on the trajectories of the five players on the ice. This is possible to do because as each event unfolds the players are positioned in different areas and patterns around the rink. For example, the dump in happens when the players are in the neutral zone, and the dump out occurs when players are in their own zone.

STATS SportVU data shows the positions of the players and the ball in NBA games. This data is used to determine the identity of a team simply by analyzing the trajectories of its players as they play the game. The data contains 137,176 possessions which start with the offensive team having the ball and end with a shot. All of the possessions in a game are used collectively to predict which team is playing. Not all possessions will be indicative of the same team so the majority wins. Whichever team the majority of possessions correlate with is determined to be the team.

Overall, the spatio-temporal patterns experiments had good results in their tasks, demonstrating that players and teams do have unique trajectory patterns.

The outcomes of these experiments could be used to learn about various team and player behaviors. Teams could use this data to determine which combinations of trajectory paths make up the winning
teams in the league. This then could be used to determine which players might be valuable in improving
the skills of their own team. Players could examine the trajectory patterns of star players in their league
and perhaps determine how to tweak their own styles to make them more effective.

Analytics methods used in this research: spatio-temporal patterns, deep learning model, 1D
convolutional network, permutation invariant sorting
The Advantage of Doubling: A Deep Reinforcement Learning Approach to Studying the Double Team in the NBA

This is a review of the deep reinforcement learning research performed by Jiaxuan Wang, Ian Fox, Jonathan Skaza, Nick Linck, Satinder Singh and Jenna Wiens.

Double teaming (two defensive players focused on one offensive player) is a strategy regularly employed in the NBA. It can result in slowing down strong offensive players; however, it can also be risky since it leaves another player open. Balancing the reward/risk is a very challenging proposition in the NBA and deciding which way to go depends on many different factors. Effective double teaming involves taking into account where and who all the players are, as well as anticipating the ball handler's next move. The offensive strategy of the opposing team must also be taken into account. A framework using the deep reinforcement learning approach has been developed to determine when it is the right time to double team another player.

This method for learning how to effectively double team uses the reinforcement learning framework. In this framework the player is interacting in their environment to receive an award, such as blocking a shot. Over 643,000 possessions from the previous three seasons were entered into the framework. Each possession began once all players crossed half court and concluded when the shot clock reset. Included are possible factors affecting the decision to double team such as player heights, weights, shooting abilities, current state of the game, shot clock and game clock. This particular framework focuses strictly on whether or not to double team the ball handler.

After running the data through the framework several observations were made. It was determined that double teaming results in a significantly lower field goal percentage for the offense overall. However, it is also more likely that the possession will end in a foul. It was also determined that when a ball handler is double teamed they typically tend to pass or dribble the ball. The observations indicate that it is more beneficial for the player to pass the ball rather than keep the ball themselves.

In the final analysis it was determined that it is important for the double teamer to be positioned in such a way that he blocks a potential pass to the open man. It also indicates that it is better to double team role players rather than star players. This goes against the typical strategy currently employed in the NBA where star players are double teamed more regularly than role players.

This approach gives analysts a more comprehensive framework to employ while working to understand and evaluate defensive plays and which offensive players perform best when double teamed.
Coaches face many tough decisions every game, just one of them being when to double team and when not to. They can look at this research to help clarify or improve their decision making skills in this area. This framework can also be modified to answer other questions regarding offensive and defensive strategy.

Analytics methods used in this research: deep reinforcement learning, neural network, expected outcome
Demand Model

Demand modeling uses statistical methods and business intelligence inputs to generate accurate demand forecasts and effectively address demand variability. Demand modeling is becoming more important because forecasting and inventory management are being complicated by the increasing number of slow-moving items, the so-called “long-tail” of the product range, many of which have unpredictable demand patterns in which the typical “normal distribution” assumption used by traditional models is totally inadequate. In these scenarios, successfully managing forecasts and inventories requires advanced demand and inventory modeling technologies in order to reliably support high service levels.


Evaluating the Effectiveness of Dynamic Pricing Strategies on MLB Single-Game Ticket Revenue

This is a review of the MLB ticket pricing research conducted by Joseph Xu, Peter Fader, and Senthil Veeraraghavan.

As much as sports teams are about winning championships, the aim of management and stakeholders is to maximize revenue. Many teams across various sporting leagues have adopted a dynamic pricing strategy for single-game tickets. However, these prices are typically based on recommendations from outside vendors. The question becomes what pricing strategy will optimize ticket sales and maintain a strong fan base.

In an attempt to answer this question a demand model for single-game ticket sales is developed. Information was generated from one MLB team to test this data against actual ticket sales in order to set appropriate parameters. The team’s pricing strategy is analyzed and three alternative strategies are developed.

The MLB team providing the data instituted two different pricing strategies over the season. Prior to the All Star break, they maintained a variable pricing strategy where ticket prices varied across games and seat sections but remained constant over time. After the All Star break a dynamic pricing strategy was implemented. The data from the first part of the season was used to determine appropriate parameters while the data from the second part of the season was used to create revenue predictions.

Customer demand for single-game tickets is modeled in three stages: game decision, ticket quantity decision, and seat section decision. Game demand is modeled with a negative binomial regression taking into account the effect of time, game characteristics, team performance, price, and occupancy. The resulting parameters illustrated the fact that higher prices result in lower demand, promotions increase demand, and opponents affect demand as well as stadium occupancy and team performance.
The ticket quantity decision, or the second stage of the customer demand, is modeled with a negative binomial regression using the same set of variables as in the first stage. The resulting parameters are the same as those generated in the previous stage.

The third stage, or seat section choice, is modeled using multinomial logistic regression. This model includes additional factors such as the time until the game and the number of tickets required. The resulting parameters included the fact that the best seat sections tend to sell sooner than other seat sections and customers looking for more tickets typically do not purchase the most expensive seats.

These three models are combined in order to make predictions for expected daily revenue for each seat section for every game. Three alternative dynamic pricing strategies are tested using the model, optimal variable pricing looking ahead at team performance, monotone myopic dynamic pricing, and unrestricted myopic dynamic pricing. The greatest positive change in revenue was created with the unrestricted myopic dynamic pricing.

Teams can use this information when setting single-day ticket prices that will maximize revenue. Ticket prices should not be changed early in the season but rather closer to the end of the season when demand tends to rise. Not all seat sections need to follow the same ticket pricing strategy. The optimal strategy for each section varies depending on factors like time until the game, team performance, and number of seats remaining in the section.

However, it is important for teams to always keep in mind their fan base - maximizing ticket revenue should not come at the cost of angering and losing the fan base.

Analytics methods used in this research: Demand Model, Negative Binomial Regression, Multinomial Logistic Regression
Descriptive, Predictive and Prescription Analytics

Three Phases of Sports Analytics

It’s a great day in sports analytics! Today, we will be looking at three critical phases of sports analytics: Descriptive, Predictive, and Prescriptive Analytics.

Before we go on, let’s briefly discuss what each of the analytics mean. Descriptive analytics is a method used by analysts to understand what has happened within the field in the past using data aggregation and mining to collect information.

Predictive analytics, on the other hand, uses statistical tools and models to provide insights into future events with the aim of making predictions.

The third, prescriptive analytics, employ the use of algorithms to optimize and simulate data and/or events with the goal of giving possible outcomes of an event.

Coaches and sport analysts employ these tools and integrate the three phases in order to understand past game events, performance of athletes or players, and the performance of a team.

Based on this understanding, they make predictions of future sporting events using predictive analysis, and then by employing the third phase, they utilize the algorithm to give possible outcomes, resulting in better decisions making.

All of these are brilliant ways of improving team performance, as they provide teams with insights, which would not be there by ordinary means of analyzing games. So let’s take a look at the three phases of this method, and how coaches and sports analysts apply them.

Descriptive Analytics

Coaches and analysts gather information about their sport and then sort out the performances of each team in their league, as well as the high ranking players. For a football analysis, for instance, it could be the top four teams, players with the most goals, coaching tactics over a period of the season, and so on.

All of these can be grouped into a column season by season to see the variation in these factors. The aggregation of this data will reveal patterns which will highlight the factors that are responsible for wins and for losses.

Predictive Analytics

Predictive analytics require the use of statistical tools and techniques. For this reason, coaches and sports analysts often consult experts who provide them with information on which certain statistical analysis can be carried out in order to make predictions about future events.

Having carried out the descriptive analytics, the factors that are deemed significant for the team or players’ performances are then used as data. This data serves as inputs in various statistical techniques and then enable the analysts to build models, which help them to determine the likelihood of one team performing better than another in future games.
Predictive analytics, as a method of sports analytics, is not only used by coaches and analysts, but also by professionals in the betting world in order to create virtual games.

**Prescriptive Analytics**

Prescriptive analytics is similar to the supervised learning method used in sports analytics. This is because it employs the use of algorithms, which are supplied to the machine by the analysts from certain factors that have been deduced from the players or teams.

The machine then optimizes and simulates the data and/or events, and provides the analysts with the possible outcomes. For instance, as we have explained earlier about football analysis and data that are used to reveal past trends in a football season, based on those factors considered, the machine can process the data and provide certain possible outcomes. This helps coaches understand the areas of deficiency in his or her team as well as strengths, and then work accordingly to improve the performance of the team.

In conclusion, these three phases of sport analytics—descriptive, predictive, and prescriptive—are often employed together as a tool in the performance analysis of a player or a team.
Econometrics

Econometrics is the quantitative application of statistical and mathematical models using data to develop theories or test existing hypotheses in economics, and for forecasting future trends from historical data. It subjects real-world data to statistical trials and then compares and contrasts the results against the theory or theories being tested. Depending on if you are interested in testing an existing theory or using existing data to develop a new hypothesis based on those observations, econometrics can be subdivided into two major categories: theoretical and applied. Those who routinely engage in this practice are commonly known as econometricians.

Source - https://www.investopedia.com/terms/e/econometrics.asp

Experience and Winning in the National Basketball Association

This is a review of the NBA research conducted by James Tarlow, applying econometrics.

Within sports, it is always assumed that experience improves a team and leads to championships. This is especially true in the NBA. Teams made up of younger players are viewed negatively for their lack of experience. When teams comprised of younger players losses a playoff series it is attributed to a lack of experience. Young coaches with less experience are criticized in the same manner.

Analysts and fans always seem to look at the experience of individual players and never the aggregate experience of the team. Teammates who have played together for a long period of time gain an innate sense of how their teammates will react during different situations. This improves team efficiency but to what extent?

An econometric study is undertaken to examine the relationship between experience and winning in the post season. It first looks at which, if any, types of experience have a positive effect on team performance and what extent that relationship has on winning games.

The study uses data from the 1979-2009 seasons including 4,020 players. Three models are developed using multiple least square regression, including a lagged dependent variable. The first model looks at player experience. Player experience is divided into two categories, NBA experience, and postseason experience. Variances of player NBA experience and player playoff experience are included in order to control for the diversity of player experience within each team. Coaching experience is examined in the second model and is also divided into NBA experience and postseason experience, but also includes the coach's winning percentage for postseason games and coach tenure. Lastly, the third model analyzes chemistry, which is defined as the number of years the five players playing the most minutes during the regular season have been teammates playing for their current team.

Results from the first model regarding player experience indicate that NBA experience does not contribute at a statistically significant level. However, postseason experience is a different story. Results indicate that player postseason experience helps a team make it into the playoffs but does not increase their ability to win postseason games.
Results from the second model regarding coaching experience are similar to those regarding player experience. Regular season coaching experience does not appear to have any effect on postseason success. However, having postseason coaching experience and the amount of that experience does appear to contribute to winning in the postseason.

The third model dealing with chemistry determines that chemistry is statistically significant in regards to postseason wins. Each year of shared experience between teammates increases the expected number of postseason wins.

All of this information helps coaches and analysts when making decisions regarding trades. Knowing the effects that the experience of player or a coach may or may not have regarding postseason success will help make wiser decisions. The chemistry model provides more potentially useful information as it indicates that teammates who play together for an extended period of time will get better and improve a team's efficiency. This would seem to be an argument against making trades simply because a team is not performing well at the moment. Leaving the current players together, rather than making trades, may prove to be a wiser decision in the long run.

Chemistry is not a variable that is typically considered by coaches when making decisions regarding trades, but perhaps it should receive more attention.

Analytics methods used in this research: Econometric Study, Multiple Least Squares Regress including a Lagged Dependent Variable
Egocentric Ghosting Model

The egocentric ghosting model generated data that is consistent with behaviors demonstrated by basketball players. It illustrates sequences of a player heading for the basket, then quickly changing directions. Evaluating the image showed that the sequence was very logical as it took into account attributes of the opposing player, including their positioning which would indicate the direction or move they planned to employ.

Source - https://www.agilesportsanalytics.com/sports-analytics-egocentric-ghosting-models/

Learning an Egocentric Basketball Ghosting Model using Wearable Cameras and Deep Convolutional Networks

This is a review of the egocentric ghosting model research conducted by Gedas Bertasius, Aaron Chan, and Jianbo Shi.

The use of data-driven ghosting models is increasing as player tracking data becomes more widely available. The issue with the models that have been previously constructed is that their analysis of the players' decisions is based solely on the tracking data. The only predictor of future behavior is the player's position on the court. However, a player in a game bases his decisions on more than just his own position; he also takes into account a variety of other factors. These factors include information such as where his teammates are located, where the defenders are located, and the physical characteristics of the defenders to name just a few. In order to account for this multitude of factors an egocentric basketball-ghosting model is developed using data gained from cameras worn by players during their games.

The data collected consisted of 988 basketball sequences from a game of one-on-one basketball played by nine college players. The one-on-one format provides better data for studying the decision-making process of basketball players. When a team is on the court, a player's decisions are often based on the decisions and time of the teammate who has the ball. In a one-on-one game, the players are constantly making the decisions themselves, as they do not have any teammates to rely upon.

All sequences were mapped on a single map, with blue dots indicating the beginning of a sequence and yellow dots indicating the end of a sequence. The map clearly showed that the sequence covered a wide variation of possibilities, therefore creating a strong database on which to build the model.

The egocentric ghosting model generated data that was consistent with behaviors demonstrated by basketball players. It illustrated sequences of a player heading for the basket, then quickly changing directions. Evaluating the image showed that the sequence was very logical as it took into account attributes of the opposing player, including their positioning which would indicate the direction or move they planned to employ.

Looking at various sequences allows a coach to determine where there are open spaces on the court that could be used by their players more effectively. The sequences could also be used as a coaching
tool to enable players to learn how to read their opponents more effectively. It would also aide in learning how to how and where to position themselves in different situations which would allow them to maximize their scoring potential. Coaches of younger players could use the sequences to teach their players how professional basketball players make decisions while playing games. The young basketball players could then improve their own skills by imitating the professionals.

Analytics methods used in this research: Egocentric ghosting model, deep convolutional networks
**Encounter Detection Algorithm**

With encounter detection algorithms, a visual analytics process to detect encounters between players and position data and an archive of historical position records is pre-processed and filtered to provide input. The algorithm arranges the position records into a set of sorted lists (SSL) so that only a minimum number of records need to be compared. The algorithm performs a single sweep over the record set to arrange it into a SSL and simultaneously find the encounters. To avoid problems due to discrete sampling, an interpolation of the data is performed when the sampling is too sparse. To accommodate large data sets, a divide-and-conquer approach using a sliding spatial window is developed. In post-processing, the elementary encounters are grouped into composite encounters by collecting elementary encounters occurring between the same players. Additionally, the composite encounters are input into a visual analytics tool where each composite encounter is represented as a layer on a map.

**Esports Analytics through Encounter Detection**

This is a review of the e-sports analytics research conducted by Matthias Schubert, Anders Drachen, and Tobias Mahlmann.

The field of e-sports, computer games played competitively, is growing around the world, with millions of competitors and millions of viewers. As a result, e-sports analytics have emerged, which relate to both sports analytics and digital game analytics. E-sports analytics have grown in demand to the point where most professional teams now hire analysts to analyze their opponents in order to gain a competitive edge.

Multi-Player Online Battle-Arena games (MOBAs) involve two teams of five players. Each player controls one character and the object of the game is destroying the enemy's structure. In order to win players must work together strategically in both offensive and defensive actions.

In order to analyze the game each action must be evaluated. The difficulty is that actions can be taking place at the same time in different areas of the game. Therefore, it is first necessary to break down the game into the separate actions. Each action is defined as when two or more characters from opposing teams are close enough to have an effect on each other. They involve both space and time.

Each individual player is labelled by position, team, and unit. At every point of time it must decide whether players are close enough together to interact. This distance depends upon the skills of each player. Those who fight hand to hand must be closer than those who shoot arrows. One action can possibly cause a following action, either immediately or later on, and this link is taken into account as well. This information is inputted into an algorithm to determine all actions that take place within the game.

Once each action has been recorded, the next step is to determine what started the action and the result of each action. The type of unit is involved in the action must be determined, as different unit types are capable of different actions. A combination of points earned and/or number of kills can
determine wins. Putting this together allows an analyst to determine the performance of each team during the action.

The dataset used for the experiments included information from 412 different matches. Through these experiments it is possible to determine which strategies are more likely to lead to a win. Strategies can include when to collect points and which type of points is the most effective, depending on the type of game. Information regarding how the majority of kills are made and how those kills affect the outcome can be analyzed. The time that each action typically requires is evaluated, as well as if each action is effective and leads to further actions or stalls at that point. Putting all of this information together allows analysts to evaluate a team's performance and develop strategies to improve their game play in order to have a better chance of defeating their opponent. Opposing teams can be analyzed in order to decide which type of actions at which point in time would be the most effective in countering their type of play.

Multi-Player Online Battle Arena games are constantly evolving so this field of analytics must be constantly evolving as well in order to take into account updated versions of games as well as new games as they are released.

Analytics methods used in this research: encounter detection algorithm
Euclidean Distances

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as the Pythagorean metric. A generalized term for the Euclidean norm is the L2 norm or L2 distance >


Acceleration in the NBA: Towards an Algorithmic Taxonomy of Basketball Plays

This is a review of the NBA acceleration research conducted by Philip Maymin, applying Euclidean distances.

Basketball coaches illustrate plays using x's and o's on a white board along with verbal instructions regarding the pacing of the play. This creates a barrier to creating a method to objectively measure execution and the contribution said execution has to winning a game. In order to combat this issue a dynamic algorithmic approach is created.

Optical tracking for 233 regular and post season games from the 2011-2012 NBA are analyzed for half-court situations which begin when the last player crosses the half court line and ends when the offense no longer has possession of the ball. This limitation is set in order to filter out improvised breaks and include only those possessions where the player's movements are the result of intentional practice. The resulting data set included 30,950 plays, each lasting an average of 7 seconds. The set includes the location of all 10 players as well as the basketball.

Acceleration is calculated as the second difference of the Euclidean distances between sequential moving average positions of a player on the court, divided by the standard of gravity.

Analysis shows that acceleration is rare and events of greater acceleration occur even less often. The frequency acceleration by position is charted with positions organized by average height of players in the position. Examining the chart shows that centers accelerate more than guards do. Bigger players tend to have more episodes of extreme acceleration. Players seem to accelerate based on who they are and not based on whom they are guarding. The spatial distribution of all players and the spatial distribution of accelerating players are quite different. Acceleration typically takes place in one of three areas: the paint, the top of the key, and the combined area of the extended elbows and wings. With each area on the right and left hand side of the court that comes to six locations where extreme acceleration typically occurs.

Proportions of co-accelerating players are illustrated in pie charts, with each chart based on a different acceleration rate. The lowest acceleration rate has the greatest proportion of all five players accelerating together, indicating that mild acceleration is the norm for most players the majority of the time. As acceleration increases, the likelihood of co-acceleration drops dramatically.
Analyzing the acceleration of each offensive and defensive player helps analysts define basketball plays, perhaps resulting in a playbook of plays for each team. Past data regarding acceleration could be used to predict future performances, helping scouts analyze plays performed by future opponents in order for coaches to create and practice the best strategical plays possible. Analysts could look for any possible relationships between acceleration and co-acceleration and field goal percentage, or any other statistic. Those relationships could then be used in training players how to best implement acceleration and co-acceleration into their plays to make them even more effective.

Analytics methods used in this research: Euclidean Distances, Dynamic Algorithm
Event Tree Models

An event tree is an inductive analytical diagram in which an event is analyzed using Boolean logic to examine a chronological series of subsequent events or consequences.

Accounting for Complementary Skill Sets When Evaluating NBA Players' Values to a Specific Team

This is a review of the NBA research conducted by Joseph Kuehn applying event tree modelling and conditional probability.

A key to the success of any NBA team is the ability to put together a lineup of players who work well together. Success depends on the whole being better than the parts.

A framework is developed to determine the value an individual player contributes to a particular lineup as well as which players help their teammates' production and which hurt their production.

Play-by-play data incorporated included the 250 players with the greatest number of possessions from the 2014-2015 NBA season, including all 10 players on the court and a detailed result of the possession. A basketball game is modelled as a series of possessions with each possession being a series of events, which have the potential to gain points for the offensive team. Each possession is represented on an event tree model. The event tree model uses probabilities for each action taken by each of the offensive players on the court to calculate the expected number of points associated with each action and the expected number of points for each possession. This demonstrates how the probabilities of particular actions taken by a team affect the expected number of points per possession.

Conditional probability is used to model individual players taking into account the player, his teammates, opposing players, and the particular event. In the first step, the maximum possibility is used to map observed probabilities into three player scores, which measure the player's tendency to accomplish an event, the player's effect on teammates' abilities to accomplish an event and the player's defensive effect on his opponents. A least-square approach is used to match the scores created in step one with a player's rating is utilized in the second step.

Substitute players can be inserted in a lineup using the above probabilities to determine what effect that player has on the lineup. Different strategies can also be substituted for other plays to determine the effect that has on the expected outcome of the possession.

The player model also provides information to let coaches and analysts know which actions players have a tendency to fall back on in any given situation, the probability they make or miss a shot from a particular location and the probability they get an offensive rebound from a missed shot from a particular location. Defensive player ratings tell coaches how a player affects the probability of the actions taken by the opponent in these situations.
Coaches can use this information in putting together lineups of teammates who complement each other, increasing their effectiveness both offensively and defensively. Players that are not as highly skilled as others often play to a higher level when put on lines that work to their strengths. Highly skilled players can play below their capabilities when paired with non-complement teammates. This knowledge enables coaches to put together the strongest line-ups possible, building on their players’ strengths and minimizing weaknesses. Coaches and analysts, when evaluating individual players, can also use this information. It is important to assess the individual for their own abilities and not those gained while working with their teammates. The abilities of the individual player will be transferred to a new team while skills gained from working with particular teammates will not be transferable. Players can be evaluated as to whether they increase or decrease their teammates’ expected points. All of this allows better trade and draft decisions.

Analytics methods used in this research: Event Tree Model, Conditional Probability, Least Squares

**Expected Possession Value**

Using film or player tracking data, a quantitative analysis representative of the entire possession will occur where each moment in the possession is summarized on the basis of the expected value of the possession at each moment. For each moment of a possession a value is assigned to each of the individual tactical moves a player can make, allowing analysts to evaluate each decision that a player makes. Under this model, dishing the ball to an open shooter at the key or near the basket is worth more expected points than to a covered player in the corner. Expected Possession Value is an extremely new avenue of basketball analysis focusing on decision-making, opportunity creation and prevention.


**POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data**

This is a review of the optical tracking research conducted by Dan Cervone, Alexander D’Amour, Luke Bornn, and Kirk Goldsberry.

The key moment of an offensive play in a basketball game may not be when the points are scored at the end of the possession. It may have occurred earlier in the possession with a pass or a move to elude a defending player. The points could not have been generated without the offensive strategy that put the ball in the shooter’s hands. However, traditional basketball statistics do not take this into account. Stats tend to look at quantifiable data like points, rebounds and turnovers. These evaluate the skill of the shooter but fail to take into account the skills of the players whose actions lead up to the shot.
In Pointwise, an optical tracking framework is built using player tracking data to develop a quantitative representation of the entire possession as a series of summaries for each moment of the possession in terms of the number of points the offense is expected to score – expected possession value or EPV. The model is able to determine how the ball handlers make decisions based on where the players are positioned on the court. Every possible option a player has is given a point value in order to evaluate how each move added to the possession. It is also possible to look at alternative moves the player could have made to determine if they made the best choice.

This possession model using optical tracking data gives coaches the ability to estimate the probability that the player will make a particular choice in a particular situation as well the resulting EPV of the resulting possession. Player’s options include single moves such as passing or shooting as well as longer moves such as moving to the left or right.

EPV is calculated continuously through the possession, increasing and decreasing as the expected value changes based on the actions of the players on the court and whether these actions are more, or less, likely to lead to points being put on the scoreboard.

Analysts can evaluate a player’s value in certain situations by replacing him with another player and observing how the EPV changes. Plays can be analyzed to determine if a player passing the ball creates a higher EPV than taking a shot. This allows analysis of whether the player is making good choices for the team or selfish choices to pad his point total. Coaches can look at the plays their team typically employ during a game to decide if those plays are generating a high enough EPV or if they should be modified.
Expected Rewards

The ultimate goal of decision making is to find an optimal behavior subject to some optimality criterion. Optimizing for the infinite-horizon expected discounted total reward is one of the most studied such criteria.


How to Play Strategically in Fantasy Sports (and Win)

This is a review of the expected rewards in fantasy sports research conducted by Martin B. Haugh and Raghav Singal

Daily Fantasy Sports is an ever-growing industry with millions of users participating each year. DFS covers a wide variety of sports including football, basketball, baseball, soccer, and golf. Each competitor puts together a fantasy team of real world players within a designated league. Typically, there are restraints such as budget; a constraint that real world teams face as well. Additional constraints include positional constraints, as the user is restricted to a certain number of players chosen for each position. The users construct portfolios putting together a team they feel has the best expected outcome of winning. These users can be split into two groups - those who use statistics to form their portfolios and those who do not. The success of a user's portfolio is determined by the success of the real players in their actual games.

Success is based on the amount of money the user is awarded which depends on the payoff structure. One type of payoff structure is the double-up payoff structure. In this configuration, the top number of players each receives an equal payoff. A second type of payoff structure is the top-heavy one. Within this framework, the amount of the cash payoff increases with the ranking of the portfolio. How can users optimize their decisions when putting a portfolio together? Optimal decisions will ultimately lead to a higher expected reward.

While a user's success is dependent upon the success of the real world players, it is also dependent on the outcomes of their competitors. Final rankings are based on well users did compared to the hundreds or thousands of competitors who put together teams for the same league. Therefore, optimizing one's portfolio includes not only choosing the best combination of players but also choosing the combination that will outplay the teams put together by other competitors.

The first step to answering this question is to determine a ranking system for portfolios. This ranking is accomplished by looking at the performance of a user's portfolio against the performance of their competitors' portfolios. The portfolios are ranked based on their point's total and cash payoff.

In order to be more effective in putting teams together a user must have some idea of the team his opponents will put together. Analyzing the teams an opponent has put together on prior occasions will help predict teams they will put together in the future.
Research was conducted using DFS contests on FanDuel during the first 12 weeks of the 2017-18 NFL season. The expected rewards outcomes clearly indicate that portfolios that take into account their predicted competitors' portfolios perform better than those that do not. How much better they perform of course depends on how accurate the predictions regarding the competitors' portfolios actually were.

Analytics methods used in this research: expected reward, ranking
Experiments, Events, Probabilities and Odds

Applying Experiments, Events, Probabilities and Odds to Sports Analytics

This chapter briefly outlines the theory of probability as an approach to sports analytics. An understanding of the theory of probability is important to properly apply it in the field of sports. Coaches and analysts employ probability as a tool to determine what aspect of their team sporting activity need improvement. Their aim of doing this is to increase their chances of winning a competition as well as improving their abilities.

This method starts with an experiment. In this context, we describe an experiment as any course of action where the outcome is random. An experiment can be very general or more specific. An example of a general experiment is a football game while a quarterback pass to a particular receiver can be explained as a more specific experiment. These two types of experiments are widely used in the analytics of sports.

After an experiment has been determined and chosen, the next step is to look at specific events. Events are the outcomes of the performed experiment. In a football game for instance, the event could be the final score, the number of passes attempted, or the number of passing yards.

In terms of the quarterback throwing a pass experiment, the event could be whether the pass attempt was a success or failure and the number of yards gained.

Events have corresponding probabilities. Probability can be explained as a measure of the likelihood that a particular event will happen when the experiment is done repeatedly. Probability refers to what will occur in a theoretical situation. If a coin is tossed for an infinite number of times, 50% of the time, a head shows up. However, if a coin is tossed only twice, there is no probability that a heads will be flipped once and a tails twice.

The probability that something will not happen can be determined by subtracting the probability that the event will happen from 1. The chance that an event will happen is most often referred to as probability.

However, in some cases, it is easier to use odds rather than probabilities. Odds can be determined by taking the probability of that event happening and dividing it by 1 subtracted by the probability of the event. In cases where the probabilities are very small or very large, odds are preferably used because too small or too large probabilities are often difficult to analyze.

When odds are applied in the world of sports, they can indicate relative difficulties of an event. This is because, the greater the odds, the harder it is for success to be achieved and less likely it is to occur. Sports analysts utilize probabilities and odds to make predictions concerning outcomes of games and the performance of various players during the game of sports.
Extra-Trees Classifiers

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.


This is a review of the sports metrics research conducted by Shouvik Dutta, Sheldon H. Jacobson, and Jason J. Sauppe applying Extra-Trees Classifier and Balance Optimization Subset Selection.

The National Collegiate Athletic Association annually holds the men's college basketball championship tournament known as March Madness, which attracts attention from around the country. People cheer for their favorite team and attempt to predict winners of each game. In 2015, approximately 70 million brackets were submitted to various competitions hosted by different sites, cumulating in a total of around $9 billion. Pride and financial gain are strong motivators for developing models to predict the outcomes of the various games played in the tournament.

A technique is proposed to select potential upsets in the tournament by identifying match-ups set for the tournament that exhibit similar characteristics to upsets in past tournaments. The differences in the season statistics between the two teams in each past tournament upset are used to build a profile, which is then compared to first round games in the upcoming tournament.

Examining upsets that occurred over past years allows the model to identify characteristics that allow a weaker team to defeat a stronger team. Looking for these characteristics in upcoming games identifies potential upsets.

This method defines an upset as a team seeded 13, 14, or 15 winning a game in the round of 64. The 16th seeded teams are excluded as to date no such team as ever won a game in the tournament. Between the years 1985 and 2015, 52 of the 372 games played resulted in the teams seeded 13-15 upsetting their opponents, an average of 1.7 upsets per year. Therefore, the aim is to find two possible upsets for the current tournament. The model involves four steps.

The first step is computing match-up statistics done by comparing the two teams playing in each game using team statistics. Observing which statistics have gaps which lead to upsets in the past allows future games to be identified that may also result in an upset.

The second step is identifying useful match-up statistics. An extra-trees classifier creates decision trees, using the different match-up statistics that are strong indicators of past game upsets. Fifteen statistics were determined to be important features including effective possession ratio, games played, extra score chances per game, opponent floor percentage, and percent fouls per possession as the top five.

The third step is finding similar match-ups using Balance Optimization Subset Selection to identify three match-ups that are similar to past upsets in regards to the match-up statistics. The control pool consists
of the teams seeded 13, 14, and 15 while the treatment group is made up of past upsets. The control groups chosen would be those most similar to the statistics of past upsets.

Finally, the fourth step narrows down the three choices from the third step to two by identifying combinations of match-up statistics that were strong indicators of upsets in the past. The combinations are identified by evaluating the performance of each combination of statistics using historical data and choosing those with the best past performance. Performance is measured as the number of upsets selected correctly over the entire range of past data.

The technique is not 100% accurate, which should be obvious as the definition of an upset indicates that it is something unexpected and therefore cannot be predicted precisely. However, this provides coaches and analysts with a unique tool to look for possible upsets within the March Madness Tournament. If coaches of underdog teams find that their match shows a stronger possibility of an upset, they could use this as a strong motivating factor for their players.

Analytics methods used in this research: Extra-Trees Classifier, Balance Optimization Subset Selection
Frequencies

In statistics the frequency (or absolute frequency) of an event is the number of times the event occurred in an experiment or study. These frequencies are often graphically represented in histograms.


No Referee Bias in the NBA: New Evidence with Leagues' Assessment Data

This is a review of the NBA referee research conducted by Christian Deutscher.

The National Basketball Association hires referees to be impartial judges. However, refereeing is by its nature subjective, which means it is possible that there is some bias involved. Biases may be based on favorite players or teams, hometown fans cheering for their team, or in some cases, bribes.

In an effort to ensure that referees are making fair calls, the league spends a great amount of money to monitor their actions. Promotions and job security is based on this monitoring, providing an incentive to the referees to remain objective. Referees have a difficult job in that the expectation of the league, players, and fans are all different.

In order to investigate the possible presence of any bias in play calling, differences between actual calls and league judgement is analyzed. These differences are compared for each individual call made on the court, which can be sorted by each player involved. Knowing which players are involved allows for studying any possible biases regarding individual differences between the players.

The NBA reviews all calls made by referees during crucial game situations, which is defined as games in which the point differential is at most five points with less than two minutes to play, or overtime. A senior referee manager or basketball operations manager reviews each call and these reviews are posted the day after the game. 1229 calls from the 2014-2015 NBA regular season are examined in this study.

The first step for assessing referee bias is finding the frequencies. In this case, 496 out of 619 fouls the league identified were called correctly by the referees and 593 of 610 no-foul situations the league identified were not called by the referees. However, the variations between players in the no-foul situations were too small so this data was excluded from the remainder of the study.

In order to look for possible biases additional information is included regarding each occurrence. This includes whether the call was against the home or away team, if the player committing the foul or the player being fouled is considered a superstar, if the player is American or not, and which team is considered to be the underdog. Also included is the time remaining in the game when the call is made and crowd attendance. Crowd attendance is included as ardent fans cheering for their team could possibly affect a referee's decision-making process.

Classifying the fouls by these controls resulted in 19.7 percent of the fouls being classified as including superstars and 38.1 percent including a non-American player.
To test for referee bias a logic approach is applied in which the independent control variables deals with potential biases towards home teams, superstar players, players of US origin and favorite teams. The time remaining in the game and crowd attendance were determined to be insignificant and thus left out.

The results indicate that there is little referee bias in the NBA with the only resulting bias being referees showing a weak preference for the underdog.

This study is not infallible as it relies on the judgements of league employees who may be unwilling to release information that shows the league and its referees in a bad light. In addition, limiting the analysis to crucial game situations provides a bias in itself. As the NBA allows video reviews during the latter part of the games, referees would naturally be more prone to ensure their calls were not biased.

Refining this information to include all calls made in a game would provide the league with information regarding possible biases of the referees. Knowing this, they could implement training specific to each individual referee.

Analytics methods used in this research: Frequencies, Logic Approach
Frequency Distributions

Frequency distribution is a table that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

Source - https://www.tutorialspoint.com/statistics/frequency_distribution.htm

Frequency Distribution as a Method of Sports Analytics

The use of data in sports analytics is becoming more and more popular as new methods are discovered each day. Today, we will discuss the application of frequency distribution in sport analytics.

In frequency distribution techniques, coaches and analysts employ the use of a frequency table to organize data. This table provides information such as number of games played, number of wins and losses and scoring statistics. A frequency table is a typical win/loss chart.

*Why frequency distribution method?*

A frequency table provides information in a manner that can be easily read and understood. Users can simply read along the row for the desired information.

*How it works*

In the frequency table, there are columns for number of wins and percentage of wins. The number of wins indicate the frequency of wins and the percentage of wins is the relative frequency.

Relative frequency is calculated by dividing the total number of wins (of a team e.g. football team) by the total number of games played. This way, discrepancies in the number of games played per team are not seen, thus making it easier to compare the teams in question.

Frequency tables measure qualitative (non-numerical) or quantitative (numerical) data. Qualitative, also called categorical data are described as discrete variables - they have limited number of outcomes and cannot be ranked. Number of wins is discrete, meaning that they are whole numbers as you cannot have 3.5 or 1.5 wins.

Qualitative variables can be either discrete or continuous. Continuous variables can take on any value, whole number or not. For example, the weight of a wrestler is a continuous variable as a wrestler can be 190 or 190.5 pounds.

Frequency tables are usually viewed in the form of graphs, called histograms.

Histograms make it easier to compare disparate data and look for similar patterns. They can be presented by shapes, the normal or bell curve (like the shape of a bell). The normal curve is interpreted as what happens theoretically under ideal conditions. Therefore, no data will ever form an exact bell curve.
Frequency tables or histograms can be compared by examining their symmetries. Normal distributions are completely symmetrical with the mode, or highest point, in the middle. When the highest point is facing the left side of the graph, the graph is said to be positively skewed and if the highest point is towards the right side, it is negatively skewed.

In conclusion, frequency tables and histograms have proven to be a quick and convenient way for sports analysts and teams to compare individual players as well as teams.

This method can also be used to compare the stats of a single player. For example, numbers of yards per pass for a quarterback can be listed by category such as 0 to 5 yards, 6 – 10 yards, 10-15 yards etc. The use of a histogram chart is also a helpful visual to better illustrate the strength and weaknesses of a team.

**Frequency Distributions as a Method of Sports Analytics**

Once data has been collected, it needs to be organized in a manner that makes it easy to analyze. The simplest way to do this is by summing up the information and entering the sums into a frequency table. A typical win/loss chart for any team or league is a frequency table. They are easy to read and can be interpreted by simply reading along the row for the desired information. Frequency tables provide information such as number of games played, number of wins and losses, and various score statistics.

When looking at the table the number of wins is the frequency of wins and the percentage of wins is the relative frequency. Relative frequency is found by dividing the total number of team wins by the total number of games played. This takes care of any discrepancies in the number of games played per team, thus making it easier to compare the teams.

Frequency tables can measure qualitative (non-numerical) or quantitative (numerical) data. Qualitative, or categorical data are what we call discrete variables, meaning they have a limited number of outcomes and cannot be ranked. Number of wins is discrete, as you cannot have 1.5 wins; it must be a whole number. Qualitative variables can be either discrete or continuous. Continuous variables are those that can take on any value. A fighter’s weight is a continuous variable as a fighter can be 185 or 185.5 pounds.

Frequency tables can be viewed in graph form, called histograms. Histograms make it easier to compare data and look for patterns. However, if you are looking for specific numbers, they are more easily found in a frequency table.

Histograms can take on many different shapes. One shape is the normal or bell curve - a curve that looks like the shape of a bell. The normal curve is what happens theoretically under ideal conditions. Therefore, no data will ever form an exact bell curve.

One method of comparing histograms is to look at their symmetry. Normal distributions are completely symmetrical with the mode, or highest point, in the middle. When the highest point is towards the left side of the graph, the graph is said to be positively skewed and if it is towards the right side, it is negatively skewed.
Histograms with only one peak are called unimodal while those with two peaks separated by a valley are bimodal.

Frequency tables and histograms are a quick, easy way for analysts and teams to compare individual players or entire teams. They are also used to compare the stats of a single player. For example, numbers of yards per pass for a quarterback can be listed by category such as 0 to 5 yards, 6 - 10 yards and so forth. Histograms are a helpful visual to illustrate what they are discussing with the players or other stakeholders.
Gaussian Copula Model

In probability theory and statistics, a copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. Copulas are used to describe the dependence between random variables.

The Gaussian copula is a distribution over the unit cube. It is constructed from a multivariate normal distribution by using the probability integral transform.

Source - https://en.wikipedia.org/wiki/Copula_(probability_theory)#Gaussian_copula


This is a review of the research conducted by Alexander M. Franks, Alexander D’Amour, Daniel Cervone, and Luke Bornn.

Data analysis is widely used in all sports, having become an essential tool for all stakeholders. Decision makers, to determine optimal decisions regarding their team and the probability of success, use statistical models. In response, the number of models has grown rapidly, resulting in areas of overlap and confusion as to which model is the most appropriate to use in any given situation.

In response to this issue meta-metrics are developed to describe which models provide the most useful and reliable information. The methods are easy to use and applicable to any sport. Metrics are evaluated based on stability, discrimination, and independence with a meta-metric developed for each category. The stability meta-metric measures how much a metric varies over time due to context and player skill changes. The discrimination meta-metric looks at whether the metric differentiates among players. Finally, the independence meta-metric indicates whether the metric provides new information not offered by other metrics.

The meta-metrics are all developed based on R-squared statistics as a function of the variances along the three dimensions.

The discrimination meta-metric describes variation within a single season and compares the average intrinsic variability of a metric to the total between-player variation in the metric. Most variability between players is a reflection of the true variation in player ability and not some chance variation.

Stability looks at how individual players’ metrics vary from season to season. The stability metric measures how much an individual player metric varies over time after removing chance variability. It specifically looks at the sensitivity of a metric to changes in context or intrinsic player skill over time.

Determining dependence among metrics is accomplished with a Gaussian copula model, which examines dependencies between metrics with a latent multivariate normal distribution.

In order to understand how sports models can be used it is necessary to quantify the sources of variability and how they are related across metrics, players, and time. The three sources of variation
looked at are intrinsic player skill, context and chance. Examining sources of variations allows them to be groups into two categories, signal, and noise. Those placed in the signal group will influence the decision making while those in the noise group are not included.

The meta-metrics are used to examine statistics used in the NBA and NHL, including both traditional and advanced statistics. Statistics such as rebounds in the NBA and hits in the NHL are the most discriminative and stable due to the large variance between players. Stats based on total minutes played are very discriminative but less stable, while statistics looking at per-minute or per-game measurements are less discriminative but more stable. Rate based statistics are more useful in estimating player skill while total based statistics are more useful for identifying overall end of season contributions. Steals in the NBA and takeaways in the NHL provide the most unique information in comparison to the others.

The most obvious use of meta-analytics for coaches and analysts alike is to evaluate the wide variety of models available and determine which will provide them with the most accurate results for guiding their decision-making. This would increase decision-making accuracy, leading to wiser decisions and hopefully a competitive advantage over the opponents. When the time comes to make trades, the discrimination meta-metric provides guidance as to which models should be incorporated to compare players’ ability. Looking at stability also becomes important at this time as teams want to invest in players who have remained consistent over the season, an indication that their consistency will carry over to a new team.

Analytics methods used in this research: R-Squared, Gaussian Copula Model, Latent Multivariate Normal Distribution, and Meta-Metrics
Generalized Linear Spatial Regression Model

In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.


The Pressing Game: Optimal Defensive Disruption in Soccer

This is a review of the soccer analytics research applying a generalized linear spatial regression model, conducted by Iavor Bojinov and Luke Bornn.

Soccer is the most popular sport in the world and the Barclays English Premier League has the largest fan base of all soccer leagues. Soccer is a dynamic game dependent on team strategy and individual player skills. Offensive and defensive strategies are extremely important in bringing teams to ultimate success.

At the end of every season analysts, coaches and fans analyze the teams and how the season played out, typically using summary statistics like who scored the most goals or who saved the most goals. However, in all of the analysis it is important to remember that soccer is basically a spatial game. To deal with this aspect summary statistics that quantify a team's ability to retain possession of the ball and to disrupt the opposing team when they have possession are created. After this, a map is created that lays out the strengths and weaknesses of a team's offense and defense.

A generalized linear spatial regression model is used to determine the average disruption surface over a season, mapping areas having a high probability of a disruption and areas having a low probability. A disruption of the attacking team is an action taken by the defensive team that leads to an interruption of the flow in play, including such plays as interceptions and tackles. From there the conditional probability of a disruption at a given location is determined.

Typically, teams that finish in the top half of the league have the highest controlling coefficient and an average disruption coefficient. The number of shots taken by a team and the value of both the control and disruption coefficients are clearly positively related. Intuitively, this makes sense, as a team that retains possession of the ball longer will ultimately have more shots. Next, maps are generated for a team's defensive disruption surface and one for their offensive control surface.

Coaches can use the maps to gain an understanding of their team's weaknesses and employ methods to correct them. They can also use the maps to determine their opponent's weaknesses and determine strategy to take advantage of those weaknesses. Teams with a weak defense will have a disruption surface below average. Disruption surfaces map where a team presses and where they are more passive. This provides coaches with information regarding the vulnerabilities of their opponent and how
to strengthen their own weaknesses. Looking at the control and disruption coefficients of their team allows coaches to determine their team's weakness and strengths as well as their opponents and use this in strategizing for the game.

Analyzing team disruption and control surfaces provides analysts with the opportunity to study what impact a coach has on a team's playing style. When a team changes coaches, the team typically does not instantly change their playing style, so any major changes made over time are due to strategic decisions made by the new coach. Consequently, the tactics of coaches can be analyzed from year to year and team to team.

Analytics methods used in this research: Generalized Linear Spatial Regression Model
**Graph, Slope, Correlation**

*Graph* - a diagram representing a system of connections or interrelations among two or more things by a number of distinctive dots, lines, bars, etc.

Source - [https://www.dictionary.com/browse/graph](https://www.dictionary.com/browse/graph)

*Slope* - In mathematics, the slope or gradient of a line is a number that describes both the direction and the steepness of the line.[1] Slope is often denoted by the letter m; there is no clear answer to the question why the letter m is used for slope, but it might be from the "m for multiple" in the equation of a straight line "y = mx + b" or "y = mx + c".


*Correlation* - Correlation is usually defined as a measure of the linear relationship between two quantitative variables (e.g., height and weight). Often a slightly looser definition is used, whereby correlation simply means that there is some type of relationship between two variables. This post will define positive and negative correlation, provide some examples of correlation, explain how to measure correlation and discuss some pitfalls regarding correlation.

Source - [https://www.displayr.com/what-is-correlation/](https://www.displayr.com/what-is-correlation/)

**A Development Model to Guide the Recruiting of Female Shot Putters at the NCAA Division 1 Championship Level**

This is a review of the female shot putter research conducted Donald G. Babbitt.

The NCAA Division 1 level of sports is highly competitive; resulting in the need for coaches to accurately determine which high school athletes will provide the greatest benefit to their collegiate sports program. The pressure is especially high for track and field coaches as they have a limited number of scholarships that is lower than the number of different sports competitions within their program. Thus, they need to be extremely careful in making their choices.

To date, sports analytics are extensively used in team sports as well as individual sports such as swimming and gymnastics. Analytics employed within track and field have traditionally been focused on the running events.

Shot put is a highly technical sport resulting in athletes requiring a longer period to maximize their potential. As a result, coaches need to take extra care in choosing their future athletes as their most successful years are likely to occur after they graduate. The ability to accurately assess the length of time it will take a shot putter to contribute a score at the NCAA Division 1 level would be very helpful.
This study is based on the performances of sixty-three top female shot putters from the NCAA Division 1 level from the years 2012-2017. The data from their best high school performances and collegiate performances were graphed with high school performance as the y variable, and collegiate performance as the x variable. Five graphs were generated comparing high school performance with each year of university performance. The equation of the slope of these graphs indicated the projected rate of improvement. Correlation coefficients were then calculated for each graph.

From this information the minimum high school performance was calculated which correlated with success in each of the five years of university. The distances recorded for first through eighth place in both the indoor and outdoor championships in each of the years from 2012-2017 were averaged. These averages were inserted into the slope equation as the x variable allowing for the calculation of the y variable, or the minimum high school performance required to achieve success at each year of the collegiate level. Calculations were completed for each of the 8 positions during the 5 years with results displayed in a table format allowing for easy analysis.

Results indicated a strong correlation between high school and the first three years of collegiate performance. This correlation decreased each year with the correlations for the fourth and fifth year of collegiate performances becoming insignificant.

Coaches and analysts can utilize this information when scouting high school talent and deciding which athletes should be awarded scholarships. The information can also be used to determine during which years female shot putters' performances improve the most. This would aide in determining regimes appropriate for each year that would maximize the potential of the athletes.

This analysis does have limitations, as it does not take into account the number of years of training each high school athlete had gone through. In addition, it can only predict the potential of each individual, which is not a guarantee of actual performance.

Analytics methods used in this research: Graph, Slope, Correlation
Graphing, Stratification, Correlation, and Causation

Graphing - the activity of making graphs to show how different information is related: On-screen graphing provides an instant analysis of your data. The spreadsheet has a graphing option.

Source - https://dictionary.cambridge.org/dictionary/english/graphing

Stratification - Stratification means arranging something, or something that has been arranged, into categories. By the time you figure out the complex social stratification of your high school class, from jocks to nerds and everything between, you’re ready to graduate.

Source - https://www.vocabulary.com/dictionary/stratification

Correlation - Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

Source - https://whatis.techtarget.com/definition/correlation

Causation - Connection between two events or states such that one produces or brings about the other; where one is the cause and the other its effect. Also called causality.

Source - http://www.businessdictionary.com/definition/causation.html

Fast Starters and Slow Finishers: A Large-Scale Data Analysis of Pacing at the Beginning and End of the Marathon for Recreational Runners

This is a review of the graphing, stratification, correlation, and causation research conducted by Barry Smyth.

The marathon is a grueling 26.2 miles and is often considered the iconic endurance event. Every year millions of runners, both elite and recreational, participate in marathons around the world. The goal for some is just to finish the race, for others it is to beat their personal best, and for a few, it is winning the race or breaking a record. Participants consider their pacing to be a crucial aspect of the race and carefully plan in advance what their pacing strategy will be. However, in the heat of the competition excitement often beats out strategy. While pacing has been studied in regards to elite athletes, there is no data regarding recreational runners.
The data collected included 1,724,109 complete race records of recreational runners, covering 12 years, 64 races, and 11 cities. Data included age, gender, 5km split-times throughout the race and final finish time.

Data is graphed of mean finish-time versus relative start pace. The resulting graph clearly indicates that running the first 5 kilometers at a rate faster than average is very costly at the end of the race. A starting pace 10% faster than the average adds approximately 37 minutes to the average finish time. Starting 10% slower adds approximately 29 minutes.

While this data illustrates a strong correlation it does not necessarily represent causation. To test for this the runners are stratified into three different finish-time bands. For each band, the percentage finish-time cost is calculated. All three bands hold to the idea that starting too fast or too slow adds time to their race, as compared to an evenly paced start.

To further investigate causation the focus turns to runners who run three or more races, looking at the circumstances that lead to a personal best time. From this data, it is determined that runners typically achieve their personal best times when they start at a pace within 5% of their average.

Analyzing pace at the end of the race, the average percentage of runners and their final pace is graphed. Approximately 58% of runners finish at a pace slower than their average, 19% finish at a faster pace, and 21% finish at their average pace, indicating that the pace at which a runner finishes the race is a strong indicator of how well the runner managed their pace throughout the race. However, the cost associated with finishing at a fast or slow pace is less significant than the cost associated with a fast or slow start.

Coaches can use this information to help their runners determine optimal pacing during a marathon race that will provide the greatest opportunity to meet their goal, whether it is to finish the race, achieve a personal best, or to win. Understanding that strategy often falls by the wayside when the race actually starts indicates that runners need additional tools to help them overcome this natural inclination.

Analytics methods used in this research: Graphing, Stratification, Correlation, and Causation
Grey Relational Analysis

Grey relational analysis (GRA), also called Deng’s Grey Incidence Analysis model, uses a specific concept of information. It defines situations with no information as black, and those with perfect information as white. However, neither of these idealized situations ever occurs in real world problems. In fact, situations between these extremes are described as being grey, hazy or fuzzy.


Ranking Regular Seasons in the NBA’s Modern Era Using Grey Relational Analysis

This is a review of the NBA research conducted by Sean Pradhan, applying Grey Relational Analysis.

Winning is the goal of all sports teams. While the goal is straightforward, there are many barriers in the way, from restrictions on financial resources, scarcity of talented players or difficulties signing rising stars from the free market. Sports analytics has grown in an effort to determine solutions to these issues. One benefit of analytics is providing teams with the ability to maximize player performance under the shadow of a salary cap. However, not all teams see analytics as offering practical solutions to problems. This study’s aim is to assess players in order to provide those practical applications.

A performance classification method is created by sampling current and former athletes from the NBA. A grey relational analysis is used to build a bridge between theory and practical application. Specifically, player statistics are used as the criteria to determine the most successful regular season achieved by an NBA player.

Grey systems theory is a process of studying problems involving a small sample and limited information. It accounts for both the known and unknown, using a matrix to quantify a players’ performance. The system involves several steps to normalize the data and convert the information into grey relational grades, which then generate a ranking system to determine which factor has the greatest influence. The first step is arranging the data into a comparison matrix based on performance indicators. From this, a referential series point is chosen.

The data in the comparison matrix is then normalized by determining the difference between observed values and the referential series points. Then a grey relational degree, which is a measure of similarity between a series of data, is obtained. These grey relational degrees are changed into difference scores in order to determine the difference between the referential series and observed data points. These difference scores are then changed into grey relational coefficients, which indicate the level of relationship between the idealized and actual sample data. From this the grey relational grade is calculated by adding together each indicators’ coefficient scores within a group and dividing by the ratio of the number of indicators within a group to the total number of indicators used in the analysis. The final step is to rank the data, indicating which factor reflects the best performance within the sample.

Within the analysis, both traditional and advanced statistics were incorporated. Traditional stats included such information as games played and games started, field goals made, and the list goes on. Advanced stats consisted of true shooting percentage, free-throw attempt rate, to name just a couple.
Twenty-five traditional and 20 advanced statistics were used in total. The results indicated that traditional statistics were slightly stronger indicators of performance than the advanced statistics.

The GRA and scouting can be used in conjunction with each other when determining which prospects the team should focus on. Being able to determine the efficiency of current and prospective players will provide scouts with better information to discern which players are better prospects. This information can also be used as an aide during contract negotiations, offering insights as to appropriate salaries and contract length. It also provides a method for comparing similar players within a team or across teams.

Coaches can analyze their team to determine their strengths, weaknesses and which players are contributing to each of those areas.

Often, there is a gap between theory and practical implications and this is an example of how that gap can be bridged.

Analytics methods used in this research: Grey Relational Analysis
Heat Maps

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. "Heat map" is a newer term but shading matrices have existed for over a century.


From Sports to Science: Using Basketball Analytics to Broaden the Appeal of Math and Science Among Youth

This is a review of the STEM sports science research conducted by John F. Drazan, Amy K. Loya, Benjamin D. Horne, and Ron Eglash.

Sports analytics give us a tool that can be used to reach out and help youth people gain skills through Science, Technology, Engineering, and Mathematics (STEM) education that will help them throughout their lives. It is possible to introduce data gathering and statistical analysis to students while they work to improve their own skills in sports through STEM sports science. Many young people have a dream of playing professional sports, however, statistics show that only a very few will attain this dream. It is possible to harness this desire to improve their ability in sports to teach them other skills that prove useful in other more attainable job opportunities.

Analysts have been working to make NBA statistics easier for people to understand. One method is graphical representations, specifically shooting and defensive efficiency maps. Basic stats such as shooting percentage and scoring efficiency can be gathered the old-fashioned way, writing it down and working out the basic arithmetic. This can then be combined with heat maps, which show the students how they need to improve their shooting skills, where they should be shooting from, and how they can become players that are more effective.

The players involved in the study took part in a shooting clinic. The leaders introduced basic statistics and then the students practiced taking shots from various positions on the court and recording their results as to how many baskets they made from each location. The data from their results was entered into a shooting analytics program to build individual heat maps.

Two separate heat maps were developed for each player. The first showed baskets made per location. Colored dots represented the percentage of baskets made at each location. This gave the students a visual representation of where they were strongest in their shooting ability and where they needed more practice.

The second map showed point efficiency. The colored dots indicated the point value of the shooting percentage at each location. Players could see from which position they were able to score the most points and use that information to their advantage in their games. Coaches could use this to discuss offensive and defensive strategies based on the players' strengths and weaknesses. They could also use this type of map to evaluate the strengths and weakness of their players and build programs specifically designed to help each player improve in their weaker areas.
The participants completed a survey prior to the shooting exercise and again after the discussions regarding the heat maps. The survey indicated that students had a positive response to the study and an improved perception of science topics. The participants also gained confidence in their knowledge of their strengths and weaknesses, and a greater understanding of how their shot performance varied, depending on where they stood on the court.

This approach would be a valuable tool for coaches of younger players. It provides a visual representation, which would be easier for the players to understand. This understanding could lead to a greater willingness to practice in order to improve their weaknesses and continue to build up their strengths.

Analytics methods used in this research: shooting percentage, scoring efficiency, heat maps
Heuristic Algorithms

Heuristic Algorithms in Sports Analytics
It’s a great day in sports analytics! Today, we will take a dive into the concept of heuristic algorithm and how it has impacted the world of sports.

The major benefit of using heuristic algorithms in sports analytics is that it enables coaches, teams, analysts and other field persons to make quick decisions that can improve team’s performance.

Heuristic algorithm techniques often employs the rule of thumb technique in making sporting decisions. A simple rule of thumb used is recognition heuristic. It states that where there are two available options and one is recognized and the other is not, the option that is recognized is to be chosen. Rules of thumb allow athletes, referees, coaches and fans make effective decisions under internal and external constraints.

During sporting events, there are important decisions that will influence team performance and ultimately, the outcome of games. Heuristic algorithms help coaches decide whether to sell a player or keep them on the team, or whether to acquire a new player. They can also help fans predict the outcome of a competition.

The application of heuristics makes use of a collection of strategies, known as minds ‘adaptive toolbox’ from which the best strategies can be determined.

Heuristic algorithm can determine player actions in basketball games. It can help players make decisions by analyzing whether they should have passed the ball or taken the shot, which player they should have passed to, etc.

The heuristic method commonly used in this type of decision-making is called sensorimotor behavior (take-the-first and take-the-best)

Take-the-first is simply to choose the first alternative that comes to mind. Research has shown this to be the best action in most of the cases.

Heuristic models enable coaches to accurately judge the performance of individual players during the course of a game.

Heuristic algorithm and strategies help sports officials, such as referees make split-second judgment on rule violations, and assists judges in rating athletes’ performances.

The knowledge of heuristics can also help sports managers and fans make near-accurate analysis of which players on a team will perform better in a championship competition.

For instance, the analysis of the hot hand phenomenon (a notion that a person who makes the most successful outcomes at random event has a greater probability of success in future attempts) has helped fans and analysts compare the performance of a player with the player’s average performance, rather than with other players’ performances, or worse, depending on chance.
These decision-making processes help managers choose if a player should be sold as well as which players to acquire.

In conclusion, the application of heuristics algorithm method in sports analytics has changed how athletes, coaches, referees, analysts, and fans perceive field activities, especially in basketball games. This method is regarded as one of the best approaches used to make quick sporting decisions on the field.
Hidden Markov Model

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states. The hidden Markov model can be represented as the simplest dynamic Bayesian network. HMM is closely related to earlier work on the optimal nonlinear filtering problem.


Counterpoints: Advanced Defensive Metrics for NBA Basketball

This is a review of the NBA research conducted by Alexander Franks, Andrew Miller, Luke Bornn, and Kirk Goldsberry, applying a hidden Markov model.

Basketball is a game based on both offensive and defensive skill. However, to date most basketball statistics deal with the offensive side of the game. This limits the ability to evaluate teams and players to strictly one dimension. In an attempt to fill this void, five new defensive statistics are created: volume score, disruption score, defensive shot charts, shots against, and counterpoints.

The first step is creating a model to estimate defensive matchups at every moment in a basketball game. This process was completed for all basketball games played during the NBA 2013-2014 season. The results provide information regarding who is defending who at every point of a possession by estimating an average defender position as a function of offender, ball, and hoop locations. A hidden Markov model is then employed to determine the progression of defensive matchup over the course of a possession.

From this model, the volume score and disruption score statistics are determined. The volume score is the measurement of total attempts a defender faces in a game, which is computed with a multinomial logistic regression. Disruption score measures the ability a defender has in reducing the effectiveness of the opponent he is defending, calculated using a logistic regression to predict shots made and shots missed. From there a defensive shot chart is created to visualize the volume and disruption scores.

The issue with these statistics is that they are static, while the defenders are not. The defense is a continually flowing process with defenders changing which member of the opponent's team they are guarding. To deal with the dynamic nature of the game, counterpoints are assigned. Counterpoints are a weighted average of points scored per 100 possessions against a defender. Three methods are created for counterpoints, or points against. The original matchup method assigns counterpoints to the defender who was guarding the shooter at the beginning of the possession. The pre-shot matchup method assigns counterpoints to the defender who was guarding the shooter when the shot was taken. The fractional method assigns counterpoints proportionally with each defender receiving points based on what fraction of the possession they guarded the shooter.

All counterpoints look at the number of shot attempts made and the average number of points scored against a defender. These measures provide coaches and analysts with a way to rank players based on
their defensive skills. This would be useful when looking at making trades and looking for players to fulfill specific defensive needs.

Taken together, these defensive statistics provide methods to quantify the defensive value of individual players. High ranked defensive players can be analyzed to determine what makes them effective. This information can then be used to create other strong defensive players as well to increase the offensive’s ability to counteract those skills.

Combining both offensive and defensive statistics provides a well-rounded evaluation of each player and their contribution to the team and the game.

Analytics methods used in this research: Volume Score, Disruption Score, Defensive Shot Charts, Shots Against, Counterpoints, Hidden Markov Model, Multinomial Logistic Regression, Logistic Regression
Image Transfer Model

Image transfer models identify factors that influence the creation of an event’s image by using theoretical perspectives from the celebrity endorsement literature to suggest that an event’s image associations are transferred to the sponsoring brand through event sponsorship activities. Addresses moderating variables impacting the strength of the meaning transfer and attitude towards the brand.


Fans' Responses to the National Basketball Association's (NBA) Pilot Jersey Sponsorship Program: An Experimental Approach

This is a review of the NBA research conducted by Dae Hee Kwak and Sean Pradhan, applying an image transfer model.

Advertising or placing a company's logo on team jerseys is a common practice in many European leagues; however, North American leagues are more hesitant about the practice. Advertising on sports' jerseys is win/win situation - the company gains brand recognition and the team earns additional revenue. The NBA approved a three-year pilot program starting in the 2017-2018 season allowing NBA teams to sell jersey sponsorships. The program is designed so that the team who sells the sponsorship for their jerseys retains 50% of the revenue while the other 50% will be split equally among the all 50 teams in the league. This will help smaller market teams gain revenue, as they will likely have a harder time selling sponsorships.

The question is how fans will respond to the jersey sponsorships. This study is designed to determine how fans would respond to jersey sponsors; taking into account various market, manufacture, team, and individual factors.

An image transfer model is created to test this idea. The model examines whether a potential sponsor would want to invest into positive images and loyalty associated with a team by pairing their logo with the jersey. It is expected that positive image transfer will occur due to repeated exposure to the fans.

Six possible positive influences on the positive image transfer are performance, team identification, market size, sponsor brand prominence, manufacturer logo, and fan ethnicity. One influence is the team's performance. It seems logical that teams who are performing well will help boost a positive image transfer. The closer a fan identifies with a team the more likely they will feel favorably towards the sponsor. Larger markets increase the exposure of the brand and therefore may increase image transfer. It is possible that brands that are more prominent will gain greater image transfer as might including a manufacturer's logo. The question also arises as to how fans of different ethnicities will respond to the practice.

The study used a 2 (manufacturer logo present or absent) x 2 (sponsor prominence high or low) between-subjects design. Participants who identified themselves as NBA fans were recruited through Amazon Mechanical Turk to participate in the study. The participants were asked which team was their favorite and were then shown their team's jerseys in a variety of images with combinations of high
sponsor prominence (Fortune 500 companies) and non-Fortune 500 companies, with manufacturing logo and without.

Gender was used a covariate in a series of analyses of covariance (ANCOVAs) measuring the impact of the participant's ethnicity, play-off status and market size of the identified team, and level of participant's team identification. Fans were categorized as either avid or casual.

International participants scored higher rates on brand attitude, awareness and credibility as did avid fans. Higher ratings also occurred when subjects were exposed to sponsors that are more prominent and when a manufacturer's logo was displayed. These same conditions also indicated a higher likelihood of purchasing the sponsoring brand products as well as the jersey.

Various leagues can use this information to determine whether they want to allow their teams to sell jersey sponsorships. Is this something that the North American audience would embrace?

Analytics methods used in this research: Image Transfer Model, 2 x 2 Between-Subjects, ANCOVA
Inertial Sensor Training Load Metrics

Technological developments have led to the production of inexpensive, non-invasive, miniature magneto-inertial sensors, ideal for obtaining sport performance measures during training or competition.


Volume and Intensity are Important Training Related Factors in Injury Incidence in American Football Athletes

This is a review of the research conducted by Patrick Ward, Michael Tankovich, J. Sam Ramsden, Barry Drust, and Luke Bornn.

Participating in sports invariably leads to injuries, especially at the professional level. A 10-year investigation of pre-season training camps in the NFL indicated that for every 1000 athletes 12.7 were injured during training sessions and 64.7 were injured during a game. Some injuries were due to contact with other players, however, many were non-contact injuries such as muscle strain. It is possible that this is a consequence of high training loads.

A study was conducted throughout 24 weeks of training during the pre-season, regular season, and playoff period for one NFL team. This time period included 76 training sessions. Players wore integrated micro technology sensors which consisted of a GPS unit to monitor running activities and inertial sensor units to monitor non-running activities including changes of direction, shuffling, and cutting. A physical therapist recorded any injury data including type and cause of each injury. All of this information was combined and evaluated.

This study focused on the relationship between training load and non-contact injuries. A non-contact soft tissue injury was defined as any injury that did not occur due to contact with another player and which resulted in the player having to sit out at least one subsequent training session or game. Statistical analysis was completed using logistic regression in order to try and understand the relationship between training load, position, and non-contact soft tissue injury.

During the 76 training sessions there were 28 non-contact soft tissue injuries resulting in time loss. Analyzing the data showed that both total playing load and a very high playing load were found to substantially increase the risk of injury on any given training day. A high IMA (inertial movement analysis) per minute had the strongest relationship with non-contact soft tissue injury. IMA includes all non-running activities that involve changes in direction.

Coaches need to be aware of the relationships between training loads and injuries in order to minimize the probability of their players experiencing injuries that cause a loss of playing time. Training days that consist of high volume and intensity are related to an increased risk of injury. Training days that consist of large amounts of low intensity training are related with a decreased risk of injury. These findings
indicate that there is a relationship between volume and intensity. Coaches and trainers need to keep this mind as they plan out the training regime for the time.

Analytics methods used in this research: inertial sensor training load metrics, logistic regression model
Injury Risk Mitigation System

An Enhanced Metric of Injury Risk Utilizing Artificial Intelligence

This is a review of the Injury Risk Mitigation System research conducted by Calham Dower, Abdul Rafehi, Jason Weber, and Razali Mohamad.

Player injuries are a key concern for all teams across all sports. Injuries can cost teams thousands of dollars every day and negatively affect the teams' competitiveness. Over the years several programs have been developed in an attempt to estimate injury risk, but to date none have been able to present precise and accurate data. The Injury Risk Mitigation System (IRMS) was designed to solve this problem. The IMRS is able to take advantage of the fact that teams are collecting more data than ever before. It helps analyze this information in order to produce estimations of an individual's injury risk on any given day by looking for patterns in the historical data.

The human body is an extremely complex system and therefore it requires a complex program that is able to analyze data for patterns leading up to any variety of injuries. An Artificial Neural Network is capable of modelling very complex systems and determining how different variables relate to each other and whether those variables are independent or dependent upon each other. Testing the program using data from two current Australian Rules Football teams from 2012 to 2017, illustrated that teams can now more accurately forecast the risk of injury for the upcoming two-week period than any other approaches currently being used.

The more data is entered into the Injury Risk Mitigation System, the more accurate its predictions. Using only one team's data means it could take years for the system to make accurate predictions. IMRS is able to minimize this issue by using data from different teams. It is able to do so in such a way that the data remains anonymous as to where it came from. This allows the system to begin making accurate predictions much sooner and help teams minimize injury risk among their players. As training patterns change from year to year, the model must be updated on a regular basis. As long as a team consistently uploads new data IMRS will be able to provide accurate predictions for the next two-week period.

IMRS does not predict injuries, but instead predicts the risk of injuries. Predicting the risk of an injury, rather than injuries themselves, is very valuable in formulating training plans for players, as these plans can be adjusted in order to minimize the chance of those injuries actually occurring. One caution to remember is that IMRS's ability to predict injury risk is more accurate during the season than in the preseason.

In summary, a player's risk of injury ebbs and flows over the season. Teams can recognize the periods when a player has an increased risk of injury. This will aid the team in managing their players in order to maximize their performance and minimize their risk of injury. This model is most useful towards the end of a season, helping teams balance performance versus risk of injury and allowing them to maintain healthy athletes as they approach the end of the season and championships.

Analytics methods used in this research: Injury Risk Mitigation System, time series and static data, data-sampling, normalization, optimization artificial neural network
Interviews and Surveys

Interviews and surveys are widely used to help us understand the thoughts and intentions of players and coaches, making them one of the most valuable sources of data for the study of athletic and sports management behavior. Interviews and surveys may be personally administered as in interviews or self-administered as in questionnaires. Despite their importance, however, the current practice of conducting interviews typically lacks quality control whereas surveys generate low response rates and tend to suffer from selection bias.

Decision-Making in the National Basketball Association

This is a review of the NBA decision making research conducted by Jonathan Mills applying interviews and surveys.

For years, coaches in the NBA have used basic statistics and experience to evaluate teams and players. More advanced analytics continue to become more readily available. The trick is deciphering what these statistics mean and how they should affect your decision-making. This research was conducted through an interview and survey process. One representative from half of the teams in the National Basketball Association participated in the process.

Results of the interviews indicated that all participants felt that analytics are an important tool for making qualified decisions. They also believed that analytics and traditional evaluation methods such as scouting players, watching film footage, or conducting workouts worked best when used in conjunction with each other. They both have their own strengths and weaknesses. Traditional evaluation tools are subjective and, as a result, are often biased. They are also limited to a small sample size. However, they provide first-hand information and context regarding the player and his background. Analytic tools are objective and, therefore, have no bias. They are also based on a large sample size. However, they are lacking context, those intangibles that make individual players stand out from the others, intangibles that can lead to a player becoming a star in the league. Therefore, it is optimal for coaches and other decision makers to develop a plan of how to include both types of information in a manner that provides optimal data to help them make informed decisions.

Historically, traditional evaluators and analysts have often not seen eye to eye and, at times, seem to work at cross-purposes. In order to maximize the potential of the information received from both sources this lack of agreement needs to be resolved. Both groups need to have an understanding of what the other is doing and why they are doing it. Outcomes will become much clearer when the two groups understand each other and are working towards a common purpose. This will give their teams the best data possible to make both major and minor decisions.

The majority of teams do not have an organizational system in place to help them balance all of the information they receive. One possible method would be to assign values to the different pieces of information. Another method would be to pull out key points from each type of information and put them together to create a more rounded picture of the player or team they are evaluating. There is not just one method for balancing the information and each decision maker will need to decide for themselves the best way to organization all of the information received, both from traditional methods
and analytics, in order to maximize their decision making process. After all, the goal is to win games and ultimately a championship.
Kelly Criteria

In probability theory and intertemporal portfolio choice, the Kelly criterion, Kelly strategy, Kelly formula, or Kelly bet is a formula for bet sizing that leads almost surely to higher wealth compared to any other strategy in the long run (i.e. the limit as the number of bets goes to infinity). The Kelly bet size is found by maximizing the expected logarithm of wealth which is equivalent to maximizing the expected geometric growth rate.

https://en.wikipedia.org/wiki/Kelly_criterion

Modified Kelly Criteria

This is a review of the research conducted by Dani Chu, Yifan Wu, and Tim B. Swartz.

Sports and gambling seem to go hand in hand; where you find one, you inevitably find the other. The Kelly criterion is a strategy used in several forms of gambling, including sports betting. It is designed to create the right balance between risk and reward while reducing volatility. To determine the Kelly criterion the gambler needs to determine the probability of placing a correct wager using a specific gambling system.

The Kelly criterion has received a great deal of attention, including some negative feedback. Some gamblers say that the Kelly fraction is too high, leading to financial losses. Consequently, suggestions have made regarding the criterion. These complaints are counter intuitive as the Kelly criterion is based on mathematical proof. It maximizes the exponential rate of growth while minimizing expected time needed to reach the desired profit. The problem arises with the probability value used to determine the Kelly fraction. Gamblers are often overly optimistic regarding their chances and use a higher probability that is accurate for the situation.

To determine the optimal wagering fraction a statistical model based on a proposed wagering system is developed. In this model, a loss function is required in order to determine the quality of the estimation. Gamblers often use historical data to determine probabilities, ignoring the fact that the true probability is unknown.

Developing a modified Kelly criterion is done by gradually increasing the assumptions, done with a decision theoretic framework with a loss function. The first approach, using the fewest number of assumptions, is based on minimax estimation. The results, however, proved that the approach is too conservative and does not provide useful output for the Kelly criterion. The second approach is based on Bayes estimation. This provides more flexibility for the gamblers as it allows for accommodation of different prior beliefs. The optimal betting fraction is obtained using a Beta prior distribution of the gambler’s probability of selecting winning wagers. When the priors in the distribution are conservative, the modified Kelly criterion tends to smaller than the original Kelly criterion.
Incorporating a loss function also is beneficial in determining a modified Kelly criterion. A variety of loss functions are possible, leaving gamblers with options as to which they prefer to use within their gambling parameters.

Analysts and gamblers can use the modified Kelly criterion to determine optimal bets for various sporting events. Analysts, in determining the likelihood of outcomes of various games, can use this information. Comparisons could also be made across various sporting events to analyze how the Kelly criterion varies by sport, by league, or even by teams.

Analytics methods used in this research: Kelly Criterion, Decision Theoretic Framework, Minimax Estimation, Bayes Estimation, Beta Prior Distribution
Kernel Density Estimation

In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. In some fields such as signal processing and econometrics it is also termed the Parzen–Rosenblatt window method.


New Metrics for Evaluating Home Plate Umpire Consistency and Accuracy

This is a review of the major league umpire research conducted by David J. Hunter, applying kernel density estimation.

Boos have been ringing out across baseball diamonds, criticizing umpire's call judgements since the beginning of the game. Pitch-tracking data provides a method to evaluate and train Major League Baseball umpires and evidence suggests that since the installation of the equipment umpire accuracy has improved. Today, players and managers accept variations in calls across umpires as long as each umpire is consistent in their strike zone.

The rulebook cites a rectangular strike zone is to be used by umpires. However, in practice, the strike zone has more rounded corners with pitches on the corners of the rectangle being called balls. In addition, pitches off the plate opposite from the batter are more likely to be strikes than pitches off the inside of the plate. This suggests that strike zones differ between left and right-handed batters.

New metrics are being created to evaluate the consistency and accuracy of an umpire's calls over the course of a game. First, the requirement for a rectangular strike zone is dismissed, and variations based on the handedness of the batter are permitted. As factors such as the starting pitcher can influence an umpire's strike zone, consistency is measured within a game and averaged over all games in a season.

Ideally, each umpire should establish his strike zone and follow it consistently throughout the game, no matter who is up to bat. Four different metrics are proposed for evaluating the consistency of calls relative to the established strike zone of the umpire.

The first two metrics are rectangular metrics. The first one looks at the smallest rectangular region that contains all of the called strikes. Any pitches called balls within this zone are said to be inconsistent. The one-rectangle inconsistency index is defined as the number of inconsistent balls divided by the total number of called balls. This index is easy to determine but is very sensitive to a single outlier. It also does not account for multiple bad calls in the same location. These issues are addressed by using more rectangles. In this case, inconsistent balls are weighted according to how many rectangles they are contained within. The more rectangles they are contained within, the greater their weighting.

The last two metrics are convex hull metrics. As strike zones established by umpires are not rectangular in nature, these metrics relax that assumption. Instead, a consistent zone is assumed to be based on the idea that any pitch landing between two called strikes will also be called a strike and therefore the
established strike zone is convex. Similar to the one-rectangle index, the convex hull index can fail to account for multiple bad calls in the same location. To combat this, the location of called balls is used to define a called ball region. Rather than counting called balls within the established strike zone, the area of overlap between the called-ball region and the convex hull of strikes is measured.

All four metrics are sensitive to a single outlying called strike and to a bad call of a ball in the middle of the strike zone. This is done so that umpires who make slightly inconsistent calls are not penalized to the same degree as those who make clearly bad calls. However, as a result these metrics are also sensitive to the number of pitches called. As the number of called pitches increases, the chances of making a poor call also increases.

A kernel density estimation is a more accurate method for determining the borders of a consensus strike zone. It can also be used to assess conformity and zone size of individual umpires. Kernel density also works well with small sample sizes.

Using one rating systems allows umpires to be ranked. Such a ranking could provide a basis for salaries, rehiring, and termination. While the evidence suggests that MLB umpires are typically quite accurate and consistent, mistakes can happen, as umpires, after all, are only human. Strike zone inconsistency could be due to game circumstances such as the strike count. The age and experience of the pitcher may also have an impact. These possible factors could be analyzed to see if any are true for an umpire, giving the umpire an opportunity to improve his call making, and avoiding changing his strike zone in any of these circumstances.

Umpires will never be 100% accurate and that is just part of baseball. However, ensuring as much accuracy as possible results in a game being won or lost based on the player performance and not the umpire’s calls.

Analytics methods used in this research: Rectangular Metrics, Convex Hull Metrics, Kernel Density Estimation
Knapsack Problem

Optimization includes finding "best available" values of some objective function. The knapsack problem involves combinatorial optimization, where given a set of items, each with a weight and a value, determine the number of each item to include in a collection so that the total weight is less than or equal to a given limit and the total value is as large as possible. The problem often arises in resource allocation where there are financial constraints and is studied in fields such as combinatorics, computer science, complexity theory, cryptography, applied mathematics, and daily fantasy sports. (Wikipedia)

Optimal Decision-Making in the NBA Free Agency Market

This is a review of the knapsack formulation research conducted by Pravek Karwe, Rishi Khera, Narain Krishnamurthy, and Sarah Tracy.

At the end of every season NBA teams need to evaluate those players who are now unrestricted free agents to determine which players would have the greatest impact on their team in future years. Simply looking at how many points a player scores will not give the team the entire picture. There are many intangible factors that need to be considered - players can positively affect game outcomes without actually scoring any points.

A starting point for evaluating a player's value is the Player Efficiency Rating (PER). John Hollinger developed the PER, an all-in-one basketball rating, which attempts to boil down all of a player's contributions into one number. The PER is standardized which makes it easier to compare players across the NBA. PER has some limitations in that it focuses mainly on offensive statistics, including only two defensive stats. It also does not include any intangible statistics.

In order to update the PER to include defensive and hustle skills, extra value will be added to each player by adding in five additional statistics: screen assists, ball deflections, loose balls recovered, ability to cause opponents to make fouls, and contested shots. This model will more accurately reflect the expected value of any player. Positional needs of the team and salary constraints are also incorporated within the model.

It must be kept in mind that while this model will determine the top choices for a team out of the group of free agents, other teams will also be trying to sign these players. As free agents are signed by other teams the model would have to be run again incorporating the remaining free agents to determine which of them are best suited to the team’s needs.

To further improve the model, the thought processes of the team should be taken into account. Depending on their previous season, teams can expect to make it to the playoffs the next year or maybe plan to rebuild instead. Teams working towards making it to the playoffs will tend to want older, more experienced players who have already developed the skills needed to help lead the team. A team planning to rebuild will be more interested in signing young talent that they can help mold and grow into future leaders. This idea of the team's thought processes was added into the model using the
difference between the average age of players in the league and the age of the individual player. Younger players would have a positive difference while older players would have a negative difference.

Possible future enhancements to the model could include analyzing player styles and looking at salary cap exceptions.

This improved PER model assists analysts and teams in determining which free agent players will be able to assist the team in achieving their future goals.

Analytics methods used in this research: optimization, expected value, weighting, player efficiency rating, standardization, knapsack formulation, linear program maximization
Latent Trajectory Model

Traditionally called latent growth curve models, latent trajectory models (LTM) are a relatively new technique to model changes of a certain phenomenon over time. The term latent is probably from the structural equation modeling (SEM) literature, where a latent variable is an abstract construct (or concept) that cannot be directly measured, but can be approximated by some measurable variables. The term trajectory is more accurate, as LTM can be used to model not only growth, but also decline and other more complex non-monotonous change patterns (such as cosine function)—so do not think LTM is only for phenomena that increase (or decrease) only over time.


A Finite Mixture Latent Trajectory Model for Modeling Ultrarunners' Behavior in a 24-Hour Race

This is a review of the ultrarunner’s research conducted by Francesco Bartolucci and Thomas Brendan Murphy, applying latent trajectory modeling.

Two hundred ninety-nine competitors ran in the 2013 International Association of Ultrarunners 24-Hour World Championships. The course consisted of a 2.314 km lap with each runner carrying a chip on their shoe in order to automatically record their lap count and the time they finished each lap. At the end of 24 hours, a bell sounded and all runners stop running immediately. Two hundred nineteen runners lasted until the sound of the bell. The fraction completed of the final lap is added to their total. Runners have the option of continually running, resting, or leaving the competition during any lap in the race.

Data from this race is studied to look at the strategies used by the runners and whether age or gender affects those strategies.

A latent trajectory model is created using a linear regression component for the speed observed for every lap completed by the runner as the response variable. A multinomial logit regression component is used for the categorical response variable that indicates if the individual is continually running, resting, or leaves the race during each lap. The model takes into account times when the runner rests or stops which can have a major influence on strategy. As a result, runner strategy is determined by speed and tendency to stop rather than just speed alone.

The model accounts for different possible running strategies and groups runners according to strategy. The number of different groups is chosen with a Normalized Entropy Criterion, which indicates that three groups is the appropriate number of groups to group runners into distinct strategies.

Results show that all three groups decreased speed over the course of the race with an increase in speed towards the end. Most runners maintain an even pacing in the early laps with higher performing runners maintaining their even pacing for a longer period of time. The pattern of speed is similar across all three groups, only the average speed differs. Gender and age do not significantly affect strategy. Finally, all three groups contain runners who drop out of the race, with the greatest number dropping
out in the middle of the race. Dropping out of the event becomes less prevalent later in the race. The largest group is made of the middle speed runners. The fastest runners form the next largest group and the slowest runners make up the smallest group.

Coaches of ultrarunners could use this information when determining strategy for a race. Runners can be trained to maintain an even pace for as long as possible during the beginning of a race. Psychological effects related to dropping out during the middle of the race can also be addressed. Analysts could develop the model further to determine if there are any particular strategies among the groups that could improve the performance of other runners.

Analytics methods used in this research: Latent Trajectory Model, Normalized Entropy Criterion, Multinomial Logit Regression, Linear Regression
Using Multi-Class Classification Methods to Predict Baseball Pitch Types

This is a review of the MLB research conducted by Glenn Sidle and Hien Tran, applying linear discriminant analysis.

Examining statistics across the years in the MLB shows us that pitchers have been getting better over time at preventing hits, which results in a lower average ERA and batting average. Batters have a difficult job when it comes to hitting pitches but is it possible that their job could be made easier by giving them a technique to determine what type of pitch will be thrown next. In an effort to determine a technique that would work in live game situations the predictive abilities of three different machine learning techniques are considered.

Data regarding pitches is much easier to collect since cameras were installed in all of the MLB stadiums to record the speed and location of each pitch. From the database, information regarding every pitch from the 2013, 2014, and 2015 seasons was obtained. Data sets were created using 22 features from each pitch, along with up to 81 additional features depending on how many types of pitches were thrown by the pitcher. Two hundred eighty-seven data sets were created for individual pitchers, 150 starters, and 137 relievers, with each data set containing an average of 4,682 pitches. Seven categories of pitches are considered: fastball, cutter, sinker, slider, curveball, changeup, and knuckleball. Due to the size of each data set, similar features are grouped together making it easier to complete comparisons and contrasts.

In order to compare and contrast the results from the 287 pitchers, predictions using three different learning techniques were produced. The first technique explored was a linear discriminant analysis because of its speed and efficiency. Multi-class support vector machines were the second technique and the third was classification trees. A committee method, using ten of each model type and taking the majority vote as output was used to decrease variance between the models and increase accuracy of the predictions. Comparing the three techniques shows that classification trees provide the greatest prediction accuracy.

In order to test the validity of the method in a live game situation models were created for each pitcher who played during the months of September and October 2016. The models were designed to predict the type of pitch thrown, the speed of the pitch, and the location of the pitch. The average accuracy across each pitcher in each game was 60%.
Teams can use this information to provide batters with an additional tool in the battle to beat the pitcher and hit the ball. The model is able to incorporate data known only moments before the next pitch is thrown, providing teams with the most-up-date prediction possible.

Analytics methods used in this research: Linear Discriminant Analysis, Multi-Class Support Vector Machines, Classification Trees, Committee Method

Using PITCHf/x to Model the Dependence of Strikeout Rate on the Predictability of Pitch Sequences

This is a review of the research conducted by Glenn Healey and Shiyuan Zhao.

A fastball typically covers the distance between pitcher and batter in 400 milliseconds. A batter's ability to synchronize visual information with physical movement must be extraordinary as decisions on whether or not to swing and at what angle must be done within 200 milliseconds, which half the length of a blink of the eye. Batting becomes more difficult as pitchers throw a wide variety of pitches with different speeds, height, location over the plate, and spins. Being able to predict the next pitch the pitcher will throw is immensely helpful for the batter.

Preventing batters from making accurate predictions is therefore an aim of all pitchers. Teams need to decide what distribution of pitches is optimal, based on the pitcher's array of pitches, strengths, and weaknesses of the batter, the score, the inning, number of outs and base runners, and the identity of the upcoming batters. Pitchers can also gain an advantage by varying the sequence of pitches. Batters' errors when swinging the bat are larger when there is a significant difference in the speed of consecutive pitches.

More than 50% of pitches thrown are a variation of the fastball. There is a positive correlation between the speed of a pitcher's fastball and his strikeout rate. A pitcher's success is also dependent on the distribution and sequencing of pitches.

Images from PITCHf/x are used to estimate the path of a pitch, its speed, and movement. These measurements are used to develop a list of descriptors for a pitcher that provide correlations between consecutive pitches in regards to location, movement, and velocity. The correlations are computed separately for each of the four configurations of pitcher and batter handedness, i.e. right-handed pitcher and right-handed batter or left-handed pitcher and right-handed batter.

A pitcher's strikeout rate is again computed separately for each configuration. The strikeout rate is the ratio of strikeouts to adjusted plate appearances, which excludes appearances that resulted in a bunt or intentional walk or when the batter is a pitcher.

A year-to-year analysis shows that velocity and movement measures are more consistent than location.

A linear regression model for each configuration is used to estimate the pitcher strikeout rate. Analyzing the results tells us that off-speed pitches have a greater impact when pitcher and batter are either both right handed or both left handed. Velocity of the ball is significant across all configurations. However,
the larger the proportion of fastballs, the lower the strikeout rate. Increased vertical ball movement and lower pitch-to-pitch correlation both result in higher strikeout rates.

This information provides pitching coaches information to help formulate pitching distributions. Keeping all this in mind while on the field will aide catchers and pitchers in making efficient decisions regarding types of pitches versus the different batters. Analysts can use this information to compare pitchers and batters, determining if the pitcher or the batter has an advantage in a particular matchup.

Other aspects such as the pitcher's ability to hide the ball are not included in this study, but could be incorporated into future studies.

Analytics methods used in this research: Correlation, Distribution, Linear Regression Model
Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

Source - https://en.wikipedia.org/wiki/Linear_regression

Drafting Errors and Decision Making Bias in the NBA Draft

This is a review of the NBA draft research conducted by Daniel Sailofsky, applying linear regression modeling.

Every year NBA teams struggle to determine who their coveted draft choices should be. This study focuses on the choices that teams have made in order to analyze decision making biases and errors. The data for this research came from the NCAA players who were drafted by NBA teams between the years 2006 and 2013. Information regarding the draftee’s court statistics, conference they played in, the year they were eligible to be drafted and the year they were rookies. Data was collected regarding their performance after they joined the NBA including where they were picked in the draft, their pre-draft physical measurements as well as statistics regarding their performance and length of their career. This information was analyzed to determine how well NBA teams were choosing their draft picks based on their performance after they joined the NBA. The conclusion was that good decisions were not being made when selecting draft choices. The research indicated that most of the factors used to determine who to draft did not relate to the player's future performance.

Teams typically want to draft players from well-known colleges that belong to conferences that are traditionally considered strong. It is human nature to avoid change and those making the draft decisions are no different. They are more familiar with the Big Conferences so they feel safer drafting those players. They fail to take into account the strengths of players from other conferences and, as a result, are often not choosing those players who have the best chance of making a difference in the NBA.

Research has shown that players who have strong skills regarding ball control and preventing turnovers do well in the NBA. However, these skills have no correlation to the players who are actually drafted. Players with a high rebound percentage and free throw rate in college do better in the NBA than others as well. However, again, these statistics do not relate to those who are actually drafted. The research seems to indicate that those players who make memorable plays in their college days are the ones who are drafted. NBA teams also tend to look at scoring and blocked shots statistics when choosing draftees, but these skills do not actually predict future NBA performance.

NBA teams tend to be more interested in the younger, less experienced players. Statistics show that teams are more willing to take on players who have the physique of a typical NBA basketball player even if they lack high-level basketball skills. NBA teams will overlook these players' weaker skills, believing
that with proper coaching these players will become better. Unfortunately, size does not guarantee skill levels and no amount of coaching will be able to make these players effective NBA players. Height is positively correlated to draft position but height is not correlated to future performance.

Research shows that scoring is positively related to a players’ draft position but is actually negatively correlated to future performance. Of course, scoring is a skill necessary to be effective at the NBA level, however, there are other factors that are related to future effectiveness. Star players in the NBA were also strong shooters in their college days, however, they would have strong skills in other areas as well.

Teams need to look carefully at how they choose who to pick in the draft. They need to determine what characteristics are related to a college player becoming a strong player in the NBA, and leave their biases behind.

Analytics methods used in this research: linear regression model, probability, percentage-based statistics, variables
Log Difference

A log difference is the difference between two logged variables. Log differences equals the logged ratio between those two variables calculated on their original metric.

Gender Differences in Marathon Pacing and Performance Prediction

This is a review of the research conducted by Calvin Hubble and Jinger Zhao. Marathon runners are a unique breed, male and female, elite and recreational. Men have the advantage when it comes to speed. The question is whether that makes them better marathoners than women. Studies suggest that men tend to be more overconfident in their athletic abilities than women, which may cause them to choose a sub-optimal pacing strategy during a race. If this is true, what effect does it have on their performance?

Data was collected from the 2013 Houston Marathon including gender, age, net finish time, splits at 5K intervals, predicted finish time, and grouping. The marathon splits runners into 4 groups; one is for elite runners, one for runners who have completed a marathon in less than 4 hours in the past, one for those whose predicted finish time is faster than 4.5 hours and the last for those whose predicted finish time is longer than 4.5 hours. From this information, a shortfall for each runner was determined by measuring the log difference between the runner's expected and actual time. Also determined was the second half slowdown by measuring the log difference between the time taken to run the first half of the marathon and the time taken for the second half.

The shortfall statistic is subjective as the runner determines the predicted time. It seems reasonable to assume that this, then, is a mark of the level of confidence of the runner, with runners who underestimate their time being overconfident in their abilities. The data demonstrates a significant difference in the shortfall values between men and women across all ages and groups. Women consistently underestimate their finish times by a lower percentage than men do.

To analyze the slowdown effect the second half slowdown measures are plotted on a histogram. These histograms show that men slow down by a greater percentage in the second half of the race than do women. Next, the split data at every 5K interval are used to create a race profile for each runner and T statistics are computed of the difference between the average male and female race profiles. Men's pace before the 25K mark is significantly faster than the women's pace and significantly slower in the second half of the race.

Finish times against second half slowdown are graphed on a scatterplot demonstrating a 32% correlation between the two values. This indicates that pacing and shortfall are related, with a slower pace in the second half of the race typically leading to a slower finish time.

Using this information male marathoners and their coaches can look at ways to alter the runner's reaction to stress and how to deal with the overconfidence factor. Pacing strategy needs to be worked on even more consistently in order to ingrain it even more deeply in the runner's psyche. Analysts can
use this information to analyze just how quickly men would be able to finish their races if they overcome this problem. Might it lead to even more broken records?

Analytics methods used in this research: Log Difference, Graph, Histogram, Scatterplot, T-Statistic
Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers.


Machine Learning as a Method of Sports Analytics

This chapter examines how machine learning analysis can be applied in sports analytics to improve team sporting performance and winning opportunities. Machine-learning analysis for sports analytics involves the use of a data-programmed machine to make analysis. The machines used in this analysis are programmed to detect similar playing patterns in individual players in the game of sports.

Machine learning can analyze complex information such as the speed of team players on the field, distance covered by players, and their level of fatigue both on the field and after a play.

This is in order to help team coaches determine the best course of action for their teams, such as which player to replace during a match and what to expect from competitors during the match.

The machine learning analysis is important for the following areas:

1. A major advantage of using machine learning is its flexibility in managing large volumes of data as well as multiple data sources. The use of a machine learning made data collection easier and more effective.
2. Machine learning is a quicker method of analytics. Using this method of analytics, coaches and sports analysts may not need to spend many hours on watching recorded game films, because meaningful insights can be derived by simply viewing the machine learning analysis. This also gives coaches the benefit of spending more time solving team problems as many solutions are already detected by the machine.
3. Machine analysis is a critical and detailed method for analyzing data of continuous sport activity, such as soccer where the data is either 0 (nil) or 1 (goal).

Machine learning analytics is fast becoming popular in the world of sports. Several sports have recorded huge success through the use of machine learning.
For example, the use of machine learning analysis has been employed by the famous Manchester City Football Club. According to research, this has greatly improved their team performance. Manchester City Football Club is also known to employ this technique in recruiting new players as well as in determining what play will suit a field position.

Machine learning analysis is utilized by Chicago-based firm STATS (Sports Team Analysis and Tracking Systems) in assembling information on various player movement.

In previous years, STATS has employed the use of cameras in Europe soccer stadiums and NBA arenas. These cameras, which are integral of SportsVU system, are programmed with the ability to track the movement of the individual players and the ball at 25 frames per second.

STATS cameras have also been programmed to analyze other individual player strengths and weakness, such as running speed, ball possession and distance/endurance.

In other words, the benefits of using machine learning analysis can be greatly improved with the application of artificial intelligence.

An Open-Sourced Optical Tracking and Advanced eSports Analytics Platform for League of Legends

This is a review of the Open-sourced optical tracking research conducted by Philip Z. Maymin, applying machine learning.

League of Nations is a multiplayer online battle arena (MOBA) game. Within a MOBA game, each player controls a single character and fight as a team. The game's publisher only provides very basic statistics such as when a character dies, who was awarded a kill or an assist, etc. This model was developed to provide statistics that are more comprehensive. The model tracks everything that happens throughout the game without affecting the game in any way.

The first statistic provided by this model is in-game win probability, which was determined by using a machine learning model. In-game win probability gives the probability of each team winning the game after each individual action takes place. It is used to measure the player's performance. Within League of Legends win probability is based on the number of kills of other characters, large monsters killed, as well as the number of towers.

The second statistic is full-game win probability. This looks at the individual player's contribution to his teams' win probability. Percentiles are computed for each player, which tells how many players he outperformed in each of the various areas.

Worthless deaths and smart kills is the third statistic looked at. A worthless kill is defined as one that did not increase your team's win probability while a smart kill did increase it. Spending time going after a weak enemy will not increase your team's win probability.

Various other basic and advanced stats can be obtained. Basic stats include total damage dealt per minute, average death time percentage, kills, deaths, and assists among others. Advanced stats include
information such as damage taken percent per death, map coverage, favorable team fights, small kills, and worthless deaths. With some adjusting, any of these can be applied to traditional sports situations as well.

After the game is over each player’s time management can be analyzed. Was time used wisely or wasted on tasks that did not contribute to the team's chances of winning? This, too, could be transferred to traditional sports to evaluate individual player’s time management skills within a game.

When looking at how an individual's performance affects the team's win probability you cannot look at simply one statistic. The statistics must be taken as a whole in order to gain an accurate picture. The important details of each player's performance are put on a fast and frugal tree analysis, which determines which decisions had a positive effect on the team's overall performance.

The insights gained from these eSports analytics can be transferred to traditional sports to answer questions such as how much does an individual contribute to the performance of the team and the probability of them winning. Ideas can be tested using eSports where data can be collected more efficiently with no chance of players being hurt. The lessons learnt can then be transferred and applied to traditional sports.

Analytics methods used in this research: Machine Learning, probability

**High-Resolution Shot Capture Reveals Systematic Biases and an Improved Method for Shooter Evaluation**

This is a review of the NBA research conducted by Rachel Marty, applying machine learning.

Shooting ability is one key to the success of a basketball player. New technology allows us to learn more about what affects a player's shooting percentage such as why shooters miss from certain positions.

High-resolution shot data allows us to analyze where players are on the court when they make their shot, as well as how the ball approaches and interacts with the basket. This data was gained from 22 million shots done in high-resolution by Noahlytics which has a sensor hanging approximately 13 feet above the basket, allowing it to capture all shots taken. The point in time when the ball reaches the rim is analyzed more in depth to determine patterns related to missed shots. The data includes shots from players in the NBA, WNBA, NCAA, high school, etc. which provides a representation of both male and female players as well as players at all levels of development.

How each shot makes contact with the basket is detailed. Does the ball reach the basket directly in front of the player or to the right or left (left/right bias)? Does the ball hit the front of the basket or the back (depth bias)? At what angle does the ball enter the basket?

Excellent shooters have very little to no left/right bias; their shots are aimed directly at the basket. However, there are two areas of the court where this does not necessarily hold true. The right side of the court has a left bias while the left side of the court has a right bias. Bank shots follow a path that leads past the basket. Non-banked shots demonstrate some similarities - short shots go deeper into the basket while long shots go shorter into the basket.
In order to gain a true assessment of a players' shooting ability you need a large data set. Players can have a good game, followed by a bad game, etc. so using a limited number of data can create a very unrealistic evaluation of the players' ability. In order to be accurate, sessions of over 1,000 shots need to be used. A model was created to evaluate player's shooting ability using rim patterns (how the ball interacts with the basket) to increase the validity of the shooting percentage. With this model it is possible to accurately determine a player's shooting percentage from small shot sessions.

Analysts and teams use shooting percentages to rank players. This approach produces a more accurate percentage helping rank players more effectively. This is especially important when dealing with trades and the draft. Teams can use this information to help their players become better shooters, giving them specific areas to work on such as a left/right bias. Players will be able to accurately follow their improvement as they compare their shooting score past scores and to other players around the league.

Analytics methods used in this research: bias, machine-learning algorithm, ranking, prediction model
Markov Chain Decision Processes

A Markov decision process (MDP) is a discrete time stochastic control process. It provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. MDPs are useful for studying optimization problems solved via dynamic programming and reinforcement learning.

Source - https://en.wikipedia.org/wiki/Markov_decision_process

Replaying the NBA

This is a review of the shot efficiency research using Markov chains conducted by Nathan Sandholtz and Luke Bornn.

The probability of a player taking a shot is not consistent throughout a basketball game. The probability of the ball handler shooting increases dramatically as the shot clock winds down. Teams would rather take a poor shot than no shot at all. This phenomenon caused by the shot clock makes it more difficult to evaluate shot selection. Simply focusing on the type of shot will not be sufficient, any research must also look at when the shot is being made and who is taking the shot.

Data from the 2015-2016 NBA regular season are combined with play-by-play data. Information including the identity of the ball carrier, their location on the court, and the defensive pressure he is facing is included in the model as well. Defensive pressure is determined by the distance between the defender and the ball handler. Time left on the shot clock is also accounted for. All of this information is inputted into the simulation algorithm.

This Markov chains study focuses on mid-range jump shots taken early in the shot-clock countdown. These shots are typically considered to be less efficient than shots taken from other regions on the court. A simulation is run in which the number of mid-range jump shots taken when there are more than 10 seconds left on the shot clock are reduced by 20%. The simulation tells us that this reduction leads to an increase in three point attempts, as well as an increase in turnovers and shot clock violations. Overall, the expected points per shot increases for almost every team. This is a clear indication that even a minor policy change can have a significant impact.

Coaches can use this simulation to determine how changes in shot policy will affect the expected points for their team in a game. This will help them decide whether such a policy change would benefit the team and should be implemented or harm the team and be scrapped. Coaches could also examine the effect injuries to key players could have on the team. The simulation could be run with substitute players in place of the team’s key players in order to determine the difference in expected value.

While this Markov chains study focused on changing the shot policy, the simulation could be used to determine the effect changes to other policies would have on the team. This would allow coaches to fine tune their policies, making large or small changes that would lead to an increased expected value.

Analytics methods used in this research: Markov decision processes, reward function, shot efficiency
Mathematical Programming

Mathematical Programming in Sports

Mathematical programming in sports refers to the use of mathematical models in making tactical, analytical, performance-based or managerial decisions in training sessions, during the field of play, or in management situations. Programming simply refers to computerizing the planning of activities.

Mathematics is pervasive; it is virtually everywhere – around us, in our day-to-day activities and, of course, in sports. In fact, they become increasingly important in this domain, be it in professional or amateur sport. For example, mathematics programming helps in the improvement of performance, new technologies, technical revolutions, etc. Mathematicians deal with the themes of sport and performance in different ways. Thus, let us discover together and more precisely, how the practice of mathematics has an effect on the sport.

From birth, children discover numbers and geometric shapes, and adults continue to use mathematics in their day to day activities.

Sports and managing high-level athletes is a big business, and thus rely on big business solutions to gain a competitive edge. This is where analytics come in. Mathematical equations are a scientific basis that have become essential in all high-level sports. They lay the foundation of work for athletes who do not stop refining their technique to obtain better results.

Breaking records with mathematics

Mathematics can be applied to sporting activities to break records. For instance, Javier Sotomayor has the record of high jump since 1993 and nobody has managed to beat his 2.45 meters until today.

Do you all remember the 2009 World Swimming Championship in Rome? The records were shattered one after another and journalists could not understand this phenomenon. In total, 43 records were broken and set during this period. You have to think that the performance of swimming is measured by the principle of submerged movement, well known in physics, and therefore involves mathematical calculations. Vertical and horizontal forces are exerted and the Archimedes principle also influences it. How is it possible for athletes to continue breaking records? Does the human body have a defined physical limit for performance?

Mathematics helps us with questions concerning how records can be broken and if there is a physical limit to the human body's performance. A study of September 2012 has shown that the classifications are governed by a mathematical law: the law of power. This law relates two elements: the frequency of an event and its size.

Technology at the service of sport

The advancement of sports equipment is making a big impact on sports science. The combinations of some equipment improve the time spent by swimmers to cross a specific borderline. It's the same thing that has happened in cycling: the technical improvement of bicycles plays a fundamental role in
improving performance. Lighter and more aerodynamic materials help in penetrating the air with minimal resistance.

**Mathematics to improve the technical skills of athletes**

If mathematics can help you manage your money and calculate percentages in your math courses, then they can also help you with sports. For instance, mathematics allows you to determine the trajectory of the ball. Therefore, it should not surprise us that sport rely heavily on algebra, geometry, multiplication tables, arithmetic, whole numbers, proportions, etc. Athletes no longer go to competitions accompanied only by their coach, their physiotherapist or their nutritionist. As times have changed, now there are other types of professionals who accompany the team. For example, the Australian or the New Zealander team in the J.J.O.O. 2016, had real mathematicians on their team. Their function was to gather data and take into consideration the context to develop statistics to optimize athletic technique in all possible parameters. The goal is clear: to achieve perfection in the discipline and optimize training.

**Mathematics in preparing football tournaments**

Mathematical analysis is also taken into account in the selections of football teams, since the selectors use it when they have to make the list for a big sports event. The analysis of team performance and the development of mathematical models allow strategic choices based more on logic than on the "instinct of the coach". It is true that it is something complex, but it can make the difference between a victory and a defeat.
Mean and Median

Measuring Mean and Median in Sports Data

Mean and median are two of the most useful methods used to summarize numerical data.

The mean of a set of data is simply the average, which is found by adding up the total of all the observations and dividing by the number of observations. The mean is used to describe the average number of goals a team scores in a game or the average number of saves a goaltender has per game. Traditional sports statistics rely heavily on the mean.

The median is found by putting the data numbers in order and then finding the "middle value". Half of the data values are above the median and the other half of the data values are below the median.

The mean and median are both useful in different situations. Deciding which to use involves determining the purpose of the analysis and the type of outcomes required. When a distribution is normal - in other words resembles a bell curve - the mean and median will be very close together in the center of the data. When the distribution leans heavily to the left or to the right, the mean and median can be very different numbers. It is important then to determine which is more accurate for your purposes as they can be misleading otherwise. The mean number of goals a team scores per game could be 4, while the median is only 2. Looking at the mean number gives inflated values for a team's scoring ability. This discrepancy happens because outliers affect the mean more heavily than the median. Outliers are the outcomes that occur very rarely, such as a 10-point game. The one 10-point game distorts the mean of the data but not the median.

The mean is only used with quantitative data, or data involving numbers, while the median can be used for qualitative data as well. Data can be assigned a value; for example, a win is assigned the value of 2, a tie the value of 1, and a loss the value of 0. Finding the median for this information is possible; however, the mean would not apply.

The mean and median have been used prodigiously in sports statistics throughout the years. They provide statistics that are straightforward so an analyst can easily discuss them with the fans and coaches can easily discuss them with their players. They are very basic tools but, at the same time, very helpful.
Mixed Integer Programming

An integer programming problem is a mathematical optimization or feasibility program in which some or all of the variables are restricted to be integers. In many settings the term refers to integer linear programming (ILP), in which the objective function and the constraints (other than the integer constraints) are linear. Integer programming is NP-complete. In particular, the special case of 0-1 integer linear programming, in which unknowns are binary, and only the restrictions must be satisfied, is one of Karp's 21 NP-complete problems. If some decision variables are not discrete the problem is known as a mixed-integer programming problem >


An Analytical Approach for Fantasy Football Draft and Lineup Management

This is a review of the research conducted by Adrian Becker and Xu Andy Sun.

Fantasy sports are ever growing in their popularity among Americans and Canadians. Of the 30 plus million people playing online fantasy sports approximately 85% play fantasy football. Despite its popularity and increased analysis by experts, there is not a comprehensive strategy for the entire fantasy football season.

In order to predict team and player performance based on historical data a mixed integer programming model is built, which uses predictions for the draft selection and weekly lineup management to create a strategy for winning a fantasy football season. The model is trained using data from the 2004-2006 seasons and then used to simulate the 2007 and 2008 seasons.

In a fantasy football game, 10 to 20 individuals act as owners to create and manage teams made up of actual NFL players. Every week, the owners are paired up against each other and score points based on the actual performance of the players on their team. The owner whose team gains the greatest number of points in that week is declared the winner. The season is made up of three phases: the draft, weekly play, and playoffs.

The draft occurs before the official start of the NFL season. Individuals are randomly assigned a pick order and then select NFL players to be on their fantasy team. Predictions and opinions of the media form the basis for the model for opponents' drafting behavior. During each round, the individual makes a draft decision using an integer optimization model whose objective is to maximize the number of wins in the season as well as the total points scored by their team.

Weekly play takes place during the first 15 weeks of the NFL regular season, with playoffs taking place the last two weeks. Each week the owners select which players from their roster will start that week based on their opponent and then score points based on the actual performance of their starting line. The top four owners advance to the playoffs while the fifth through eighth placed owners go to a consolation playoff. A simply greedy algorithm is used to select the starting lineups.
Next, a prediction method to estimate the weekly fantasy points earned by a player or defensive team and the fantasy points needed to be reasonably certain that a team will win is created. An index is used to represent the idea of the innate ability of a player or defensive team and then used to allocate experts' projections of fantasy points for the entire season to each week. A weighted average of historical performance is used with recent performance being more heavily weighted using a least squares estimation method. The least squares estimation is nonlinear and therefore an alternating minimization algorithm is employed to resolve the issue.

A simulation model is developed to measure the model parameters and test performance. The model simulates all three phases incorporating three aspects of variability: the individual's draft order, the weekly match-up schedule, and the opposing individuals' draft picks.

While these have been created for use within the realm of fantasy football they can be incorporated by management of actual teams. During the off-season managers will be evaluating reserve players from other teams, looking to fill specific positional needs. As these players are not regular playing members of a team, their experience at the professional level is typically limited. Therefore, these players need to be evaluated in terms of their opponents. In addition, coaches need to prepare their team for their weekly games, preparing for a specific opponent, in order to win games and advance to the playoffs.

Analytics methods used in this research: Mixed Integer Programming, Integer Optimization Model, Simple Greedy Algorithm, Weight Average, Least Squares Estimation, Alternating Minimization Algorithm
Mixed-Integer Linear Programming

An integer programming problem is a mathematical optimization or feasibility program in which some or all of the variables are restricted to be integers. In many settings the term refers to integer linear programming (ILP), in which the objective function and the constraints (other than the integer constraints) are linear. Integer programming is NP-complete. In particular, the special case of 0-1 integer linear programming, in which unknowns are binary, and only the restrictions must be satisfied, is one of Karp's 21 NP-complete problems. If some decision variables are not discrete the problem is known as a mixed-integer programming problem. 


Alleviating Competitive Imbalance in NFL Schedules: An Integer-Programming Approach

The National Football League has a strong fan base and generates more revenue than any other sports league in the world. During a regular season, NFL games are scheduled primarily on Sundays with the exception of two or three additional games on Thursday and Monday nights. The NFL regular season lasts 17 weeks with each team playing sixteen weeks with one bye week. Having a scheduled game on a Thursday provides a team with three or four extra days for rest and practice while a bye week provides an extra week to prepare for the next match. Due to this, teams can play against more rested opponents, giving the opponent a competitive edge.

The NFL incorporates a variety of complex rules to schedule a season's worth of games in order to maintain fairness for the teams. However, every year several teams feel that their schedule does not provide them with the same fair opportunities afforded to other teams.

When comparing average win percentages against all opponents versus average win percentages against rested opponents it becomes clear that teams are less likely to win when playing a rested opponent.

A mixed-integer linear programming model is developed to create a schedule which minimizes the number of games in which a team plays against a rested opponent as well as minimizing any long periods of consecutive home or away games.

This model incorporates the thirteen rules currently used by the NFL in determining its schedule as well as an additional nine rules to help create fairer NFL schedules. The model incorporates these rules into designing the schedule while looking to minimize the number of games teams play against a more rested opponent, the number of teams playing three consecutive road games, and the number of teams playing three different sets of back-to-back road games in the course of a season.

It is also essential that the model is able to generate alternative schedules in a reasonable length of time, which can be difficult due to the sheer amount of data and parameters. To handle this, a two-phase model is incorporated which breaks the full model into two simpler programs, solving them consecutively. The first phase assigns the games to weeks without determining the venue and also
assigns bye weeks to each team. The second phase then determines who will host each game, taking into account all of the parameters. Breaking the model into two phases increases its efficiency, thus increasing the speed with which it can generate schedules.

Running experiments with the model determines that the schedules generated by the model, as compared to past NFL schedules, treat teams on a fairer basis in regards to playing better-rested teams.

The NFL can use this model to create a schedule that is fair to all teams, minimizing the bias regarding playing better-rested teams. This would help ensure teams win due to their skill and not because they had a longer time to recover between games.

Analysts could compare schedules created by the model with past NFL schedules to determine what effect, if any, it would have had regarding which teams made to the playoffs and which team was the ultimate winner of the season.

Analytics methods used in this research: Mixed-Integer Linear Programming Model, Two-Phase Method.

Are Daily Fantasy Sports Gambling?

This is a review of the sports gambling research conducted by Todd Easton and Sarah Newell, applying mixed-integer linear programming.

We are a culture obsessed with our sports, from a very young age and throughout our lives. We cheer sometimes and boo at others. As fans, we often feel we could do better than the coaches could. This has led to the popularity of fantasy sports where fans act as an owner or general manager and select athletes to form their team. The athletes compile fantasy points based on how they perform in the real world. The fan whose team generates the greatest number of fantasy sports ultimately wins.

Many states have now deemed that fantasy sports are gambling, consequently creating additional regulations to monitor the industry.

For an activity to be categorized as gambling it must meet three criteria: players must decide to participate in the activity, players must pay to partake in the activity, and the outcome is based on chance. Fantasy sports clearly meet the first two criteria and the question is whether they meet the third criteria.

An argument can be made that fantasy sports is based on skill and not on chance. If chance is the sole factor in determining the winner, then skilled and unskilled players have an equal likelihood of winning.

This study looks at fantasy points in one particular site using random teams to model those chosen by unskilled participants. Random teams are created using simulation. Teams modelling those chosen by skilled participants are created using an integer-programming model, which optimizes the expected fantasy points earned by the team.

Each week, for ten weeks, 100 random teams are created and the average fantasy points generated is calculated and compared to the team created by the integer programming model. Comparing the outcomes clearly shows that the random teams fair much poorer than the skilled team. Analysis
determines that if each random team played the skilled team they would win only 70 out of 1000 games.

To further the study the random teams are entered into an actual fantasy sport site. Again, the hypothesis is that if winning is strictly based on chance then the random teams should have an equal likelihood of winning. Random teams were entered into 35 different contests ranging from 50 to 2500 contestants. None of the random teams won any sort of payout. The odds of this happening are astronomically small, indicating that winning is not based solely on chance, but on skill.

This information can be used by those who are opposed to the laws being enacted by the various states regulating fantasy sports as a gambling enterprise. This information clearly refutes the idea that fantasy sports is based on chance but, instead, is based on skill. Consequently, fantasy sports do not meet the three criteria of gambling and should not be regulated as such.

Analytics methods used in this research: Simulation, Integer Programming Model
Modern Portfolio Theory

Modern portfolio theory, or mean-variance analysis, is a mathematical framework for assembling a portfolio of assets such that the expected return is maximized for a given level of risk.

Three Point Shooting and Efficient Mixed Strategies: A Portfolio Management Approach

This is a review of the research conducted by Mark Fichman and John O’Brien.

Sports all have rules designed to manage the competitors’ actions and style of play. Over time, rules change to optimize competitiveness and fan enjoyment. Rule changes occur often in the NBA. The biggest change made in the last few decades was adding the 3-point shot. The rule change did not have an immediate impact and game strategy continues to adjust. Teams continue to utilize the 3-point shot more regularly.

This research analyzes this adaptation of game strategy, which is then used to predict future game strategy in the NBA with regards to 2- and 3-point shots.

The mean and variance of successful 2-point shots has remained stable over time. However, the mean of successful 3-point shots has increased while the variance has decreased. The 3-point shot variance continues to decrease but still remains higher than the 2-point shot variance. This indicates that while point expectancy is important in team strategy, so is the risk associated with each type of shot.

A finance approach is utilized to deal with the mean, variance, and covariance of the expected payoffs of successful 2-point and 3-point shots, treating team strategy like a portfolio with its unique combination of risk and return. Modern Portfolio Theory is used to develop explanations and predictions of 3-point trends for the NBA as a whole while Capital Asset Portfolio Theory is used at the team level in describing offensive and defensive strategies.

Treating 2- and 3-point shots as assets in a portfolio allows an analysis of risk and return. When the 3-point shot was first introduced, the risk was greater than the return so it was rarely used. However, over the years as the risk has decreased and the return increased the percentage of 3-point shots taken has increased.

A Sharpe Ratio is used to identify an efficient mixed strategy in regards to 2- and 3-pointers. The idea is to maximize the Sharpe Ratio, which is determined by dividing expected points by point volatility. Strategy is determined for both offense and defense. Offensive strategy benefits from a large Sharpe Ratio, which indicates that their expected points per unit of risk is greater than the opponents’. The opposite is true for the defense.

The correlation between 2- and 3-point shot successes is negative and as strategy has evolved to include the 3-point shot, the negative correlation becomes stronger. This indicates that a mixed strategy involving both 2- and 3-point shots is preferred to one that utilizes only one type of shot.
Analysts can this information when predicting which teams will experience post-season success. It is also useful when determining optimal strategies when playing against a particular team. The theory can be adjusted by teams to analyze other shot types to compare their risk and return to further develop effective strategies. Leagues can use predictions for optimal mixed strategy to predict the effect rule changes will have on competition, helping determine if the rule change will, in fact, be beneficial for the league and its fans.

It is important to understand that the idea is not to tell teams when they should make 2-point or 3-point shots but rather what percentage of shots should be 2-pointers and 3-pointers.

Analytics methods used in this research: Modern Portfolio Theory, Capital Asset Portfolio Theory, Sharpe Ratio
Monte Carlo Simulation

Monte Carlo methods (or Monte Carlo experiments) are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. Their essential idea is using randomness to solve problems that might be deterministic in principle. They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other approaches. Monte Carlo methods are mainly used in three problem classes: optimization, numerical integration, and generating draws from a probability distribution.

Source - https://en.wikipedia.org/wiki/Monte_Carlo_method

Misadventures in Monte Carlo

This is a review of the basketball research conducted by Richard Demsyn-Jones, applying Monte Carlo simulation.

Basketball coaches, management, players, and fans are invested in their team's probability of making it to the playoffs. Beyond these groups, betters wager large amounts of money based on a team's chances and academics are always looking for better, more accurate ways to calculate the probability of a team winning the championship.

Statistical models analyze individual games rather than looking at which teams will make it to the playoffs. The large number of possibilities makes it extremely difficult to directly calculate playoff probabilities. Therefore, simulating the season multiple times provides estimates that are more accurate. This is done through a Monte Carlo simulation, but there are some issues with this process. Monte Carlo estimates can be mistaken for accuracy in the underlying probabilities themselves. When running a Monte Carlo simulation, any variance in the underlying probabilities needs to go through the simulation as well; otherwise, the results can be very misleading.

A Monte Carlo simulation involves repeated random sampling to estimate expected value of a function. It is useful in situations where mean, variance, and other statistics cannot be calculated directly. The quality of a Monte Carlo analysis is dependent on the accuracy of the input. Errors in the input will result in errors in the output.

Often, information from quality models is used as input for a Monte Carlo simulation without taking into account any error correlations that always accompany such models. These simulations have been known to estimate that a team has a 100% probability of making it to the playoffs or a 0% chance. These absolutes are not viable outcomes, therefore there was some error correlation not accounted for within the simulation.

One way of predicting which teams will make it to the playoffs is to make a prediction for each game within a season and then use Monte Carlo simulations to estimate which teams are likely to have successful seasons, earning entrance into the playoffs. A problem comes about when player quality and minutes played are estimated, yet used as absolutes in the simulation. Stochastic models have an
inherent variance due to the randomness of the event itself but there is also variance within the accuracy of the estimates. If the variance of estimate accuracy is not accounted for within the Monte Carlo simulation, the simulation assumes that the estimates are accurate, and the error multiplies every time the simulation is run. This leads to results that are overly certain that likely events will occur and unlikely events will not.

In order to improve the process, an error correlation needs to be incorporated within the model. Rather than keeping the predictions regarding each player’s quality and minutes constant, they are allowed to vary, generating new error terms for each simulated season. The results are less certain but are better standardized, not creating the absolute estimates that are formed when errors are not incorporated.

This model is useful for coaches, analysts and fans alike, providing better estimates of a team’s probability of making it to the playoffs. Analysts will be able to provide better predictions at the start of a season. Coaches and teams will be able to look at training and possible player trades before it is too late to make changes that will improve the probability of competing in the playoffs.

Analytics methods used in this research: Monte Carlo Simulation
Multiple Imputation

Multiple imputation (MI) is a way to deal with nonresponse bias — missing research data that happens when people fail to respond to a survey. The technique allows you to analyze incomplete data with regular data analysis tools like a t-test or ANOVA.

https://www.statisticshowto.datasciencecentral.com/multiple-imputation/

An Examination of Statistical Disclosure Issues Related to Publication of Aggregate Statistics in the Presence of a Known Subset of the Dataset using Baseball Hall of Fame Ballots

This is a review of the baseball hall of fame ballot research conducted by Gregory J. Matthews, Petala Gardenia da Silva Estrela Tuy, and Robert K. Arthur.

We live in a world of data. Data comes in all varieties and is often easily accessible, improving the lives of many. However, as data becomes more readily available the issue regarding privacy arises. Not all data should be openly available to anyone and everyone. The question is when should data be available to the public and when should privacy be respected.

One such question arises in regards to Baseball Hall of Fame ballots. The Baseball Hall of Fame was created to honor those players who make rich contributions to the sport of baseball. A screening committee puts together the ballot of eligible players for possible induction into the Hall of Fame. The approximately 625 members of the Baseball Writers Association of America can then vote for up to 10 players on the ballot. Any player who receives votes on at least 75% of the ballots is then inducted into the Hall of Fame. The results of the ballots are released in aggregate form only, as well as the names of all voters who cast a ballot.

However, many of the voters publish their votes for the public to view. Others do not. A snooping experiment is carried out to learn as much as possible about the votes that are not disclosed to the public. Of particular interest are the voting habits of those who kept their ballots private as well as differences between the voters who publish their votes and those who do not. This is done with a multiple imputation restricted based on marginal counts and latent class analysis.

For this experiment, data was collected for the 2014, 2015, and 2016 votes. Each voter is labelled as being part of the known or unknown group. Taking the vote totals for each player as released by the Baseball Hall of Fame and the publicly released ballots, the probability that an unknown voter voted for a particular player is determined.

An odds ratio between the known and unknown groups is calculated to look at the differences between how the two groups voted for a particular player. A Fisher’s exact test was performed to test the null hypothesis, which states that there is no statistically significant difference in the proportion of votes received by a player between the known and unknown groups. Out of 57 tests performed on the 2014 and 2015 data, only eight were rejected leading to the conclusion that there is a statically significant difference.
In order to determine the overall differences in voting patterns, multivariate results are produced using latent class analysis. A chained equation approach was used to input the ballots belonging to the unknown group with each ballot seen as an observation and each player on the ballot seen as a variable. A logistic regression model is then built for each player in order to predict the probability that an unknown voter voted for a specific player. The results show that a major difference between the known and unknown voters is that those voters who kept their ballots secret were less likely to vote for players whose names were connected in some way with performance enhancing drugs.

Analysts could use this information to look more closely at other possible ways players whose names are connected with performance enhancing drugs are treated differently. This is especially important, since rumors connecting players with drugs are not always accurate.

Analytics methods used in this research: Multiple Imputation Restricted based on Marginal Counts, Latent Class Analysis, Fisher's Exact Test, Chained Equation, Logistic Regression Model
Multivariate Statistics

Multivariate Statistics as a Method of Sports Analytics

Many statistical methods can be performed using ordinary pocket calculators. But, this is not the case with multivariate statistical analysis. For its execution, a computer is needed and a statistical program, which are offered on the market. In such conditions, even untrained and inexperienced people can perform complex multivariate statistical procedures, which is a double-edged sword, because, although the result is relatively easy and fast, there is a great chance to make a mistake.

Multivariate statistical analysis has been present in the sciences for almost a century. However, its application in economic research began in the late 1950s. Eventually, applications of multivariate analysis have become more and more frequent since they were increasingly appreciated by both scientists and businesses.

Prior to multivariate statistical analysis, most researcher used analysis that treated at most two variables at the same time. As a product of such analysis, results were most commonly reported as central tendencies (arithmetic mean, modus, median ...), variation measures (variance, standard deviations, quarters...), confidence intervals and tests based on a normal schedule, t-schedule and similarly. The longest range in the study of the relationship of two phenomena was the correlation coefficient.

Multivariate statistical analysis has provided much more powerful techniques that enabled researchers to detect patterns of behavior in the interrelation of a large number of variables, patterns that would otherwise be hidden or barely noticeable.

In sports, multivariate statistical analysis can be used to determine the development of functional abilities of a group of soccer oriented athletes. In the experimental group of soccer-oriented respondents, the "circular" form of exercise was implemented for the additional 33 hours of motor exercises. Determining the load level as part of modeling the program for functional abilities development was in accordance with the individual abilities and characteristics of the respondents. Particular care was taken to ensure that the dosage of the load has a gradual and progressive character in all its components (intensity and extensiveness). The selection of the methods of exercise applied in the "circular" form of exercise for the functional abilities development was in the function of achieving goals and tasks, raising the level of preparedness, respecting the age characteristics and conditions in which the experimental program was realized. The organizational form of the "circular" form of work was carried out within homogenized groups. Transformation of functional abilities in both subunits during experimental treatment was determined by analyzing variance at the multivariate level.

The other purpose of multivariate statistical analysis in sport is determining which team has chances of winning in a competition. This is achieved by using principal component and cluster analysis based on the previous results of every sport’s team. After determining the principal components, first and second were used as new data and cluster analysis was used to divide them into two groups. The multivariate statistical analysis made it possible to crystalize the more and less successful teams within the groups.
Multivariate statistical analysis in sports is most appropriate when a researcher wants to analyze the relationships between multiple variables (more than two), and simultaneously according to the appropriate model on which this technique is based.
**Negative Binomial Regression Model**

Negative binomial regression is a type of generalized linear model in which the dependent variable is a count of the number of times an event occurs.

Source - https://www.mathematica-journal.com/2013/06/negative-binomial-regression/

**The Nature of Regional Bias in Heisman Voting**

This is a review of the negative binomial regression model research on Heisman voting conducted Nolan Kopkin. The Heisman Trophy is the top award given to a college football player every year. However, many speculate that there is a bias in the voting process. The country is split into six regions, Northeast, Mid-Atlantic, South, Southwest, Mid-west, and Far West, with each region having 143 votes to distribute among its media. Former Heisman Trophy winners are also given a vote. The public is given an aggregate vote via online polling.

Theoretically, it is expected that a regional bias exists when the finalist and voters are from the same region and that finalists will receive more votes from nearby regions than finalists based further away. This impact would decrease if a region has multiple finalists, which would split the vote. The location of the finalist’s opponents also plays a factor, as well as national media coverage.

Data was collected from 1990-2016 and analyzed to produce summary statistics. Each piece of data was weighted by the inverse of the number of finalists in that particular year so that all years are weighted equally. The statistical significance of each difference between in-region and out-of-region samples was tested using chi-squared tests. The stats demonstrate that voters do favor players who are based in their region as well as players who play more games against regional opponents.

Boxplots clearly show the bias regarding in-region players. When voting for out-of-region finalists, voters from the Mid-Atlantic are more likely to vote for players from the Northeast than the other four regions. They are then more likely to vote for players from the South than the remaining three regions. The Midwest is less likely to vote for finalists from the South and Far West voters are more likely to vote for Midwest finalists than those from other regions.

A negative binomial regression model is used to analyze regional vote tallies in order to handle the skewed, discrete data. Results indicate that finalists receive more points from their home region than from other regions. Games played in region lead to a higher tally count as well. Increased national media coverage leads to an overall decrease in regional bias.

Results from a sensitivity analysis show that the fact that players receive more votes from their own region is standard across all six regions. The idea that finalists from the Far West receive fewer votes than those from other regions is shown to be false and that, in actuality, finalists from the Northeast, South, and Southwest tend to receive fewer votes. The Northeast demonstrates the strongest in-region bias.
Analysis demonstrates that bias is definitely prevalent within the Heisman Trophy voting. The fact that national media coverage decreases the bias leads to the question of whether there are additional methods that would aide in decreasing the bias. This would be a worthwhile endeavor for analysts to research as the ultimate purpose of the Heisman Trophy is to award the best player overall.

Analytics methods used in this research: Weighting, Chi-Square, Boxplots, Negative Binomials Regression, Sensitivity Analysis
Network Analytics

Network Analytics as a Method of Sport Analytics

Network analytics has been used by various fields for a very long time. It is used in process control networks which run across the manufacturing sector, utilities, health, and even in sports. The effectiveness of network analytics is so undisputed as it makes industrial networks quite safe, offers high productivity, and is much less expensive to operate.

Before we go ahead to talk about network analytics as a method of sport analytics, let’s see some design examples of this method and how it is used.

Analytics utilizes process control which are embedded in sensors and controllers that use simple commands to provide operations. For instance, the command could be: “if the pressure in the boiler is above X bars, then shut down the boiler”. So basically, it follows an “if and then” operation. By expansion of this, one can now build a model for analytics computation, which will enable the provision of solutions to complex problems.

Network analytics allows the computer to compare data that are inputted against data sets that are pre-programmed in the operational models. This comparison helps the computer make functional decisions in a way that it improves the operational process.

Sport analysts use network analytics to build a model around sports events, which are now compared to certain information that are already pre-programmed in the computer with the aim of getting certain functional decisions. These decisions help the analysts and coaches optimize the team and understand the variation in performances of teams and the factors responsible for them.

Presently, coaches and sport analysts employ the use of machine learning, which is augmented with network analytics in other to get their intended information. This is the reason why their analytics models are frequently updated and work in real time because the conditions in the network are constantly changing. However, some analysts prefer to only apply network analytics without the involvement of machine learning.

Let’s see how network analytics is benefitting sport analysts. There are many ways this method of sport analytics has shown great advantages. However, we will only discuss the three most common uses. They are (1) optimization of network resources, (2) provisioning of new revenue streams, and (3) market acceleration.

Optimization of Network Resources

This part is particularly helpful to coaches because it helps them manage resources provided by the team or club owner. When coaches have gathered their resources, they are always looking for the best way to use them in other to get optimum output. By applying network analytics to sports, coaches are able to understand their network, weigh their resources and utilize them in the best ways that yield higher performances and lower cost.

Provisioning of New Revenue Streams
Beyond optimization of resources, teams are always in the habit of finding ways to generate new streams of income for the club. So they employ network analytics to meet their needs. Network analytics open up new income opportunities by providing mining insights, which help the managers identify new streams and then build data which are driven for the sole aim of business.

*Market Acceleration*

When coaches are in need of new resources to increase the performance of their teams, they make use of this technique, because network analytics helps them plan for growth easier and the resources needed can easily be estimated, so that as soon as the market is declared open, they already know what they need.

In conclusion, network analytics is very beneficial for coaches, analysts and owners as it provides optimal ways for upgrading their teams and revenue opportunities.
Network Diffusion Methods

Network-based diffusion analysis (NBDA) is a statistical tool to detect and quantify social transmission of information or a behavior in social networks (SNA, etc.). NBDA assumes that social transmission of a behavior follows the social network of associations or interactions among individuals, since individuals who spend a lot of time together, or who interact more have more opportunity to learn from each other.


A Network Diffusion Ranking Family that Includes the Methods of Markov, Massey, and Colley

This is a review of the sports research conducted by Stephen Devlin and Thomas Treloar applying Massey method, Colley’s method and Markov chain.

Ranking is a natural phenomenon within the world of sports. Teams are ranked and so are individual players. In leagues consisting of a small number of teams and playing a large number of games, the ranking process is quite straightforward. However, when a league consists of many teams playing relatively few games a unique problem arises when ranking teams who have not faced each other during the season nor played any common opponents. The Massey method, Colley method, and Markov method are three methods used for ranking in sports situations. A framework is created to understand the three methods and their similarities and differences.

The Massey method is based on the idea that the difference in two teams' ranks should predict the point differential in a game featuring the two teams. It employs a least squares ranking approach.

Colley's method uses a modified winning percentage. The percentage is found in the form of one plus the number of wins divided by two plus the number of games played.

The Markov chain method represents each team as a node in a network with edges between teams that play one another. The Markov rating is normalized by the number of games played by the team in order to remove any bias related to teams having played different numbers of games.

A one parameter family of rankings interpolates between the Markov method with p equal to zero and the Colley and Massey methods with p equal to one. The distribution process is determined by specifying the flow of rank from one team to another in terms of face-to-face wins, weighted by face-to-face losses, and a rank-infusion vector, which is determined by the team's overall record. This allows similarities and differences between the three methods to be analyzed especially in regards to the parameters chosen and the normalizations made by each method.

Analysis shows that there appears to be no single optimal value for p. To make face-to-face results relevant to a team's ranking, beyond their winning average, the value of p needs to be less than one. If
an emphasis is to be placed on a team's record and strength of schedule, then $p$ should equal one. It is also likely that an optimal value for $p$ will vary depending on the sport or even the league.

The Massey and Colley methods produce ratings in a perfect season that are equally spaced and highly stable while the Markov rating is non-uniform in its distribution of values and therefore less stable.

Analytics methods used in this research: Massey's Method, Colley's Method, Markov Method, Least Squares Ranking
Network-Style Model

The network model is a database model conceived as a flexible way of representing objects and their relationships. Its distinguishing feature is that the schema, viewed as a graph in which object types are nodes and relationship types are arcs, is not restricted to being a hierarchy or lattice.


A Method for Using Player Tracking Data in Basketball to Learn Player Skills and Predict team Performance

This is a review of the tracking data and network model research conducted by Brian Skinner and Stephen J. Guy.

Since the 2013/2014 season, all NBA arenas have installed a system of cameras and tracking software. These systems provide a wealth of information that can then be analyzed. This information includes both quantifiable and non-quantifiable skills. Quantifiable skills include points, rebounds, assists, steals, and blocks. Non-quantifiable skills include high quality (non-assist) passes, setting up good screens or effective rotation on defense. This tracking information is not currently available to the public. If, or when, it does become available these researchers suggest creating a network model to analyze this information. It is possible to create a network model of a basketball offense that relates players’ skills to the team’s success at running different plays. If the tracking information were inputted into such a model the output would give analysts and teams the capability to predict the ability and effectiveness of any 5-man lineups.

Traditionally, analysts spent a considerable amount of time watching players perform in order to evaluate different aspects of their performance, with an emphasis on the non-quantifiable aspects that could not be seen on the score sheet. This combination of tracking data and network model would provide analysts with unbiased information of the different players’ skills levels and how they would work together as a team. Every aspect of the game would be open to evaluation. Analysts could learn the skill levels of NBA players and use this to make predictions regarding the team’s performance in the upcoming season.

Using the tracking data and network model combination allows coaches and management to predict how their team is affected by a variety of changes, including trades, substitutions, or alternative plays. Coaches are able to input different line combinations into the model giving them the ability to determine the most effective way to use their players in order to improve their offensive performance. This system also provides the opportunity to determine which plays are the most effective against any given opponent. Teams can determine how players would work together, giving them increased insight when making personnel decisions such as who should be traded. Management can input trade prospects into the model to determine if their abilities would be effective in improving the team as a whole.
This application is just waiting for the tracking information to become available. It can also be adapted to look at defensive capabilities.

Analytics methods used in this research: network-style model, high/low model, statistical inference algorithm
Neural Network Algorithm

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Neural networks can adapt to changing input so the network generates the best possible result without needing to redesign the output criteria.

Source - https://www.investopedia.com/terms/n/neuralnetwork.asp

(batter/pitcher) 2vec: Static-Free Talent Modeling With Neural Player Embeddings

This is a review of Neural Player Embedding research conducted by Michael A. Alcorn.

The game of baseball is full of statistics. These can be very helpful but often do not accurately describe the abilities of individual players. These statistics also do not take into account the habits and idiosyncrasies of each player. The network algorithm (batter/pitcher) 2vec was created to help overcome this weakness. It has the capability to include this non-quantifiable information into its analysis process. Data was collected from observations of individual players as well as each individual play rather than from aggregate statistics. Aggregate statistics have already summarized those observations and therefore are not as detailed and accurate. (batter/pitcher) 2vec predicts the outcome of an at-bat, taking into account who is at bat and who is pitching.

The data input into the algorithm included the batter, pitcher and at-bat outcome such as strike out, home run, etc. Data was obtained from each game that was played in the 2013 - 2016 seasons. The data set included over 461 thousand at-bats, 524 different batters and 565 different pitchers.

The algorithm also has the ability to compare players and determine which players are most alike. Some of the pairings were what you expect such as Craig Kimbrel and Aroldis Chapman who are both elite closers. However, some pairings were unexpected such as finding opposite-handed doppelgangers. Mike Trout (a right-handed batter) is partnered with David Ortiz (a left-handed batter).

Networks like (batter/pitcher) 2vec are able to generalize and transfer observations to other situations. This network has the ability to determine outcome probabilities for batters and pitchers that have never before faced each other. This information can be used by analysts and coaches alike in determining which line up would be best when facing a specific pitcher.

This algorithm would be a highly beneficial tool for baseball teams. When considering whether to make a trade the team can swap out the names of the players they are proposing to trade. They can then look back at previous games to determine if they would have won or lost more games with that player in the lineup. This can then be used to predict possible future outcomes after completing the trade. If a (batter/pitcher) 2vec algorithm was developed for the minor league, teams could look at players they are thinking of calling up and swapping them into the algorithm just as they did with prospective trades. All of this would give teams another decision-making tool to aide in putting together the strongest team possible and increase their chances of making it to the World Series.
The algorithm \textit{(batter/pitcher) 2vec} is a very versatile and flexible tool. As a result it could easily be adapted for use in other sports such as football and basketball.
Neural Network Algorithm

Artificial neural networks (ANN) or connectionist systems are computing systems inspired by the biological neural networks that constitute animal brains. The neural network itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules.


Classifying NBA Offensive Plays Using Neural Networks

This is a review of the neural networks research conducted by Kuan-Chieh Wang and Richard Zemel.

Offensive strategies in basketball are complex and dynamic, ever changing in response to the defense. Not only do the plays evolve but also players can change roles as the play on the court continues. A further complication is that plays are variable in how long they take to set up and how long they take to execute as they respond not only to the defense but also to where teammates are placed on the court. A turn of events like fouls and player errors further complicate the process. Recognizing the multitude of offensive plays of a team requires a deep understanding of the game. This information cannot be recorded as easily as other data such as number of shots and shot efficiency and therefore requires a more advanced process.

Player tracking data from SportVU was entered into a neural network in picture form. Neural networks are relatively easy to use and are capable of solving difficult problems. Neural networks do not need the various individual features of each play to be coded; instead, they are capable of learning the features themselves by analyzing the data. The pictorial data was separated into individual steps and information regarding the position of every player on the court at that moment was noted. The position of the ball in relation to each player was also recorded. The individual steps were then combined in player position sequences, which are presented as lines on a map of a court. From these lines, the model is able to determine which offensive plays are being utilized. The data used came from the Toronto Raptors 2013-2014 season. Eleven classes of offensive plays were included and 1435 sequences from the data fit into these categories.

The neural network model that was developed based on the 2013-2014 season’s data was tested in the 2014-2015 season. The model proved that with limited data, it was able to distinguish the different offensive plays.

The model requires more information to analyze offensive plays than a scout does. However, they are capable of watching multiple games in a short period of time and are able to provide the usable data within a fraction of the time required by its human counterparts.

Being able to determine and classify the offensive plays allows analysts to scout for talent that would be a good fit for their team. This information will allow coaches to analyze their team's effectiveness in
using the various offensive plays, which will allow them to refine their playbook and provide information
to their players in order to help them improve their play. The information will also create scouting
reports that contain more detailed and useful information.

Analytics methods used in this research: neural network algorithm
Non-Parametric Correlations

Non-Parametric Correlations in Sports Analytics

This chapter discusses the use of Non-Parametric Correlation as a method of sports analytics.

What is Non-Parametric Correlation?

Before we get into how non-parametric correlation is used in sports analytics, we first must understand non-parametric correlations in a textual concept.

Non parametric correlation is a correlation technique that can be used with any variables that can be transformed into ranks. This means that data used for non-parametric correlation methods is often ordinal and does not rely on numbers. Instead it focuses on ranking, relation or some order of some sorts during analysis.

However, for non-parametric correlation to hold, there are two basic assumptions:

- Variables to be measured should be measured at an interval or ratio scale or at an ordinal level.
- There should be a monotonic relation between the two variables to be used.

Electronically, software like SPSS or Minitab are used to compute non-parametric correlations of variables.

Types of Non-Parametric Correlation

There are two types of non-parametric correlation: Spearman’s Correlations and Kendall’s Correlation

The Spearman’s correlation, often denoted by the symbol Rs or the Greek letter ρ, is an analytical method that measures the direction and the strength of association between two variables which are measured on an ordinal scale. For example, you could use spearman’s correlation to know the relation between time spent during training and actual performance of players.

Kendall’ correlation, commonly referred to as Kendall’s Tau, measures the ordinal association between two or more variables to be computed.

Non Parametric Correlation as it Relates to Sports

Now that we’ve established the meanings and types of non-parametric correlation, let’s see how they relate to the sports industry. Using the non-parametric method of sports analysis, analysts have and can improve performance of athletes and sustain winning streaks.
Case Study: 2005 Monaco Grand Prix

During the 2005 Monaco Grand Prix, non-parametric correlation method was quickly employed to enable the McLaren team to win. Here is how it happened:

Well into the race, Kimi of McLaren team was leading with a little gap. The third driver, Schumacher accidentally smashed into the second driver and both cars ended up needing repair. Other following drivers approached the turn that was debris filled and the race marshals deployed the safety car.

The ideal thing for drivers to do during the safety car period was to pit, change tires, refuel and continue. However, the McLaren team radioed Kimi, asking him not to pit. Obeying, Kimi fired in a few quick laps and increased his lead with 35 seconds. On the forty second lap, he pitted and came out still with a 13 seconds lead and won the race.

With quick thinking, the McLaren team applied non-parametric correlation to make a decision and win the race. They related variances like fuel, time of tires to last and lap time variances in their ranks to ensure success.

Advantages of Non-Parametric Correlation

1. Non-parametric method of sports analysis can be done quickly. As seen in the case study above, the method was applied during the course of the race and still ensured victory.

2. Can be computed without relating numerical variables. In non-parametric correlation, analysis can be done without having to relate numerical variables. Events carried out can be used in its computation.

3. No need for multiple data. Using this method, few data are required to carry out an analysis.

Summarily, with the use of non-parametric correlation methods in sports analytics, sports analysis can be done quickly and still maintain expected performance results.
Normal Likelihood Sampling Model

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model, given observations. The method obtains the parameter estimates by finding the parameter values that maximize the likelihood function. The method of maximum likelihood is used with a wide range of statistical analyses.

Source - https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

Position Importance in NCAA Football

This is a review of the NCAA football research conducted by Zachary Knowlton and Gilbert W. Fellingham.

College football coaches are always looking for those players who will give them the best chance of winning a championship. Knowing which positions are the most important on the team would aide coaches in this process. This research develops a process for evaluating all players on a football team to determine which position has the greatest impact on the team's performance, specifically in the areas of scoring points and prevent the opponent from scoring points.

The required data must include a way to quantify the team's production during each play and a method for rating each player involved in each play. Quantifying team's performance is accomplished through an expected points model built using data from NCAA Football Bowl Subdivision from the years 2005-2013. The BYU football team is used to determine player importance.

In order to rate players and estimate position importance, BYU coaches graded each of the players on the team. Defensive players were graded on a plus/minus scale with players receiving a minus if they did not fulfill their assignment on a particular play and a plus otherwise. The minus was translated as a zero and the plus into a one for the model. Offensive players were graded on a three-point scale with a zero meaning the player clearly did not complete his assignment, a one meaning the assignment was adequately completed and a two indicating an exceptional performance.

A model was built for each system with expected points used as the dependent variable and player ratings the independent variable in order to determine the contribution of each position or player to the team's point production.

To model player performance a normal likelihood sampling model was created to examine each BYU game and the complete season. In order to model position performance a Bayesian hierarchical model was created.

The results indicate that the most important offensive position is the quarterback, which is what you would expect. When the quarterback completed his assignment, it had the largest positive effect on expected points. However, that effect demonstrated a great deal of variability, perhaps due to the quarterback’s reliance on other positions to complete the play. After the quarterback, the wide receiver and running back were the next most influential offensive positions.
On the defense the boundary corner, mike (inside linebacker who calls the plays), and right end were the top three most influential positions. Defensive positions were grouped more closely during a win than a loss, indicating a less consistent defensive effort in losses.

Coaches can use this information to gain an understanding of how the performance of specific positions relates to gaining points or losing them during a game. Comparing positions provides coaches with a guideline for which position players should receive the greatest focus in the recruiting process. It also provides an objective method to compare players competing for one position.

Once the importance of position has been determined, this process could lead to determining which skills have the greatest impact on the team's performance. This would allow coaches to maximize practices by focusing on those skills.

Analytics methods used in this research: Normal Likelihood Sampling Model, Bayesian hierarchical model
Normality Tests

Tests of Normality as a Method of Sports Analytics

This chapter delves deeper into sports analytics by looking at one of the methods of analysis which is now being used in sports – normality tests.

Sport coaches, managers and analysts are often found to attribute the performance of players in a team to various factors such as speed, height, body weight, jump power, etc. This enables the analysts to ascribe a player’s performance to certain factors, and then look out for a trend to create a distribution. However, because the distribution may not be normally well-modeled to truly represent the factor responsible for the performance of the players, a test of normality is carried out to establish the acceptability or rejection of the hypothesis made about each of the players. This allows analysts to determine the particular factor or skill that is largely or lowly responsible for the performance of the players.

Various Techniques Used to Test the Normality of a Distribution

There are many techniques adopted for normality tests, which includes the Chi-square test, Kolmogorov—Smirnov, Lilliefors, Shapiro—Wilks, and the Anderson Darling test. However, in sport analytics, Kolmogorov-Smirnov and the Shapiro—Wilks tests are most often preferred.

The difference between the two tests is the number of values that are present in the overall samples. Shapiro-Wilks test works for values below 50, while Kolmogorov-Smirnov test is used for values greater than 50. For example, after a football game, if the performance of players in two different teams is to be statistically estimated based on certain factors, Shapiro—Wilks test is used because the totality of the players is 22.

How Normal Distribution is Adopted for Sports Performance Testing

Normal distribution is often used to represent random variables whose distributions are not known. For instance, say a rating system has been established for players and by statistical sample, the rating shows that the performance of the bulk of the players in a team is average (50%), while smaller number of players have a rating of 70% or 30%. And even smaller percentage of players have a rating of 90% or 10%.

This statistical data creates a kind of symmetrical graph where half of the data falls to the right of the mean, and the other half to the left. Mind you, this rating system is based on factors such as speed, height, body weight, and jump power. Now, what is the evidence that the normal distribution of the players is true based on the factors mentioned above? This is why sport managers go for the test of normality to determine if the data set such as above has been modeled correctly by the normal distribution.
Shapiro – Wilks Test Applied in Sports Analytics

Shapiro – Wilks test follows a concept of null hypothesis, which works based on two factors—alpha level and p-value. In sport analytics, an example of null hypothesis is: “the height of a player doesn’t determine how many goals the player scores in a match.” This hypothesis is subjected to acceptance or rejection. If perhaps the null hypothesis statement is rejected because it is found to be true, the probability of rejecting such hypothesis is called “significant level” which is set to be 5%. That is, it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis. This significant level is called “alpha level”.

The p-value is called the “probability value” and it is the probability that when null hypothesis is true, the statistical result (for instance the sample mean different between two teams) would be greater than or equal to the actual observed results. It is this p-value that is used to quantity the statistics of normality. If the p-value is less than the alpha level, then the null hypothesis is rejected.
Observation Chi-Square

A chi square (χ²) statistic is a test that measures how expectations compare to actual observed data (or model results). The data used in calculating a chi square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample.

Source - [https://www.investopedia.com/terms/c/chi-square-statistic.asp](https://www.investopedia.com/terms/c/chi-square-statistic.asp)

Side-Out Success and Ways that Points Are Obtained in Women's College Volleyball

This is a review of the research conducted by Jose M. Palao.

In the game of volleyball, the actions of the players on both teams interact together to form a complex system with performance dependent on a variety of factors. Points can be attained in a variety of ways. There appears to be a difference in the probability of success when teams are side-out as opposed to when teams are in serve and defense. An analysis of side-out success tells us that its components are the in service opponent error, first attack side-out success, attack and block counter-attack success and opponent errors. The success of the counter-attack component is dependent on the points in service, first attack opponent error in side-out, attack, block success in counter-attack, and opponent errors.

This study looks at possible ways women's college volleyball teams can achieve success in side-out as well as ways that points are obtained in correlation to the result of the game.

The sample used for the analysis was composed of 2435 rallies from 48 games played in the Missouri Valley Conference in 2008. Information was gathered through observation by an expert observer with coding done in a spreadsheet. A Chi-square test and likelihood ratio regarding the data was completed with respect to descriptive and inferential analyses. The likelihood of data to increase or decrease the probability of winning or losing was symbolized with a plus or minus symbol.

Results showed that winning teams had a great percentage of side-out success. Winning teams also had more success with the counter-attack. The length of the rally had no effect on the outcome. Reception and side-out are related, as when the reception was better so was the side-out. Winning teams experienced greater side-out success using all types of attacks with the exception of the second contact attack, which was equally successful for both teams. Successful attacks not only result in points, they also prevent the opposition from forming a successful counter-attack.

In regards to defense, winning teams gained more points from counter-attacks and blocks, as well as from opponent errors.

Taken together, the data points to side-out and defense strength being strong indicators of success.

Coaches can use this information to structure team practices to build the skills most related to winning success. In the majority of rallies, the ball crosses over the net fewer than two times, which indicates that coaches need to focus on improving players' concentration and quick reactions.
Analytics methods used in this research: Observation Chi-Square, Likelihood
Ordered Logit Model

An ordered logit model (also ordered logistic regression or proportional odds model), is an ordinal regression model—that is, a regression model for ordinal dependent variables—


Does Artificial Grass Affect the Competitive Balance in Major League Soccer?

This is a review of the MLS artificial grass research conducted by Matthew J. Trombley, applying an ordered logit model.

Soccer is a game that has traditionally been played on a natural grass surface. However, an artificial surface is becoming more common. Many players and coaches maintain a negative attitude towards the artificial surface. This attitude stems from three perceptions. The first is that artificial surfaces increase the risk of player injury, to the point that some star players are known to refuse to play games taking place on an artificial surface. The second is that it is more tiring to play on. The third perception is that the ball acts differently than it does on a natural grass surface by moving faster and bouncing higher. These perceptions may influence players’ actions during games played on artificial turf, whether or not the perceptions are true.

A descriptive analysis is completed comparing results of home and away games played on natural and artificial surfaces. Data includes points earned per game, the probability of a win, tie, or loss, goals scored, and goals conceded. The statistics display no significant difference between teams whether they have artificial or natural turf. However, a couple of factors complicate this observation. One is that the greatest influence on the probability of a team winning a game is the skill of the players. Other factors include the fact that most artificial surfaces are in the stadiums of expansion teams, typically weaker in their first years of play and that teams often display a competitive advantage at home. These variables need to be studied through a rigorous regression analysis to determine any true relationships.

The sequence of win, lose, draw and the assumption that the relationship between each outcome is the same is tested through an ordered logit model. The chi-squared statistic indicates that this model is appropriate for use in estimating match outcomes. The estimates determined by the model indicate that an artificial surface is not correlated with the outcome of any game.

A restricted generalized Poisson model is utilized to analyze goals scored during games. While the results indicate that artificial turf is negatively correlated with goals scored by visiting teams and positively correlated with goals conceded by visiting teams, both results are statistically insignificant.

Overall, the results indicate that players’ negative connotation of artificial turf are based on false assumptions. Given this, coaches can use this information to work on realigning the perceptions of their players, encouraging a consistent level of play, whether on natural or artificial surfaces. Analysts can compare these outcomes with the effects the negative perceptions have on the play of individual players. It needs to be determined what effect these negative perceptions have on the psyche of the
player, and consequently on their level of play. Leagues will then need to decide if using artificial turf is of any advantage in the long run.

Analytics methods used in this research: Ordered Logit Model, Restricted Generalized Poisson Model
Ordinary Least Squares Regression

In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function.


Heterogeneity and Team Performance: Evaluating the Effect of Cultural Diversity in the World's Top Soccer League

This is a review of the soccer research conducted by Keith Ingersoll, Edmund Malesky, and Sebastian M. Saiegh applying ordinary least squares regression.

Diversity is a hot topic across all aspects of society, including sports. Teams need to balance the costs and benefits of diversity. Costs include dealing with language and cultural barriers while benefits are a diversity of talents, perspectives and experiences that help facilitate creative problem solving in team situations. The question ultimately is whether diversity affects soccer teams' performance.

Attempting to answer this question is done by examining the teams in the top European soccer leagues, namely England, France, Germany, Italy, and Spain and their achievements at the Union of European Football Associations Champions League tournament between the years 2003 to 2012.

Critical to this research is the fact that players from fifty countries play for these teams. Teams who qualify to play in the tournament are the most influential, wealthiest, and talented soccer teams in the world. This wealth allows these teams to build scouting programs that scour the world for the best possible players.

Players from different parts of the world experience different playing styles, coaching styles and strategies as they grow in the sport through the years. Each player then brings this unique set of abilities with them to their team, broadening the knowledge base of the team, providing new ideas that could benefit the team and give them that needed edge to outscore their competitors.

There are also costs associated with hiring players from other countries. The most obvious one is the language barrier, necessitating additional personnel and training to overcome this obstacle. The language barrier also increases the possibility of miscommunication among the players on the field during a game. Teams consisting of players from a wide variety of ethnicities may experience greater discontent within the team as conflicts arise due to different beliefs and cultural backgrounds.

As most soccer games are low scoring, a goal differential is used as the main variable, rather than goals scored. Not every team plays the same number of games during a tournament and in order to combat...
any biases this can create the average goal differential per game played during the year is calculated for each team. Average points obtained and winning percentages are also taken into account.

Examining diversity goes beyond looking at the numbers of foreign players on a team to the scope of their differences. This was determined using linguistic difference as a measurement within the model.

The model determines the per-game goal differential of each team, taking into account the variables of diversity, average linguistic distance between the players and wealth of the team. This measure of team performance is assessed using ordinary least squares regression. The results of this regression clearly indicate that diversity is strongly correlated with team performance and that teams with greater diversity outperform their less diverse counterparts.

Teams can look at this information and examine the diversity within their own teams. It will help them decide if gaining greater diversity among their players would provide greater benefits than costs, allowing the team to reach even greater heights.

Analytics methods used in this research: Ordinary Least Squares Regression
PageRank and Elo

Elo rankings, which were designed to rank players in a fair manner, start off equal and iteratively change two ratings at a time when two players change each other. The strength of a players' opponents is taken into account in Elo rankings, but only in sequence. The key difference between Elo and PageRank is that Elo does not change the ratings of every player after every competition.


Ranking and Prediction of Collegiate Wrestling

This is a review of the collegiate wrestling research conducted by Kristina Gavin Bigsby and Jeffrey W. Ohlmann, applying PageRank and Elo models.

Rankings within college wrestling are very important as they determine seeding and which teams will advance to the postseason. Rankings also provide benchmarks of success for individuals, teams, and programs as well a basis for predicting future outcomes.

Folkstyle, or scholastic wrestling, is divided into ten weight classes, ranging from 125 pounds to 285 pounds. There are several different types of victories, which can be determined by point differentials, match-ending maneuvers, and if a match ends by default, forfeit, or disqualification. To complicate things further two types of competitions exist, dual and tournament. Consequently, the ranking process is confusing and, therefore, not always accurate.

To start, an evaluation of the current ranking systems indicates that none are overly accurate and therefore there is room for improvement. Consequently, two new methods, PageRank and Elo, are analyzed regarding their effectiveness in ranking individual wrestlers.

For evaluation purposes, information on all Division 1 college wrestlers and matches from the 2013-2014 season is utilized. The 149-pound weight class is used for the training process as it contains the largest number of matches and wrestlers. Organizing wrestlers into weight classes causes a problem in itself, as wrestlers are known to compete in more than one weight category throughout the year.

PageRank is flexible, simple to use and takes into account the strength of schedule. It divides importance in a network proportionally among the nodes according to the number of links. In this case, a wrestler gains importance in the network by having many wins and few losses against wrestlers who also have many wins. Wrestlers who competed in more than one weight class receive a ranking in each class. Two processes are used for weighting links. The first is simple point differential in which the link weight is calculated as the difference between the points awarded to the winner and the loser. The second is a discounted point differential in which the time of the season the match took place is included in the analysis.

Elo is also flexible and easy to use. Elo uses prior ratings as a substitute for player strength and the difference in rating between competitors at the time of their match is used to determine an expected win probability. Ratings following the match are calculated based on the difference between this
expected win probability and the actual outcome with higher-ranking points awarded when the underdog wins.

Both PageRank and Elo are improvements over current ranking systems. They are capable of ranking all wrestlers competing in a weight class and can rank a wrestler in more than one class. Additional strengths are that they deal with strength of schedule, are flexible, and can be updated after each match, week, or any other time period.

This ranking system can be used by various wrestling conferences to rank their wrestlers. Analysts can use rankings to look at predicted outcomes of matches. This information provides coaches with a greater ability to prepare their wrestlers for their upcoming matches.

Both PageRank and Elo are improvements over current methods with Elo showing greater strengths that can be built upon in the future.

Analytics methods used in this research: PageRank, Elo
Pairwise Function

Pairwise comparison generally is any process of comparing entities in pairs to judge which of each entity is preferred, or has a greater amount of some quantitative property, or whether or not the two entities are identical.

This function transforms data that are given in wide or long arm-based format (e.g. input format for WinBUGS) to a contrast-based format that is needed as input to R function. The function can transform data with binary, continuous, or generic outcomes as well as incidence rates from arm-based to contrast-based format.

Source - https://en.wikipedia.org/wiki/Pairwise_comparison

Modeling the NCAA Basketball Tournament Selection Process Using a Decision Tree

This is a review of the NCAA research conducted by Shouvik Dutta and Sheldon H. Jacobson.

The NCAA's basketball championship tournament garners a great deal of attention from media and fans alike. However, the method used to determine which teams will play in the tournament is not fully understood. The 32 teams who win their conference tournaments are guaranteed entry while a selection committee chooses the remaining 36 teams. The procedure the committee uses is known but the exact details regarding how the teams are ranked are not. This research looks at a decision-making method that attempts to copy the method used by the selection committee over the past years, predicting which teams will be selected to play.

The committee members have every possible statistic available to use as a guideline but it is unknown which statistics the committee members place a stronger emphasis on when choosing between teams.

In order to build a model that will predict which teams will be selected to play in the tournament, it must be decided which pieces of information should be included. This information will then be used to determine the relative strength of each team. This model uses both expert rankings and performance metrics. Specifically, in terms of expert rankings the Rating Performance Index and Pomeroy rankings are used. In choosing which performance measures to incorporate, the statements made by committee members in the past are analyzed. As a result the number of top 50 wins, number of 200+ losses, strength of schedule, number of games played against the top 100 teams and the record of the team in the last 12 games played prior to the tournament are used.

The selection process is done in two steps. The first step is comparing teams in a pair-wise function, determining which team is the stronger of the two. A team is given 1 point every time they are chosen, and the teams with the most pair-wise points are chosen to compete in the tournament. The decision was made to focus on those teams that may or may not make it into the tournament, excluding those teams who were certain to participate. The evaluation method for determining which team was the stronger in the pair-wise function is modeled as a decision tree.
To evaluate the accuracy of the tree, it was compared to the teams chosen during the years 2012-2016. The model let in one team each year that did not actually participate but was accurate concerning all other chosen teams. When the model was adjusted to fit one year correctly, it was even less accurate for the other years. It appeared that some teams were chosen to play in the tournament with no statistical basis. This is probably due to the human factor and personal opinions and choices.

The factors that appear to be the most important to the committee are the Rating Performance Index combined with the top 50 wins and 200+ losses.

This process would provide analysts with a clear idea of who is likely to be playing in the tournament. Coaches, knowing the factors considered the most important by the committee, would have a chance to work on improving their team's strength in those areas.

A future expansion for this work would be to predict the seeding as well as the selection.

Analytics methods used in this research: Pair-Wise Function, Decision Tree
Panel Fixed Effects Model

A fixed effects model is a statistical model in which the model parameters are fixed or non-random quantities. This is in contrast to random effects models and mixed models in which all or some of the model parameters are considered as random variables. In many applications, a fixed effects model refers to a regression model in which the group means are fixed (non-random) as opposed to a random effects model in which the group means are a random sample from a population. In a fixed effects model each group mean is a group-specific fixed quantity. In panel data where longitudinal observations exist for the same subject, fixed effects represent the subject-specific means. In panel data analysis the term fixed effects estimator (also known as the within estimator) is used to refer to an estimator for the coefficients in the regression model including those fixed effects (one time-invariant intercept for each subject).


When Do Soccer Players Peak?

This is a review of the soccer research conducted by Seife Dendir, applying a Panel Fixed Effects Model.

Soccer teams are often wary when offering a contract to a player over the age of 30, as it is typically felt that the average soccer player peaks in their mid to late 20s.

In order to determine the peak age of professional soccer players, performance ratings of players from the top four European leagues, Bundesliga, Premier League, Series A, and La Liga, are used, focusing on the years from 2010-2015. A panel fixed effects model is utilized, as it incorporates the longitudinal variation in age and performance in order to determine the peak age. The three outfield positions of defense, midfield, and forward are analyzed via separate models.

Level of performance is determined from their WhoScored.com rating which ranks each player out of a possible 10 for every match they play based on each recorded play within the match. This rating is utilized as it will help smooth out variability in performances during the season due to unpredictable factors such as injury and luck.

The first step is to stack the data in panel form for each player across the years. From there, players are classified as a forward, midfielder, or defender. The sample included 1721 observations for forwards, 2779 for midfielders, and 3468 for defenders.

Before completing the estimation process, two other methods are used to determine peak age for comparison purposes, namely age distribution and bivariate analysis. The first method is a simple process of plotting the players’ age distribution and determining the mode. The idea is that most players of all abilities play at a professional level when they are at their peak level and therefore the modal age, which is the age at which most players compete, is the same as the peak age.

Two bivariate approaches are used, the first using playing time as a substitute for performance while the second one uses the WS rating. Playing time is used based on the idea that if coaches choose players...
simply based on performance, then the players who play the greatest number of minutes are the best performers. Therefore, the age when most minutes are played is the peak age. The second bivariate approach defines peak age as the one when average performance of the players is the greatest.

All three approaches indicate that the peak age for forwards and midfielders is approximately a year younger than that for defenders. However, while all the methods are informative and simple to use, they are easily affected by bias as they are impacted by other factors such as marginal players and assumptions that players are chosen based strictly on performance.

In modeling the relationship between age and performance and also estimating peak age, the issue of possible selection bias needs to be addressed. Selection bias is typically based on undefined differences between players indicating heterogeneity needs to be addressed. A longitudinal dimension of the data is used to combat this issue.

The results indicate that soccer players peak between the ages of 25 to 27, depending on which position they play, with forwards reaching peak age first, then midfielders and finally defenders.

This information is useful to coaches when negotiating contracts with current or prospective players. It provides a statistic on which to base decisions regarding how the player's age is likely to affect future performance. This is especially important in soccer where wages account for a large portion of a team's budget.

When looking at this information, it is of upmost importance to remember that a player's performance remains highly dependent on external factors such as his teammates, and is not simply an indication of individual skill.

Analytics methods used in this research: Panel Fixed Effects Model, Age Distribution, Bivariate Analysis, Estimation
Play Sketching

In play sketching, coaches can sketch plays and instantly see how their opponent is likely to respond.

Bhostgusters: Real-time Interactive Play Sketching with Synthesized NBA Defenses

This is a review of the play sketching research conducted by Thomas Seidl, Aditya Cherukumudi, Andrew Hartnett, Peter Carr and Patrick Lucey.

Tracking data in sports has become a staple for teams and coaches as they can now study their upcoming opponents in order to determine the best strategy possible. However, coaches do not have access to this data during the game itself. Instead, coaches typically rely on their experience and instincts.

Play sketching might be a solution. In play sketching, coaches can sketch plays and instantly see how their opponent is likely to respond.

Player tracking data from the 2016-2017 NBA season was used. Games were divided into possessions. Each possession had ten two-dimensional trajectories (one for each player, referred to as ‘ghosts’) and one three-dimensional trajectory for the ball. Information such as game clock, shot clock, player fouls, and the number of seconds each player has played in the game to this point was included. The data consisted of 30,764 possessions.

When tested against an actual game, the positions of the players on the court and the ghosts on the screen were not identical but the expected outcomes were very similar indicating that the ghost behavior was similar to the behavior of the players on the court.

Inferred tracking data makes it possible to visualize a sketched play as an animation. In this format, the user gets immediate feedback about the design and timing of a play. Additionally, the animation makes it easy to communicate the intention of the set play to those without the required experience to interpret a sketch directly.

Because sketches can be generated directly from tracking data, a user can edit any offensive play that was run in any game. For example, players can erase the actual pass that took place, and visualize how the defense would have reacted if the ball had been passed to a different teammate. Similarly, the routes of the teammates can be modified to better spread the defense.

Up to this point insights could only be gained from player tracking data after the game was finished. This framework allows coaches access to the data for use in in-game decisions by combining ghosting with a digital sketching interface. This framework is highly intuitive, allowing anyone to draw a play and easily understand how a team is likely to defend against it. It is also very quick in responding to questions.

Analysts, coaches and fans can all use this tool to explore an endless variety of scenarios, looking for weaknesses and strengths, and determining if they can find a better play for that scenario.
Analytics methods used in this research: ghosting, deep imitation learning, prediction error, expected values
Player Injury Forecasting

Player Injury Forecasting as a Method of Sports Analytics

As a result of increased popularity in the use of sports analytics, several methods of sports analytics have been discovered. This is all in an attempt to enhance team’s performance and improve on their chances of winning.

The Sport analytics method we’ll be considering today is Player Injury Forecasting.

Why Player Injury Forecasting?

Injuries suffered by players, either on the field or outside sporting activities has been identified as one of the major reasons for a team poor performance due to an inability of the player to function effectively. It is therefore important that professional teams be able to determine or quantify the likely injury-burden that would be encountered throughout a sporting season.

In an attempt to solve this, a player injury model was invented. This model helps to predict the likelihood that any given player will be injured during an upcoming game.

From a research conducted using SportsVU data model (a tracking technology that can collect positioning data of players during a game) and team information, like player workload and measurements, it was discovered that several factors can greatly increase the risk of player injury, they are:

- The total number of games played;
- The average running speed of a player during games;
- The average distance covered by a player;
- The average number of minutes played; and
- The average number of field goals attempted.
- The total number of games played, average number of field attempted and average number of minutes played were identified as workload. Increased workload is naturally connected with greater risk of injury.

Also, the average speed a player ran and average distance covered which is determined by playing style could cause an increased risk of injury.

The player injury model, SportsVU data, can also indicate the probability of a player being injured in upcoming matches. The knowledge of this can help coaches and teams decide when best to schedule their games and when to rest their star players, thus reducing the risk of long-term injuries.

For instance, when the model indicates that a player has a probability of 0.15 or higher percentage of being injured, that player can be advised to rest for the next game.
Analysis has shown that if the top 20% of high risk players were rested for a set time, it could be possible to prevent 60% of all injuries. An alternative approach to this could be to decrease the number of minutes a player plays on the field rather than resting them for an entire game.

In conclusion, team performance would be greatly enhanced if the risks of player injuries are reduced. The use of this method in sports analytics has also been proven to minimize financial costs needed to treat injured players.

In addition to this, a detailed player injury forecasting can help analysts and coaches make the best possible decisions on the field or court.

Apart from improving team sporting performance, player injury forecasting also promotes fan enjoyment and engagement, increase fan loyalty and can increase revenues from ticket pricing.
Poisson Factorization

Poisson factorization (PF) is a probabilistic model of users and items. It associates each user with a latent vector of preferences, each item with a latent vector of attributes, and constrains both sets of vectors to be sparse and non-negative. Each cell of the observed matrix is assumed drawn from a Poisson distribution, whose rate is a linear combination of the corresponding user and item attributes.


A Generative Model for Predicting Outcomes in College Basketball

This is a review of the research conducted by Francisco J. R. Ruiz and Fernando Perez-Cruz.

Predicting probabilities regarding outcomes of sporting events is difficult as it is often not clear which variables actually affect the outcome and what information is known before the event begins. Predicting outcomes for team sports is even more difficult as now there is the additional information regarding the individual players that will affect the outcome predictions.

A model is developed to estimate probabilities for the March Madness Tournament in college basketball. The model combines soccer models, which identify teams by its attack and defense coefficients, and Poisson factorization in which the parts of a matrix are assumed to be independent Poisson random variables given some hidden attributes. The structure of the March Madness Tournament is also taken into account.

To start, an attack and defense coefficient is defined for each team and for each NCAA conference. The conference coefficients replicate the overall behavior of each conference and the team coefficients demonstrate differences within each conference. Each coefficient represents a particular strategy, as some teams may be successful at defending some strategies while less successful at defending others. The same idea is used with attacking.

The posterior distribution of the attack and defense coefficients are approximated with a mean-field inference algorithm, as is the home coefficient. These inferences are converted into a non-convex optimization problem using variational algorithms which are simpler to compute than Markov Chain Monte Carlo methods and do not create the same limitations.

The variational algorithm is applied to four years’ worth of data from the tournament. The output of the algorithm is used to estimate the probability of teams winning in each tournament game. The predicted probabilities are averaged for 100 independent runs of the variational algorithm, implementing different initializations.

One advantage of this model is that it not only generates results, it also supplies explanations for the results. Experiment outcomes indicate that there is some advantage to playing at home, but that advantage is not as strong as it is for soccer teams. The model can also rank conferences. Ranking of teams indicates that teams that lose a few close games and win many games by a large margin will finish
higher than teams that win all their games by a small margin. Last, the model can predict the expected results of each game. These predictions were derived after averaging the expected Poisson means for the 100 independent runs of the model.

This model is general enough to be implemented for most team sports during regular season or playoff games. Consequently, analysts can rank teams by the strengths and weaknesses of the offense and defense. They can also compare different conferences within a league.

Analytics methods used in this research: Poisson Factorization, Mean-Field Inference Algorithm, Variational Algorithm
Populations and Samples

Populations and Samples in Empirical Research

It’s a great day in sports analytics! Today, we’re going to discuss the use of populations and samples, which are an integral part of probability and statistics. And we will examine populations and samples, as they are used in empirical research as a method of analyzing sports.

Establishing populations and samples through an empirical approach in sport analytics pave the way for sports analysts and coaches to represent the behavioral pattern of a team, or a sport in general, in a concise quantitative analysis.

This analysis helps analysts recognize the variability that appears in the team through measures of dispersion alongside measures of location. This measure is estimated by first describing the samples which are taking into consideration. Let’s see how samples are taken for probabilistic studies.

Before samples are selected in statistical analysis, population of category or topic under consideration is first drawn. Population refers to all the sets of a defined group that are collected for studying in order to make a statistical decision. For instance, a soccer analyst may decide to analyze the performance of players or teams in the league. Let’s take Premier League as a case study.

A soccer analyst who wants to study the Premier League will have to consider the twenty teams who play in the league, and then the total number of players in the league. In the case of the Premier League, the population of the statistics is 20. It is based on the performance of these twenty teams that a prediction or statistical decision is made.

A sample is a set of observations that are drawn from a population. Now, supposing the analyst wants to focus on the number of goals that are scored in the league each season, then the number of goals becomes the samples.

Empirical Research in Sports Analytics

Sports analysts employ empirical research to find deep information regarding the number of goals scored in one season of the Premier League—more closely, the number of players that score the most goals, the average number of goals scored by a team or a player, the relationship between goal scorers and the top teams, and so on. Empirical studies help analysts quantify all of these variables by using populations and samples, and then making measures of variability using various statistical techniques.

The process of using populations and samples for empirical research in sport analytics can be grouped into five areas. They are:

1. Abstraction: Abstraction represents the human behavior in numerical variables under the samples drawn from a population. This abstraction can help sport analysts quantify how often a player scores in a particular match relative to the strategy and strength of the opponents.
2. **Sampling**: Sampling is the collection of data, which represents a sample of the population of interest. An example of this is the number of goals scored per team in a season of the Premier League.

3. **Summarizing**: Summarizing determines the sample parameters, which describe the quantities of an average participant in the sample data. For instance, a soccer analyst may use summarizing techniques to determine the average number of goals scored by each team in a season.

4. **Analysis**: Analysis is where the statistical procedures developed in abstraction, sampling and summarizing are used to determine relationships between the variables used in the statistics, such as the relationship between goal scorers and the top teams.

5. **Generalization**: Having analyzed the statistical values, generalization is where the analysts make assumptions. For instance, a sports analyst may assume that the team that has the top scorers in the Premier League are the ones who end up in top four.

In conclusion, populations and samples in empirical studies are used in sport analytics to make certain deductions about a particular factor of a team or a set of players.
Prediction Model

Predictive modelling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect patterns, after the game has taken place. In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data.


Going Inside the Inner Game: Predicting the Emotions of Professional Tennis Players from Match Broadcasts

This is a review of the sports analytics prediction model research conducted by Stephanie Kovalchik and Machar Reid.

Tennis is often considered to be a mental game as it is an individual sport and the majority of the time is spent preparing for the next play. Consequently, coaching is often focused on the players’ mentality. The purpose of this study to develop a method to predict emotions in sports. It is believed that emotions can affect player performance, both positively and negatively. In order to determine the effect emotions have on a player’s performance it must be possible to measure emotion during competitions. Facial recognition programs aide in this process. This study focuses on seven emotions: anger, annoyance, anxiety, dejection, elation, focus and fired up.

Images were gathered from a variety of data sets. Faces of players were isolated and assigned an emotion. In total there were 7,952 images. Workers were found through Amazon Mechanical Turk who were asked to rate the intensity of each emotion from 0 to 10 with 10 being the most intense. Each image was rated by 5 people and the median (middle number) of the ratings was used as the final intensity. This information was then used to train the model to detect emotions and how to label its intensity.

This model was then used to analyze the facial emotions of Novak Djokovic, Roger Federer, Andy Murray and Rafael at the 2017 Australian Open including two matches per player. From this analysis profiles were created for each player. These profiles indicated that each player experienced unique emotions throughout a competition. Anxiety was the most common emotion among all the players and the most predominant emotion of Rafael Nadal. Roger Federer was the player whose emotion was typically either neutral or focused. Andy Murray and Novak Djokovic both experienced high level of anxiety. Djokovic experienced fired up, dejection and anger more often than the other players indicating that he demonstrated the most emotions during the competition among the four players. Elation and annoyance were rarely seen, with Andy Murray expressing these emotions the most. The fact that he experiences these two opposing emotions with a greater frequency corresponds with the reputation he has for being one the most volatile players on tour.
This data was used to test two commonly held beliefs. The first is that players’ emotions are a response to how they perform and the second is that a player's emotional reactions influence how they perform throughout the rest of the match. These beliefs were tested by combining each facial image with game context including the score, the winner of the point, the server of the point and the importance of the point. Three of the players showed strong emotional responses to their performance, the exception being Roger Federer. To test the belief that a player's emotion affects their play was tested by looking at the play following each facial image. Emotions affected the next play significantly for Nadal and Djokovic with much weaker links for Federer and Murray.

Coaches can use this information to help their players understand the relationships between their own emotions and how they play. If players were able to focus on 'letting go' of the emotions that hurt their performance while 'hanging on' to those emotions that improved their performance it is likely that they would improve their overall performance and chances of winning.

Analytics methods used in this research: data sets, prediction model, prediction method
Predictive Model Weights

Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights

This is a review of the predictive model weights research conducted by Joel Brooks, Matthew Kerr, and John Guttag.

Soccer is the world's most popular sport, however, its statistics are not as sophisticated as those used in other sports. Players are evaluated using simple stats like number of goals, shots and assists. A new method for player evaluation would be to measure their passing performance. Currently, players who are credited with assists are considered to be good passers. This needs to be expanded to include all passes, not only those that lead to goals. In order to do this sets of passes that lead to a shot on goal are grouped together. Then the location of where each pass originates and ends is taken into account as well as the player involved and the outcome. In order to incorporate pass locations the field is divided up into 18 zones.

After the data is inputted into the model that relates passing strategy in a possession to shots taken. Looking at the data shows that many shot opportunities originate in the zone centered in the field in front of the penalty box. The corner areas are also featured prominently in those possessions that lead to shots on goal.

The model is then able to rank players based on their ability to complete passes that lead to a shot. In soccer most possessions do not lead to shots on goal. Therefore, players will typically have a negative ranking as the majority of their possessions will not lead to shots on goal. Only a few outstanding players will manage to achieve a positive ranking.

Using predictive model weights facilitates analysis of key areas on the field that lead to shots on goal. If a team were to collect large amounts of data regarding their own plays and enter them into the model, they would be able to build a team specific model. This would allow an in-depth analysis regarding which strategies are the most effective for their situation and players. Teams then can write their playbook to feature these strategies more prominently. The team specific model would also outline which areas of the field lead to the greatest number of shots on goal for their players. Players could then be coached in how to utilize these areas of the field more effectively when they have the ball as well as when they are not directly involved with the ball. As very few players will achieve a positive ranking it will be obvious which players to focus on during a game and which players' skills should be analyzed in order to improve the skills of other players. Soccer is a game of strategy and this model would determine the best strategy for all players on the field to ensure the greatest number of shots.

Analytics methods used in this research: ranking, expected value, machine learning model, probability, weighting
Price of Anarchy

The Price of Anarchy (PoA) is a concept in economics and game theory that measures how the efficiency of a system degrades due to selfish behavior of its agents. It is a general notion that can be extended to diverse systems and notions of efficiency. For example, consider the system of transportation of a city and many agents trying to go from some initial location to a destination. Let efficiency in this case mean the average time for an agent to reach the destination. In the 'centralized' solution, a central authority can tell each agent which path to take in order to minimize the average travel time. In the 'decentralized' version, each agent chooses its own path. The Price of Anarchy measures the ratio between average travel time in the two cases.


The Price of Anarchy in Basketball

This is a review of the basketball research conducted by Brian Skinner, applying the price of anarchy concept.

A basketball game can be seen as a series of intertwined networks. Each possession has a starting point, a path to follow, and an ending point. Each possession, or pathway, would have its own unique probability of scoring points. A pathway that is used more regularly would naturally have a lower probability as the more often one type of play is used the easier it is for the opponent to be prepared. In order to describe the entire offense of a team in a basketball game as a network, every possible pathway would need to be created along with its unique efficiency, a process that is impossible in practical terms.

However, a network can provide useful information regarding a basketball offense, especially in looking at the difference between a team's efficiency and maximum possible points. In this case, a basketball game is compared to a simplified traffic network.

The difference between a team's efficiency and maximum possible points can be described as the price of anarchy, which measures how the efficiency of any system degrades due to selfish behavior.

In basketball, a possession is like a car on the road and the different plays are different roads that can be chosen to reach the destination. The more often a play, or road, is used the lower its efficiency. If all cars take the fastest route, the route will be clogged and time increases. However, if some cars take a slower route the average speed for all cars increases. This is the difference between selfish and unselfish behavior. Likewise, in basketball, always implementing the play with the highest probability of scoring can lead to an overall decrease in efficiency.

A simplified network can be created for a basketball offense with each line in the network connecting the beginning of a possession to the shot attempt. Each player is assigned a scoring efficiency dependent on how often that particular play is used. In this network the efficiency of a player is defined
as the player's true shooting percentage as a function of the fraction of the team's shots he takes while he is on the court.

The optimal strategy for a team would be the one that maximizes the team's overall efficiency. This optimal strategy can be calculated using Lagrange multipliers for constrained optimization.

Coaches and analysts could take this information to determine how often each player should take possession of the ball and how often a particular play should be utilized in order to maximize the overall team's efficiency. This may indicate that star players should not be given the ball as often as they currently are as such 'selfish' behavior may actually decrease the team's overall scoring potential. If a star player makes fewer plays, he will receive less defensive attention, leaving him more open to make the points when he does have the ball.

In fact, in a traffic network removing a road can actually increase efficiency, indicating that removing a player could also increase efficiency. This obviously is not logical and is referred to as Braess's Paradox.

Analytics methods used in this research: Price of Anarchy, Network, Braess's Paradox.
Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

Source - https://en.wikipedia.org/wiki/Principal_component_analysis

A Scalable Framework for NBA Player and Team Comparisons Using Player Tracking Data

This is a review of the NBA research conducted by Scott Bruce, applying principal component analysis.

Volumes of NBA statistics are available to the public through the player tracking camera systems. These cameras record the location of the players on the courts and the position of the ball. From this information, a wide variety of statistics can be calculated expanding the traditional stats available in the past. This allows the categorization of shot attempts and points by shot selection. Rebounding ability for contested and uncontested boards can be assessed and statistics such as average speed, distance, and opponent field goal percentage at the rim can be calculated.

However, the immense amount of data available makes it difficult to efficiently analyze the data and interpret findings. The dimensionality of the data set needs to be reduced while maintaining the ability to compare player performances. This can be done through Principal Component Analysis which separates out which dimensions of the data contain the maximum variance in the data set.

In running the analysis, four components stand out which account for 68% of the total variance. Each component has a vector containing the influence each variable has on the component, which makes interpretation possible. Variables can have either a negative or positive contribution which aids in determining what affect the variable actually exerts on the component.

The first components accounts for 42% of the total variance. The players with positive scores are skilled at securing rebounds and defending the rim and close shots. Players with negative scores tend to be outside players who drive the basketball to the hoop, take pull up, catch, and often shoot. They also tend to retain possession of the ball more often and generate assists.

Twelve percent of the variance is determined by the second component. Players with positive scores typically generate offence through assists and driving shots. They often have possession of the ball, either passing the ball to a teammate or driving the ball to the basket. A negative score is associated with those who mainly provide offense with catch and shoot shots, tending to be highly efficient scorers, especially from behind the 3-point arc.

The third component explains 9% of the variance with those having a positive score exhibiting quickness on both defense and offense, covering a lot of territory while playing. Those with negative scores are high scorers, providing a majority of the offensive production.
Finally, the fourth component explains 4% of the variance. Players with a positive score typically pass or convert catch and shoot shots when they receive the ball. A negative score is associated with players who drive the ball and score efficiently when they receive the ball.

These dimensions can also be used to profile teams using a weighted average.

Calculating the Statistical Diversity Index for pairwise player combinations provides a quick method for pinpointing players with similar statistical performances.

Analysts can use the team profiles to compare team playing styles and which style tends to lead to victories. The Statistical Diversity Index is useful in developing lists of players with similar profiles. Such a list can be used when looking at free agency or trades. It makes it possible to compare players with the playing style a team is looking for and perhaps replace a high paid player with a lower paid one who is still able to provide the necessary skills. It provides a method to help identify which players are best suited to fill holes within the team structure.

Overall, this method provides coaches and general managers with a tool to compare players and evaluate which are best suited for their team.

Analytics methods used in this research: Principal Component Analysis, Statistical Diversity Index
Probability

Probability is the measure of the likelihood that an event will occur. Probability quantifies as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur.

Source - https://en.wikipedia.org/wiki/Probability

Measuring Excitement in Sport

This is a review of the tennis research conducted by Graham Pollard, applying probability methods.

Excitement levels vary within every match of tennis observed by spectators. Excitement of a point within a tennis set can be defined and the mean and variance of the total excitement experienced in a tennis set can be determined, allowing comparisons between different sets.

The excitement fans demonstrate when a point is scored is not the same as the importance of the point; however, they are related. In fact, excitement is dependent of four factors. The first is at what stage of the match the point is made. Secondly, a game with many deuces raises the excitement level as points are scored. If the players rally for a longer than normal period of time the eventual point scored creates more excitement. Finally, excitement is influenced by which players are the playing the game.

This research focuses on the first two factors, namely stage of the game and number of deuces. Every point involved in a scoring system has its own level of excitement and that excitement is defined as the absolute value of the change in expected value of a player’s probability of winning the game as a result of the current point being played. Adding together the excitement for each point during the game calculates the total excitement of the match.

The importance of a point within a game is equivalent to the importance of the point within the set multiplied by the importance of the set within the game. The number of deuces is clearly demonstrated as a factor in excitement as enthusiasm of a tiebreak point exceeds the expected value of excitement for any other point. Using this information allows the average value and variance of the total excitement in a set to be determined.

It is important to realize that excitement in matches that can end in a tie cannot be measured on the same scale as excitement in matches that must produce a winner. This is not really an issue, as the two types of matches would never really need to be directly compared with each other.

Analysts could use this information to objectively compare games within a tournament, within a year or even throughout history in order to determine which game was the most exciting. This statistic could be added to summary statistics typically produced after a game of tennis.

While this paper focused on the game of tennis, the concept could be applied to other sports as well, in order to determine an objective analysis of the excitement of different games within different sports.
Analytics methods used in this research: Probability, Expected Value, Average Value, Variance

**Applying the Rules of Probability to Sports**

Sports statistics regularly use the theory of probability. It is important to have a good grasp of the concept in order to use the information appropriately and understand the meaning behind it. The starting point in sports probability is an experiment. An experiment is any course of action where the outcome is random. An experiment can be very general such a football game or more specific such as the quarterback throwing a pass to one of his receivers.

Once the type of experiment has been determined, the next step is looking at the specific events. Events are the outcomes of the experiment. In terms of the football game, the event could be the final score, the number of passing yards, or the number of passes attempted. In terms of the quarterback throwing a pass, the event could be whether it was a success or failure or the number of yards gained.

Each event has a corresponding probability. Probability is defined as a measure of how often a particular event will happen the experiment is done repeatedly. Probability tells us what will occur in a theoretical situation. If you toss a coin an infinite number of times, you will get heads 50% of the time. However, if you toss a coin only twice you are not guaranteed to get a head once and tails once. Therefore, a probability is not a guarantee of an outcome, but an indicator of what should happen if everything follows the same pattern as it has in the past.

At times, we might want to look at the probability that something will not happen. This is found by subtracting the probability that it will happen from 1.

The chance that an event will happen is most often expressed as probability. However, sometimes it is easier to use odds rather than probabilities. To find the odds of an event happening you take the probability of that event happening and divide it by 1 minus the probability of the event. When probabilities are very small or very large they become unwieldy to work with and difficult to interpret. In these instances, it is easier to work with odds rather than probabilities.

When odds are used within the world of sports, they are able to indicate the relative difficulty of an event. The greater the odds the harder it is for success to be achieved and consequently is less likely to occur.

Analysts use probabilities and odds to make predictions regarding outcomes of games and the performance of various players during the game. Coaches use probability as a tool to determine what areas their team needs to work on in order to increase the probability of success.
Probabilistic Graphical Model

A graphical model or probabilistic graphical model (PGM) or structured probabilistic model is a probabilistic model for which a graph expresses the conditional dependence structure between random variables. They are commonly used in probability theory, statistics—particularly Bayesian statistics—and machine learning. Generally, probabilistic graphical models use a graph-based representation as the foundation for encoding a distribution over a multi-dimensional space and a graph that is a compact or factorized representation of a set of independences that hold in the specific distribution. Two branches of graphical representations of distributions are commonly used, namely, Bayesian networks and Markov random fields.


Graphical Model for Basketball Match Simulation

This is a review of the basketball research conducted by Min-hwan Oh, Suraj Keshri, and Garud Iyengar, applying probabilistic graphical models.

With any sporting event, it is natural for analysts, bettors, and fans to make predictions regarding the outcome. This is certainly true within the National Basketball Association.

A simulation infrastructure is developed to bring together player identity and team level network. A basketball game is modelled using a probabilistic graphical model, which illustrates every touch and event during a game as a sequence of transitions between discrete states. The progression of a game is treated as a graph in which each node is a network structure of players, the actions, and events while the edges illustrate possible moves in the flow of the game. The conditional probability of the edges is learned while ball movements between players, how likely a player is to take shot, and how defense and teammates affect the dynamics of the game are all simulated.

The start of a possession is modelled as a multinomial distribution between players on the court. The probability of a field goal attempt is modelled as a Bernoulli distribution based on the idea that the likelihood of a player taking a shot depends on his tendency to shoot, his defender, and also the tendencies of his teammates. Shot efficiency is modelled as a function of the offensive player’s skill, the defender at the time of the shot, and the location of the shot. Passes between players are modelled as a network. Shooting fouls are modelled as a function of the shooter’s skill at drawing a foul, the defender’s foul tendencies, and the location of the shot. A free throw percentage for each player is used to sample a free throw success event. Rebounds are modelled as a competition between the players on the court. Possession start begins with an inbound pass, a defensive rebound, or a steal. Two types of turnovers are considered, stolen balls and all other turnovers that result in an inbound pass. The average probability of turnover per touch for each player is calculated from historical data.

The model is used to simulate the 2013-14 season record for each NBA team resulting in good estimates of the teams’ actual win percentages. For the simulation, the lineup is the input parameter. Matches are simulated multiple times to estimate the expected statistics for both players and teams. Results indicate
that changes in a team’s lineup or the opponent's lineup has a significant impact on the dynamics of
how the game progresses.

The model can be used to evaluate the effect different lineups have on game results. Coaches can
evaluate hypothetical lineups against an opponent in order to determine which lineup will be the most
effective. Coaches can also model hypothetical opponent lineups in order to work out the best defensive
strategy. Specific player performance can also be evaluated. This will help when looking at making
trades and determining which players would bring the best improvement to the current lineup.

Analytics methods used in this research: Probabilistic Graphical Model, Multinomial Distribution,
Bernoulli Distribution
Probabilistic Physics-Based Model

The purpose of the probabilistic physics-based model is to determine the probability that the team in possession of the ball will score with their next on-ball event.


Beyond Expected Goals

This is a review of the probabilistic physics-based model based on player tracking data research conducted by William Spearman.

Soccer is more a game of strategy than scoring as there are relatively few goals made in a soccer game. What are the other players on the field doing when they do not have the ball and what value do they contribute?

The development of tracking data allows us to look into these questions. A probabilistic physics-based model is built to determine the probability that a player who does not currently have control of the ball will score a goal. The data comes from a 14-team professional soccer league during the 2017-2018 season which included 58 matches. Events were determined as any on-ball actions that occurred during a match. Each event was labeled with the time the event occurred, the player in possession of the ball, and the type of event (pass, shot, goal, etc.). The position of all players on the field were included in the data.

The purpose of the probabilistic physics-based model is to determine the probability that the team in possession of the ball will score with their next on-ball event. In order to do this the probability that the team successfully passes to each point on the field and scores is determined. Three individual probabilities are taken into account - the probability that the ball is passed to one point on the field, the probability that the ball will be controlled at that point by the passing team, and finally the probability of scoring a point from that point on the field. These three probabilities are then combined to represent the probability that a goal will be scored with the next on-ball event at one specific spot on the field.

There are several possible applications of this data for analysts and coaches. Analysts will be able to quickly determine which were the key moments in the game - which moves ultimately lead to the score or offensive chance.

Are players passing to the points on the field where their teammates have the best chance of scoring? If not, then strategies need to be reviewed in order to maximize scoring chances. Coaches can look at how their opponents use the space on the field in order to develop the best possible defensive strategy. Coaches can examine the data to determine player and team preferences. This can be used to strengthen their own team by adapting offensive strategies to maximize these preferences. Defensive strategies can also be developed to minimize the chances of the opponent scoring.
After the data was analyzed for all matches it was found that teams who score more goals also generated more opportunities to score. Coaches and players need to be aware of where the best opportunities to score exist in order to maximize their chances of scoring.

When teams are looking to make trades they can analyze the data of the players they are looking at to determine which one will be the best fit for how the team uses space on the field.

Analytics methods used in this research: probability, parameters, off-ball scoring opportunity, probabilistic physics-based model
Probability Data Aggregation

Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis. A common aggregation purpose in sports analytics is to get more information about players, teams and their actions on the field, court or ice.

Source - https://searchsqlserver.techtarget.com/definition/data-aggregation

Preventing In-Game Injuries for NBA Players

This is a review of the player injury research conducted by Hisham Talukder, Thomas Vincent, Geoff Foster, Camden Hu, Juan Huerta, Aparna Kumar, Mark Malazarte, Diego Saldana, and Shawn Simpson

Player injuries significantly affect the overall performance of a team and therefore are very concerning for team management and fans. A player injury model has been developed to help teams predict the likelihood that any given player will be injured during an upcoming game. This is done through a quantitative and systematic approach.

Data for the analysis was gathered using play-by-play game data, SportsVU data, player workload and measurements as well team schedules for two years. Testing the model using this data proved that it is able to predict the probability of a player being injured in the upcoming week.

The research demonstrated that the most relevant factors that increase the risk of injury were (in decreasing order):

(1) the average speed at which a player ran during games;
(2) the total number of games played;
(3) the average distance covered by a player;
(4) the average number of minutes played; and
(5) the average number of field goals attempted.

Total number of games played, average number of minutes played and average number of field goals attempted are all related to workload. Increased workload would naturally be associated with a greater risk of injury. The average speed a player ran and average distance covered deal with playing style which is also related to an increased risk of injury. However, the number of back-to-back games and the number of games played during a 14-day period did not significantly increase the risk of player injury. This contrary to the thought process of many, including NBA officials. Acting on this data could change the direction the NBA has been taking to schedule fewer back to back games in order to minimize injuries.
Combining these results with team schedules would allow team management to identify when would be the best time for a team to rest their star players and reduce the risk of long-term injuries. If the top 20% of high risk players were rested it could be possible to prevent 60% of all injuries.

Rather than resting players randomly during the year, it could be done strategically using this model. When the model indicates that a player has probability of 0.15 or higher of being injured that player should be rested for the next game. Resting for one game dramatically reduces the risk of injury. The team could take into account both the probability of injury and the importance of the upcoming game in making decisions regarding resting players. It is not always possible or feasible to rest a player for an entire game. An alternative approach would be to decrease the number of minutes a player plays over several games rather than resting them for an entire game.

In conclusion, reducing the risk of player injury would enhance team performance and fan enjoyment as well as minimize the financial cost associated with injured players missing games.

Analytics methods used in this research: machine learning, probability, data aggregation, variables, optimization
Probability Distributions

In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events. For instance, if the random variable $X$ is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of $X$ would take the value 0.5 for $X = \text{heads}$, and 0.5 for $X = \text{tails}$ (assuming the coin is fair). Examples of random phenomena can include the results of an experiment or survey.

Source - https://en.wikipedia.org/wiki/Probability_distribution

Point Distributions and Expected Points

One probability distribution that is especially useful in sports statistics is the distribution of the points scored. This distribution analyzes the probability of scoring points in any given situation. In a football game, a team could be on second down at their opponent's ten-yard line. The distribution would give the probability of the team scoring 6 points, 3 points, or no points in that situation.

The distribution of points is very detailed, which is not always required. At these times, the expected value, also known as expected points, is a more useful tool. Rather than a distribution of points, the expected points boils the chances of scoring points in a particular scenario down to one number. For example, a baseball team has a runner on third base, with one out. The expected runs in this scenario would be seen as a single probability such as 0.415. Expected points are a single number summary of a point distribution. It can be useful on its own but you must be careful, as it does not contain all of the information often required in an analysis. Relating it back to the distribution can prove to be very eye opening in looking at the expected points of scenarios that are very similar to each other.

There are two basic methods used to determine the probability distribution of points. The first approach sifts through all of the historical data available combining similar scenarios together to determine the probability distribution. This approach assumes the current scenario will follow a pattern similar to previous scenarios. The second approach is to use a probability model based on the probability of certain outcomes occurring and creates the probability distribution from that information. The type of approach used depends on the type of game being analyzed. Neither method is able to guarantee the expected points in a given situation, as it is unable to take into account the intangibles of the individual players on that specific day.

Analysts can use the probability distribution to look at the probability of the impact different plays can have on the outcome of the game. This helps in ranking which type of play in a given scenario is most likely to gain the outcome desired. They can use expected points to analyze a coach’s choice in which play to call and how effective it will be in that situation or whether a different play would be more likely to have a positive outcome.
Coaches can use the probability distribution to evaluate strategies in different situations giving them information to adjust their playbook to better fit specific scenarios when facing different opponents. Looking at expected points helps them decide which play is most likely to result in a score.
Probability Regression Model

In statistics, a linear probability model is a special case of a binomial regression model. Here the dependent variable for each observation takes values which are either 0 or 1. The probability of observing a 0 or 1 in any one case is treated as depending on one or more explanatory variables. For the "linear probability model", this relationship is a particularly simple one, and allows the model to be fitted by simple linear regression.

Source - https://en.wikipedia.org/wiki/Linear_probability_model

Spatial Analysis of Shots in MLS: A Model for Expected Goals and Fractal Dimensionality

This is a review of the MLS research conducted by Alexander Fairchild, Konstantinos Pelechrinis, and Marios Kokkodis, applying a probability regression model.

A simple way of analyzing the performance of a soccer team is to look at the number of shots taken. However, this can be highly inaccurate as not all shots have the same probability of scoring. To rectify this, the idea of expected goals has been utilized.

To build the model, data was collected by watching and manually noting the coordinates of non-penalty shots that were on target from 99 MLS games during the 2016 regular season. Additional information came from a pitch tracker including shot's location, outcome, game state, assist type, and phase of play that produced the shot. Analyzing the data determines that there are no statistically significant differences in goal probability when focusing on various aspects of a feature. Different phases of play produce the same baseline probability as does shot angle.

A goal probability regression model is built to determine the probability of a shot ending up as a goal. The information obtained above was input into the model, which determined that only the distance from the goal and angle of the shot had a statistically significant impact on the probability of the shot leading to a goal.

Using this model the expected number of goals for a team or player can be estimated either for a game or across several games. A sequence of shots is modelled as a Poisson binomial distribution, which is a sum of independent Bernoulli trials.

Using the expected goals calculated by the Poisson distribution the offensive and defensive efficiency can be determined based on how much better or worse they perform compared to the expected goal count. Teams are plotted based on their offensive and defensive outcomes, which provides an easy classification system. Teams in the top right hand corner are over achieving in both offense and defense. The top left hand quadrant contains teams with strong defenses, bottom right have strong offenses, and those in the bottom left quadrant are underperforming in both areas. Players' efficiencies can also be plotted, again providing an easy classification system for each player.

The next step is to determine if there is a relationship between the spatial distribution of shots and the offensive efficiency of a team, which is achieved by fractal dimension. If a small value is calculated this
indicates the team's shots originate from a small area of the field. A large value indicates the team's shots are more uniformly spread across the field.

Teams are divided into two groups, the top 50th percentile, and the bottom 50th percentile. Teams in the bottom half exhibit a positive offensive efficiency while teams in the top percentiles have a negative offensive efficiency. These results indicate that highly efficient offensive teams focus their shots from a small area of the field. Teams who spread their shots out from various areas of the field end up taking many lower probability shots, thus achieving fewer goals.

This information provides coaches with a focus for training, specifically focusing offense on utilizing a smaller portion of the field for shooting on net in order to increase scoring probability. Players' movements can be analyzed through this process, allowing coaches to work on improving the skill set of individual players. Teams with lower efficiency ratings can analyze teams with positive ratings to determine what successful strategies other teams are utilizing.

Statistics continue to grow and while looking at the number of shots provides information, comparing offensive efficiency with spatial distribution provides practical information.

Analytics methods used in this research: Probability Regression Model, Poisson Binary Distribution, Shot Charts
Probit Regression Model

In statistics, a probit model is a type of regression where the dependent variable can take only two values, for example a made shot or missed shot. The word is a portmanteau, coming from probability + unit. The purpose of the model is to estimate the probability that an observation with particular characteristics will fall into a specific one of the categories; moreover, classifying observations based on their predicted probabilities is a type of binary classification model.

Source - [https://en.wikipedia.org/wiki/Probit_model](https://en.wikipedia.org/wiki/Probit_model)

The Advantage of Experience: Analyzing the Effects of Player Experience on the Performances of March Madness Teams

This is a review of the NCAA basketball research conducted by N. David Pifer, Timothy D. DeSchriver, Thomas A. Baker III, and James J. Zhang, applying a probit regression Model and least squares regression model.

Every March the men's Division I teams from the NCAA compete in the NCAA Men's Basketball Championship, which a six-stage, single elimination tournament. The pressure on teams and individual players is immense as there is a great deal at stake including funding and possible future careers. There are also the internal pressures of wanting to win and a strong desire to demonstrate your best skills to a large national audience. The standard belief is that the experienced player, whether in terms of playoff experience or in terms of age, will rise to the occasion while inexperienced players will falter. This assumption has not been tested so this study looks at a statistical analysis to analyze the statement.

This study used two different measures of experience. The first was prior experience in March Madness games determined by the number of minutes a team member had played in previous March Madness competitions. The second measure was the age of the player, determined by their class ranking. Other factors included were the win-loss percentage, strength of schedule, offensive and defensive ratings, coaching experience, and player height.

To test the impact experience has on winning a game, a probit regression model was developed. In the early stages of the tournament, the strength of schedule had the greatest impact on a team's probability of winning with win-loss percentage, offensive and defensive ratings having a significant impact as well. Prior March Madness experience did not significantly influence the probability of winning in the early or late rounds. In fact, results show that experience, or age, actually has a negative effect in the later rounds of the tournament. Actually, a strong defense had the greatest effect on winning in the later rounds of the tournament.

An ordinary least squares regression was developed to look at the impact experience has on the margin of victory. Again, win-loss percentage, strength of schedule, offensive, and defensive ratings were the
only variables having a significant impact on the margin of victory. Greater experience has a slight positive impact, which could prove helpful, especially in closely contested games in the later rounds.

Overall, the assumption that age and experience increase a team’s probability of winning the tournament appears to be false.

Analysts can use this information to re-evaluate their predictions regarding who will ultimately be the winner of March Madness. Teams can use this information when forming their teams, taking a closer look at the younger players to determine if they have the ability to increase their probability of winning. Obviously, while some experience is definitely necessary it is important to ensure that a team does not consist entirely of seniors, but includes a strong selection of younger players as well.

This study demonstrates the importance of testing what appears to be the most logical of assumptions to ensure their accuracy.

Analytics methods used in this research: Probit Regression Model, Least Squares Regression Model
Probit Stepwise Regression

In statistics, stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Usually, this takes the form of a sequence of F-tests or t-tests, but other techniques are possible, such as adjusted R2, Akaike information criterion, Bayesian information criterion, Mallows's Cp, PRESS, or false discovery rate.


An Easily Implemented and Accurate Model for Predicting NCAA Tournament At-Large Bids

This is a review of the NCAA tournament bid research conducted by B. Jay Coleman, J. Michael DuMond, and Allen K. Lynch applying probit stepwise regression analysis.

The NCAA Tournament garners more than 60 million Americans completing a tournament bracket. It is estimated that anywhere from $60 - 70 million is bet legally every year with another $3 billion wagered through pools. ESPN held a tournament projection contest in 2014, which gained more than 11 million participants. Obviously, the NCAA Tournament is the focus of attention from the time the playoff bracket is announced until the final championship game.

During the time prior to the announcement of which 68 teams will be competing players, coaches, directors, media and fans devote a lot of time and effort into attempting to determine which teams will be invited to participate. Thirty-two teams are invited based on their winning record while the NCAA Selection Committee, which consists of ten people, namely athletic directors and conference commissioners, determines the remaining teams. They not only determine which teams will be invited but they also seed each team and place them in the playoff bracket.

Gaining an invitation to the tournament is of major importance to the teams. Many view being invited to the tournament as the indication that they have had a successful season. In addition, there are many benefits to be gained for schools, players, and coaches. Each school receives $1.58 million, paid out over six years, for each tournament game their team plays in.

The Selection Committee holds closed session when determining their selections and do not typically reveal why certain teams are chosen while others are not. It is not known which factors the committee considers to be of highest importance in their decision making each year. While the factors they focus on are not known, there is a list of all the factors that the NCAA provides to the committee for their consideration.

This research is designed to improve on earlier processes used to predict which teams which will be selected. It is based on the Rating Percentage Index which is the weighted average of: the team's winning percentage versus Division I teams, the team's opponent's winning percentage, and the winning percentage of the team's opponents' opponents. The RPI is determined for each team using a formula
that weights all wins and losses equally. The outcome is measured based on a binary variable of whether or not a team was invited to play at the tournament.

The first step involved performing a probit stepwise regression. These results were used as a basis for further analysis. The probit stepwise regression analysis selected the following predictors: RPI ranking, in-conference losses below 0.500, wins against the RPI top 25 teams, wins against the RPI teams ranked 26-50, games above 0.500 against teams ranked 51-100, as well as the number of road wins.

Next, two additional model selection processes were used to sort through the data to create alternative factors for comparison. Neither alternatives were better predictors, and therefore the original method was retained.

The obvious use of this statistic is that everyone involved will gain a greater ability to determine who is likely to be selected to join the tournament. Coaches can decide which factors they need to focus on in order to increase their chances of being chosen.

Finally, the fact that this model is able to accurately duplicate the decision of the Selection Committee over the past years is a clear indication that the process employed by the committee is very consistent.

Analytics methods used in this research: Rating Perception Index, Stepwise Probit
Production Function

In economics, a production function gives the technological relation between quantities of physical inputs and quantities of output of goods. The production function is one of the key concepts of mainstream neoclassical theories, used to define marginal product and to distinguish allocative efficiency, a key focus of economics. One important purpose of the production function is to address allocative efficiency in the use of factor inputs in production and the resulting distribution of income to those factors, while abstracting away from the technological problems of achieving technical efficiency.

Source - https://en.wikipedia.org/wiki/Production_function

Diamonds on the Line: Profits through Investment Gaming

This is a review of the research conducted by Clayton Graham.

As long as humans have been competing against each other, others have wagered regarding the outcome. Sports and betting go hand in hand across time, sports, and societies. In order to profit in the baseball marketplace, baseball game modeling and analytically based gambling are amalgamated. Investment gaming's objective is to maximize the expected profits of a baseball season, following investor risk guidelines.

The ultimate goal in baseball is to score runs. In order to score runs a batter must first reach base. Therefore, the production function is built using singles, doubles, triples, home runs, and base on balls as inputs. The output is the number of runs per out. Runs per out are incorporated rather simply the number runs scored as this effectively cancels out any effects caused by a varying number of innings between games. A formula for expected winning percentage is created as a combination of density functions related to runs scored and runs allowed in a Pythagorean Theorem. The formula is run through a Monte Carlo simulation to create an accurate representation of the distribution of winning percentages and winning margins. The results then need to be adjusted for batter pitcher matchup and ballpark factor. The ballpark factor measures the difference between runs scored in a team's home park and road games.

The betting line determines the cost of the bet, the resulting payoff, and an implied probability of winning. The most common form of betting on baseball is money line bets. Bringing in the idea of economic consequences involves including the money line's cost, payouts, implied probabilities of winning along with the production function's expected scoring, and the game's predicted probability of winning.

Several conclusions are drawn after running the model. Only a small percentage of games are worth an investment. The overall winning percentage is typically 68% and the average return on capital is approximately 35%.

The methods discussed here can be used by coaches and analysts to determine the value of various batters and pitchers. They can also help identify strengths and weaknesses in various teams in the
league, which provides tools for developing strategies to improve one's team and gain a competitive advantage against opponents. Finally, the methods included here would aide coaches in deciding which players should be included in a game's lineup depending on the opponent and the ballpark where the game is being played.

Analytics methods used in this research: Production Function, Pythagorean Theorem, Monte Carlo Simulation
Pythagorean Formula

This theorem states that, for independent random samples, the square of the standard deviation is their sum is the sum of the squares of their standard deviations. Variances: It is easier to write the Pythagorean Theorem of Statistics using variances: \( \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) \)


The Shrinkage of the Pythagorean Exponents

This is a review of the NBA research conducted by John Chen and Tengfei Li, applying the Pythagorean formula.

The Pythagorean exponent was a major step forward in the ability to predict outcomes of sporting events. The Pythagorean formula tells us that the ratio of a team's total points scored to the total points allowed is a more accurate statistic than the ratio of total games won to total games lost. The expected winning percentage determined by the Pythagorean formula is the percentage of games that a team should win. When a team has more wins than the expected number of wins they are overachieving and it is expected that their future performance will revert to the expected number of wins.

The problem with the formula is determining the appropriate exponent to use, which varies between sports. In fact, it can even vary between different leagues of the same sport.

In order for the Pythagorean formula to be utilized in the NBA, the optimal exponent needs to be determined. This is done by analyzing data from 21 NBA seasons, from 1993 to 2014. This data provides the statistics needed to determine team strengths through the beginning of a season. These are then compared against the future win-loss ratio of the remaining games in the season in order to determine the optimal exponent for predicting future winning percentages within the NBA.

Team strength is determined using point ratio, point difference, ratio of offensive/defensive ratings, and win-loss ratio. Logarithms are used to place the average team at a score of zero, making comparisons easier. An ordinary least squares approach is then applied to the information to determine the Pythagorean exponents.

Analysis of the outcomes states that the Pythagorean exponent needs to be shrunk in order to obtain a more accurate prediction of the winning percentage for the games remaining in the season. The shrinkage factor is defined as the value of the Pythagorean exponent at a time in the season in the prediction scenario divided by the value of the Pythagorean exponent at the end of the season in the fit scenario. At the beginning of the season, the shrinkage was significant and gradually decreased later in the year. Finally, predictions are less accurate towards the end of the season. The law of large numbers explains these findings, which states that the average of the results obtained from a large number of trials will be close to the expected value, and will tend to become closer as more trials are performed.

Analysts can use this information to predict outcomes of future games, leading to predictions regarding expected outcomes of the season such as which teams will make it to the playoffs and which will
eventually win the championship. Teams that are winning more games than expected can be analyzed to determine what factors may be behind the greater than expected strengths. Coaches may then be able to use those factors to adjust their training regimes in order to improve their own team’s strength.

Analytics methods used in this research: Pythagorean Formula, Pythagorean Exponents, Shrinkage
Pythagorean Theorem

An Improvement to the Baseball Statistic "Pythagorean Wins"

This is a review of the baseball Pythagorean Theorem research conducted by Jay Heumann.

In baseball, the general belief is that a team's ratio of runs scored to runs allowed is actually a better predictor of a team's future performance than their winning record. This is based on the baseball Pythagorean Theorem, which uses the number of runs a team has scored and the number of runs a team has allowed to calculate the probability that a team will win a future game.

However, many feel that the formula can be improved in order to calculate a more accurate probability of a team winning. One suggestion is changing the exponent from 2 to 1.83. This research demonstrates a new way to improve the Pythagorean wins statistic, called a pairwise Pythagorean formula. In this formula, the Pythagorean wins are calculated on a team-by-team basis, and not cumulatively for all teams. While the original formula calculated the sum over all teams first, this method determines each teams' Pythagorean win total against other individual teams and then calculates the sum over all teams.

In order to test the validity of the pairwise Pythagorean formula it was put to the test using actual data. Data from the 1960-1990 Major League seasons were used and both the traditional Pythagorean Theorem and the pairwise Pythagorean Theorem were calculated for each year along with their root mean square error. Both formulas were calculated using the exponent of 2 and 1.83. The resulting chart clearly indicates that while changing the exponent from 2 to 1.83 in original formula is more accurate, the pairwise formula is even more accurate whether using 2 or 1.83 as an exponent. In 36 of the 60 seasons, the error of the two pairwise formulas was lower than that of both of the traditional formula. In 22 seasons, both pairwise methods were more accurate than their traditional counterparts were.

To improve the formula even further it is suggested that other variations should be tested to examine their validity. Variations could include changing the exponent to something other than 2 or 1.83. It is also possible that the formula should not use fixed exponents, but rather exponents that are a function of some other variable and would therefore vary from team to team.

Analysts and coaches can use this new pairwise Pythagorean formula to gain a more accurate reading of the number of games their team is expected to win over the course of a season. Analysts would have greater accuracy in predicting which teams will make it into the playoffs. Coaches will have the opportunity to make trades that will better improve their chances of winning.

There are many possible variations to the Pythagorean Theorem. It is conceivable that one of them might produce more accurate results than the tradition theorem currently being used. Further exploration is needed to determine which variation will ultimately create the best prediction of a team's expected win-loss totals.

Analytics methods used in this research: Pythagorean Theorem, Root Mean Square Error
Pythagorean Win Expectancy Model

Decomposing Pythagoras
This is a review of the baseball research conducted by Edward H. Kaplan and Candler Rich, applying the Pythagorean Win Expectancy Model.

Bill James developed the Pythagorean win expectancy model as a way to convert total runs for and against a baseball team as an estimate of that team's seasonal winning percentage. Many others have studied and analyzed the model since and adapted it for use in other sports, including basketball, football, and hockey. It has also been adapted for use in overtime in various sports. The exponent varies from sport to sport but there is no explanation for why this occurs. This research looks for an explanation to this problem.

The Pythagorean model relates a team's seasonal win percentage to that team's total runs scored and runs against. The original model used an exponent of two, but since then experts have determined that 1.8 is a better exponent when applying the model to the sport of baseball. A general form of the equation uses a variable as the exponent, allowing the formula to be applied to other sports where instead of total runs scored and runs against it now related to total points scored and points against.

Running a simple linear regression of win percentage versus the difference between runs for and against teams, results in slope estimates very close to the value of the exponent in the theorem, indicating that the theorem accurately represents the relationship between winning and scoring.

Using an exact win expectancy model and decomposing the Pythagoras model, it becomes evident that the exponent increases with the average number of points scored per game. It also increases along with the empirical average of the exact model slopes from all teams, which declines with the scoring margin. In other words, better teams have higher average net scores, larger margins of victory, and smaller margins of defeat.

The conclusion reached is that it depends on the typical scoring margins of victory and defeat in a game, as well as the average number of points scored in total. Both of these vary across sports, which is why the Pythagorean exponent also varies across sports.

Testing this idea with other sports tells us that the Pythagorean model accurately relates win percentage to both scoring total and margins of victory for the sports of baseball, basketball, and hockey, but not as well with the game of football.

The ratio of the Pythagorean exponents for two different sports is approximately equal to the ratio of average points over average winning margin for the first sport divided by the average points to winning margin ratio for the second sport. The different exponents for different sports provide information regarding how scoring margins and total points combine together to produce winning records.

Analysts can use this information as way to compare very different sports and look at how scoring margins and total points affect each sport's winning probability. Knowing what the exponent represents should allow analysts to apply the Pythagorean Theorem to other sports and to compare those sports to each other.
Analytics methods used in this research: Pythagorean Win Expectancy Model, Wimple Linear Regression, Exact Win Expectancy Model
Qualitative Exploration

Probably the oldest of all scientific techniques, qualitative exploration is a method used extensively by scientists, researchers and analysts to study human behavior, opinions, themes and motivations. Different from quantitative data, which relates to quantities, amounts and measurements that can be expressed in numbers and manipulated statistically, qualitative data is concerned with the features, attributes and characteristics of phenomenon that can be interpreted thematically.

Source - https://explorable.com/qualitative-research-design

Social Talent Scouting: A New Opportunity for the Identification of Football Players?

This is a review of the social talent scouting research conducted by Elena Radicchi and Michele Mozzachiodi, applying qualitative exploration.

Teams are continuously on the lookout for new talent. A variety of tools are used to facilitate this process. First is human expertise - scouts have built up years of experience watching players at all levels, looking for those who demonstrate the skill necessary to become a professional. Statistical data is prolific, covering every imaginable quantifiable aspect of the game. New technologies such as play tracking software are now available to provide an additional layer of knowledge, allowing increased efficiency in forecasting of a player's future. Social media can now be added to the list as a possible tool for the scouting process.

The goal of scouts and teams is to identify talented athletes before anyone else does. Social media is one way that more young players can gain exposure in order to be evaluated. Scouts have been able to start a dialogue with talented youngsters by using social media such as Twitter or Facebook. They can also monitor the social media accounts of these young players, giving them unique insight into their lives. This can be useful in monitoring behavior and personality and whether these increase their potential. Text mining could be used to extract data that can be analyzed regarding the attitudes of these young players.

There are a multitude of web tools available, one being the FB Player - Social Football Talent. This platform was developed to connect players and scouts/coaches. It provides a spot for the young players to display their skills for a wider audience. The FB Player database was analyzed, focusing on those who were registered as players. The majority of these players were 18 years of age or less. Players came from many different countries, including countries that are extensively scouted as they tend to have a greater proportion of talented players compared to other countries. This would indicate that even in these countries there are still young players who are not in a position to be seen by the scouts so they are looking to alternative methods to be noticed.

The players were typically motivated by the desire to have their skills noticed, a passion for the game, a need for a sense of belonging, or a desire to learn new skills.
Social media provides scouts with the opportunity to start the scouting process sooner and to do it more efficiently. Young talent can be followed through social media and other online tools, decreasing the need for travel.

While social media will never replace the traditional scouting process, it can be used as an additional tool to improve the process and give scouts access to a larger pool of young talent.
Random Coefficients Model

Random coefficient models are intended for settings with two or more sources of random variation. The widest range of applications is found for them when observational units form natural clusters, such that the units within a cluster are more similar than units in general. Models for independent observations have to be extended to allow for within- and between-cluster variation.


Effort vs. Concentration: The Asymmetric Impact of Pressure on NBA Performance

This is a review of the NBA research conducted by Matt Goldman and Justin M. Rao, applying a random coefficients model.

Athletes are often praised or criticized for their performance under pressure. They are said to rise to the occasion or buckle under pressure. Studies have shown that amateur athletes do tend to perform worse when playing in pressure situations. The drop in performance is often caused by a tendency to self-focus, or focus on each individual motion in an action rather just letting the body complete the action naturally or automatically. The question in this study is how elite athletes react to pressure during game situations.

In order to accurately study the questions four criteria must be met: a well-defined action with a clear positive or negative outcome, a precise measurement of the importance or pressure of the situation, a large sample size of action in both high and low pressure situations, and actions made up of several different motor tasks. To meet these requirements the study looks at free throws and offensive rebounds in the NBA. In order to determine a precise measurement of the importance or pressure of the situation a model is created to estimate the value of a point scored at each moment in the game. The value is defined as the impact the point has on the probability the team will win the game.

The model determines the win probability of every regular season game from the 2005-2010 seasons during which time 360,000 free throws were attempted. Determining the probability a free throw will be made is done with a random coefficients model, which allows parameter values to fluctuate between players. The key points analyzed are if the shot is taken at home, the win value of the point, the point importance combined with the home dummy, which applies the differential impact of home pressure and the time remaining in the game.

The outcomes show that average NBA players perform slightly worse in moderate pressure level situations at home. However, increased pressure levels do not affect the visiting team in those same situations.

The home dummy coefficient provides an estimate of how home players respond in non-pressure situations, which is both positive and significant. In non-pressure situations, home team players are more accurate in their shooting ability than those playing for the visiting team. However, as the pressure increases visiting players fare better than the home team players. Overall, the performance of the better
free-throw shooters is less likely to be affected by pressure, whether playing for the home or visiting team.

Offensive rebounding is looked at next as it takes place during the course of play, rather than when the game is stopped, as it is for free throws. Therefore, there is not the time to over think the process. Looking at offensive rebounding shows that it does not vary with increased pressure for either the home or visiting team. As points become more important later in the game, the home team gains a significant advantage over the visiting team in this area.

The results clearly indicate that pressure while playing at home does affect a player’s ability to make free throws. This would indicate that players would benefit from working with a sports psychologist to help become better able to handle the pressure and not let it affect their free throw ability in pressure situations.

Analytics methods used in this research: Random Coefficients Model
Random Decision Forests

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.


"The Thin Edge of the Wedge": Accurately Predicting Shot Outcomes in Tennis using Style and Context Priors

This is a review of the golf research conducted by Xinyu Wei, Patrick Lucey, Stuart Morgan, Machar Reid, and Sridha Sridharan, applying random decision forests.

Strategy is the name of the game in tennis with players continually attempting to maneuver their opponents into weak positions in order to score points. The final shot is not the most important element of the rally, but rather the strategy preceding the final shot. This research looks for the strategy behind the progression of a tennis rally used by top players.

Players systematically hit the tennis ball in a manner that will put their opponent in a weaker position, thus gaining dominance for themselves. Players that use similar strategies can be grouped together according to style. It is also important to look at the series of shots in a rally to determine context.

A player will work towards moving their opponent to a specific position on the court, and once this is achieved, they gain an advantage and are able to win the point.

Three years of data from matches in the Australian Open Men's singles draw, including 2292 winners and 37727 shots are analyzed, focusing on the top 10 players who played the most matches. Information regarding the position of the player, position of the ball, current score, point duration, server, and receiver identity are included.

Raw features such as the trajectory of the ball are included as well as dominance features including ground stroke speed ratio, ground stroke weight ratio, and lateral player movement ratio.

A Random Decision Forest is used to predict the probability that the next stroke is likely to be the winner. This provides a universal model for all tennis players. However, individual players each have their own unique style. A tennis player's style is discerned through learning a dictionary of shot trajectories from data by optimizing prediction performance and reconstruction error. This dictionary includes single shot elements as well as shot combinations. Shot combinations are grouped using a K-means algorithm. Evaluating style is then conducted via a reconstruction error and prediction loss process based on both single shot and shot combination information.
Context information is used to describe elements of the rally occurring before the point is scored. This includes the score line, elapsed match time, environment conditions such as wind and temperature, and the court surface.

Style and context combined together provides superior information for predicting the winner of the rally.

Analysts can this tool to make predictions during a rally as to the probable outcome. They can also compare and contrast the strategies used by the various players. Even when two players have rarely met before, or never met, analysts can look at players with similar styles to predict the outcome of the game. Coaches and players can use the style groupings to analyze future opponents, allowing them to prepare the best possible strategy.

Incorporating style, context, single shot, and shot combinations provides an increasingly accurate ability to predict the outcome of a shot in a tennis game.

Analytics methods used in this research: Root mean square error, ground stroke weight ratio, lateral player movement ratio, dictionary learning, random decision forest
Ranking

A ranking is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. In mathematics, this is known as a weak order or total preorder of objects. It is not necessarily a total order of objects because two different objects can have the same ranking. The rankings themselves are totally ordered.


What a Fairer 24 Team UEFA Euro Could Like

This is a review of the UEFA Euro research conducted by Julien Guyon, applying ranking strategies.

The UEFA European Championship, also called the Euro, is a competition among 24 men's national teams. It starts with a round robin stage, with 16 teams going on to the knockout round. Tournament rules dictate that 6 group winners, 6 runners-up, and the 4 best third-placed teams advance to the knockout round. The 6 third-place teams are ranked according to number of points attained; goal differential; number of goals scored; player conduct during the tournament, and position in the UEFA national team coefficient rankings in order to determine which 4 will move on to the next round.

As this tournament structure is asymmetrical due to the number of teams involved, it is not an easy task to create a fair knockout bracket. The current bracket has several strengths including balance - each half of the bracket consists of 3 group winners, 3 runners-up, and 2 third-placed teams. Group diversity is also a strength in that the 3 group winners and 3 runners-up in each half of the bracket come from 6 different groups.

The current structure also has several weaknesses such as the fact that it is advantageous to compete in certain groups in order to advance as far as possible. The structure is also arbitrary in choosing which group winners will play against third-placed teams and which will face runners-up. Finally, there can be a lack of incentive for teams to win, as it is not always clear whether it is best to finish first or second within a group.

It is possible to create fairer brackets if the teams who make it to the knockout round are ranked from 1 to 16 instead of using their group ranks. This eliminates any group advantage and increases the incentive to win, as it is better to win the group than be the runner-up as the runner up will have to face a higher ranked team in the knockout round. The format of the round robin stage remains the same and it retains its strengths regarding balance and group diversity. In the first round the number 1 ranked team plays the 16th ranked team, the second ranked team plays the 15th ranked team, and so forth.

The structure is more balanced as each half consists of 3 group winners, 3 runners-up, and 2 third-based teams. It is set up so that the top 8 ranked teams do not meet before the quarterfinals and the top 4 ranked teams cannot meet before the quarterfinals. It also eliminates any group advantage. The biggest flaw of the proposed structure is that it does not guarantee diversity; teams from the first group may face each other again in the knockout round, which can be resolved with some minor revisions.
However, this structure creates a logistic problem in that all teams will have to wait until the round robin stage is finished to know what their rank will be going into the knockout round. This then creates difficulties with scheduling games, travel plans for teams and makes it difficult for fans to make decisions regarding attending games involving their team. However, it is felt that the improved overall fairness of the new structure outweighs any potential problems.

Analysts can use this information to look at past tournaments and compare the actual knockout round with the proposed structure. The two structures could be analyzed to predict if the new structure would have changed the outcome.

As initially incorporating the new structure could create problems as unexpected problems arise, the format could be tested on a youth competition that follows a similar format.

Analytics methods used in this research: Ranking
**Regularized Adjusted Plus-Minus Rating**

With “Regularized Adjusted Plus-Minus” (RAPM), the goal is to provide more accurate results by employing a special technique called “ridge regression” (a.k.a. regularization). It significantly reduces standard errors in adjusted plus-minus (APM).

**Modelling the Financial Contribution of Soccer Players to their Clubs**

This is a review of the soccer research conducted by Olav Drivenes Saebo and Lars Magnus Hvattum, applying a regularized adjusted plus-minus method.

Soccer is widely played around the world. During the last ten years, European association football clubs have seen steadily rising revenues, particularly from media coverage. However, even though revenue is increasing teams typically do not maximize their profits as their balance sheets are in the red, showing high levels of debt.

European leagues do not have the limitations regarding salary caps, player drafts, and roster limits experienced by North American leagues. Squad size limitations are being put into place in several competitions but highly paid players are routinely traded. Teams could gain a competitive edge if they assessed the consequences of cost incurred in player trades.

This research looks at a framework that evaluates player transfers in these European soccer leagues by estimating the influence an individual player has on the team’s performance. This will give the club an indication of how much they should be willing to pay for that player. This framework evaluates each player in regards to their contributions to the team, predicts outcomes of future games based on the players involved, and predicts competition results.

The first step is to evaluate the players through a regularized adjusted plus-minus rating system. These ratings look at all the games in which the player participated. The rating is dependent on the other players participating at the same time such that a positive plus/minus rating against strong opponents is weighted more heavily than a positive rating playing against a weaker opponent. If a team consistently has lower scores while an individual player is on the field that player receives a lower rating than his teammates. The ratings do not rate the individual player's ability, but rather rates their relative performance during the matches.

Match outcomes are predicted using a probit regression model based on a single variable, which is the difference of the average plus-minus rating for the home team players and the average plus-minus rating for the away team players.

Next, competitions are simulated, taking into account player ratings and match outcome probabilities. In order to incorporate player ratings additional information is required regarding team squads and a model for player selection. Due to injuries, etc. player availability is uncertain and therefore the process needs to be fluid, as teams need to be chosen in regards to putting together a combination of players who have the greatest probability of beating their opponents during the competition.
All matches are evaluated separately with teams given two, one, or no points depending on the outcome. The points are totaled for each team and at the end of the simulated season, each team is given a conservative estimate of the expected financial income based on their positioning in the league. A Monte Carlo simulation is repeated 100,000 times in order to improve accuracy of the outcomes.

Coaches can use this model to run simulations based on different selections of players to determine the best combination for a competition. By putting in and taking out individual players, that player's financial contribution to the club can be determined, which is an indication of how valuable that player is to the team. This is especially useful when looking at making trades and acquiring new players.

The model as it currently stands has many uses but could be improved by adding other factors such as travel distance, importance of the match, playing surfaces, player fatigue, and long-term injuries.

Analytics methods used in this research: Regularized Adjusted Plus-Minus Rating, Probit Regression Model, Monte Carlo Simulation
R-Squared of a Linear Regression

In statistics, the coefficient of determination, denoted R2 or r2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. There are several definitions of R2 that are only sometimes equivalent. One class of such cases includes that of simple linear regression where r2 is used instead of R2. When an intercept is included, then r2 is simply the square of the sample correlation coefficient between the observed outcomes and the observed predictor values. If additional regressors are included, R2 is the square of the coefficient of multiple correlation. In both such cases, the coefficient of determination normally ranges from 0 to 1.


Can’t Buy Much Love: Why Money is Not Baseball’s Most Valuable Currency

This is a review of the baseball research conducted by Martin Kleinbard, applying R-Squared of a linear regression.

Over the years, there has been a complaint among stakeholders in the MLB that a payroll inequity creates an uneven playing field among the teams. Is there any validity to this argument?

If money plays a role in winning then there should be a trend over past years between winning and payroll, with teams that pay out larger salaries consistently being champions. To determine if this is true an R-squared linear regression of payroll on regular season wins is run. The higher the R-squared value the greater the relationship.

In order to create a competitive league it is necessary to minimize any impact afforded to a team who has the budget and willingness to spend large sums of money on players' contracts.

The first method used to test this idea is a cross-sectional analysis comparing the MLB with the NBA and NFL. The Win Buying Index is compared to the Payroll Inequality Index of each league through the 2001 to 2012 seasons. MLB, the league with the fewest restrictions on payroll, has the highest index in both categories while the NFL, the league with the greatest payroll restrictions, has the lowest. However, the Win Buying Index is below 20% for all three leagues, indicating a poor relationship between payroll and winning. The Win Buying Index and Payroll Inequality Index for the NBA and NHL, both of which institute salary caps, have a negative correlation clearly indicating that a reduced payroll inequality is not the cause for their lower indices.

The second method used to test the idea is a longitudinal analysis looking strictly at the MLB over its history. Free agency was instituted in 1977, however, prior to this was an Under the Reserve Clause system stating that any team that initially signed a player had total control over his contract throughout
the player's career. This created a low payroll inequality, which did not create a low Win Buy Index, but instead a larger one. This was likely due to the fact that while teams had control over contracts, only the wealthiest teams were able to maintain extensive scouting networks, giving them the ability to find the best talent available.

When the draft was implemented in 1965, the average Win Buying Index dropped across the league. All teams now have the ability to acquire talented young players who will receive a lower payroll. This raises the question of how dominant young players are relative to veteran players. To answer this, a new statistic, Youth Dominance Index, is created. It measures the percentage of strong, dominant seasons produced by younger players. Indeed, there is a negative correlation between the Youth Dominance Index and Win Buying Index, indicating that youth dominance, rather than payroll inequity, has a strong influence on the league and competitiveness of teams.

Coaches and analysts can use this information when looking to put a strong team together. As youth is a stronger predictor of winning than payroll, the focus should be more on growing and strengthening young players, rather than paying enormous sums of money to veteran players.

Like everything else there needs to be a balance. A team consisting strictly of young players without any veterans to guide them is not likely to win, so teams must look for the optimal balance between young and veteran players.

Analytics methods used in this research: R-Squared of a Linear Regression, Cross-Sectional Analysis, Win Buying Index, Payroll Inequality Index, Longitudinal Analysis, Youth Dominance Index.
Sales forecasting has been a growing trend in the world of sports, where it has been used in an attempt to predict outcomes of games. Today we will be stepping into sales forecasting and the business model relating to sports analytics.

Sales forecasts are quite essential in sports analytics. The prediction markets offer a very promising approach for the prediction of future events. They are now increasingly being deployed to collect information on sporting events, fan behaviors and memorabilia, etc.

Throughout our lives, we make thousands of forecasts, most of the time without realizing or giving the deserved importance to this activity. In the case of sports organizations, this cannot be the case, as they generally cannot make sound decisions without making informed forecasts.

Sports sales forecasts are the estimate of future sales, taking into account the environmental conditions. It is carried out based on the analysis of the trends, trying to deduce to what extent the historical data of invoicing will influence in the future, and of the analysis of other influential factors: factors of the environment, evolution of the sector, potential of sales of the company, the quality and price of the product or service, analysis of the competition, etc.

When you make a sales forecast, remember that it is very important that you consider three possible scenarios:

1. A positive one where sales are increasing
2. A neutral one where sales are maintained
3. A negative one in which sales are nosediving

The general sales forecasts seek to find out what part of the market potential we can achieve for our company, given our economic, technical and human possibilities. The forecast for sales is a process that consolidates the participation and influences of other functional areas within the organization.

Such influences can occur in two ways:

- **Unidirectional**: This refers to the fact that forecast sales influence the decisions taken by other functional areas
- **Bidirectional**: This is where the forecast is used to quantify the market effects according to the changes foreseen by other functional areas.

Therefore, we intend to define our sales potential based on:

- The total capacity of the market.

We need to ask the following while making a sales forecast.

Is there a potential market?
Is the market accessible to us? Can we overcome barriers to entry in the market? Can we achieve a sufficiently good positioning that allows us to sell?

How is the competition doing? Does it satisfy its consumers? What part of this market could we take away? Will our survival depend on just a general market growth?

How are the forecasts?

- According to time: Immediate, short, medium or long term.
- According to the type of data: Subjective (on opinions or personal intuitions), statistics (on internal historical data), economic (on external historical data).
- According to the nature of the product: The provisional methodology is different if it is a pre-existing product in the company, or in the market, or it is totally new.

According to the amplitude: We can make forecasts only about our products and sales, but also especially on the market with more or less breadth.

Benefits of carrying out a sales forecast

- It improves interdepartmental communication
- Improves the profitability of procedures
- Improves customer service
- Helps anticipate and solve future anomalous situations
- Improves the forecast of human capital
- Reduces product rotation
- Avoids breakage of stock
**Scatterplots**

A scatter plot (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are color-coded, one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.


**Evaluation of Pace of Play in Hockey**

This is a review of the hockey research conducted by Rajitha M. Silva, Jack Davis, and Tim B. Swartz.

The pace of play is a characteristic that influences the playing style in many sports. Typically, when the pace of a game is high, the game is smooth with greater opportunities to score. In hockey, a high-paced game is viewed as one that moves from end to end. However, there is common measurement for the pace of play.

The pace of play in hockey is regarded as fast when players are rushing back and forth from end to end, first attacking, and then retreating. There is little opportunity to become organized in the offensive zone. Counter attacks often occur due to the aggressive pace. Teams moving sideways across the ice carrying the puck or passing backwards are slowing down the pace.

In order to measure pace of play, pace contribution is defined only when the team is moving forward and in possession of the puck. Dumping in the puck or moving sideways does not contribute to the pace. From this, the total attacking distance in a game is determined as the sum of forward attacking distances (in feet) covered by both teams in a game. As tactics are different when teams are shorthanded, total attacking distance is determined only when teams are playing 5 on 5.

Scatter plots are created for total distance versus total goals and total distance versus total shots. Neither graph demonstrates a positive correlation, which is not what one would expect. The graphs indicate that high-paced games actually create fewer shots and fewer goals. To further test this outcome scatter plots were created for overtime 3 on 3 play and for play in the first period. The outcome remained the same.

To further investigate, attacking distance was only counted when it was done at sufficient speed, taking into account both distance and time. The definition of sufficient speed was adjusted in different trials, yet the results remained the same, there was no positive correlation.

A third attempt was made where attacking distance was modified to include only distanced covered between the blue lines. There was still no change in the outcome.

Finally, instead of measuring distance, pace was measured based on the number of times alternate zones were entered. This scatter plot showed a slight positive correlation but not significant enough to warrant further investigation.
Against our intuition, high-paced games do not result in greater chances for scoring. In fact, the opposite is true, more scoring chances are generated in slower paced games. This could be because teams have time to set up in the offensive zone in a slower paced game.

Coaches can use this information in formulating strategies, realizing that a slower pace will provide their team with more scoring opportunities. However, they will also have to keep in mind that the slower pace will also give their opponent greater scoring opportunities. They will need to determine what the optimal pace for their team is in terms of optimizing scoring chances for themselves and minimizing score chances for their opponent.

It is surprising how often ideas that we consider must be true are proven false when examined more closely.

Analytics methods used in this research: Scatterplots, Correlation
Shot Chart Analysis

Shot Chart Analysis as a Method of Sports Analytics
In this chapter we look at a very important technique that has proved to be an effective tool for improving team sporting performance - Shot Chart Analysis. This analytic method is commonly used in the game of basketball, a game involving 10 players. The shot chart analysis method can be used to determine the strengths and weaknesses of various offenses and defenses performed on the sports field. It compiles much of the game’s important information in one place.

Application of Shot Chart Analysis

1. Shot charts can help sport coaches understand how their several decisions in a game can influence team’s performance. For example, how the team’s field goal percentages are affected when different players are substituted into the on-court lineup and how a player affects the shooting percentages of his teammates.

Two of the most popular shot charts used are provided by NBA.com and ESPN.com. NBA Hotspots and ESPN Shot Chart. These shows how well an individual player or team shoots against another team as a whole.

NBA Hotspots splits the court up into various regions to allow users view how well a player or team shoots from the different regions. Thus, helping them understand how often a particular player or team shoot from a particular position.

ESPN Shot Charts marks the court as a map of made and missed shots. It provides more information about specific shots of players, such as additional information about the shooter, the distance of the place of shot from the basket, and the time of the attempt.

Teams can use these charts to change their line-ups based on the opponent’s players on the floor.

2. Through the use of shot charts, analysts can determine how players’ field goal percentages vary between several locations. This information further guides coaches in choosing preferred shot location for different players.
3. It can also be used to evaluate team defense by analyzing their points distribution as well as the most common locations where the opponent scores. This provide the team an opportunity to evaluate their strengths and weaknesses of offenses by determining if its opponent spreads across the floor with its shots, thus figuring out areas of high percentage shots when a type of player combination is on the court.

The shot chart can also provide player interactions but it demands that more information be provided. Information like, which player made the shot attempt, his location on the court, and the location of other nine players when the shot was made.

Tracking specific individual contributions on the court and their interactions with the other 10 players creates a dataset that allows users view team interactions in a new way.

In conclusion, the use of shot charts has greatly improved the performance of teams on the basketball court. With this analytic method, coaches and teams can determine scoring averages, strengths and advantages of certain players, thereby, determining what actions are best to take on the basketball court. Increasing the detail of information on shot charts would allow users understand a team in even greater depth.
Shot Chart Framework Encoding

This analytic method is commonly used in the game of basketball. The shot chart analysis method can be used to determine the strengths and weaknesses of various offenses and defenses performed on the sports field. It compiles much of the game's important information in one place.


Information Visualization in the NBA: The Shot Chart

This is a review of the shot chart analysis research conducted by Stephen Chu.

Using statistics to analyze NBA game performance is becoming a higher priority with many teams hiring more statisticians, hoping that this will give them a competitive edge. Current NBA statistics focus mainly on the production of the individual player and do not look at how players interact.

One possible useful basketball analytical tool is the shot chart as it compiles much of the game's important information in one place. It can also be used to determine the strengths and weaknesses of various offenses and defenses. Increasing the detail of information on shot charts would allow users to understand a team in even greater depth. One possible application would be understanding how the team's field goal percentages change by substituting different players into the on-court lineup which would indicate how a player affects the shooting percentages of his teammates.

Current shot charts do not allow for the in-depth analysis of how well specific combinations of players work together or play against each other. Two of the most popular shot charts are provided by NBA.com and ESPN.com, which show how well an individual player or team shoots against another team as a whole but do not provide data on specific lineups.

NBA Hotspots splits the court up into various regions. This allows the user to see how well a player or team shoots from the different regions. In turn, it allows the user to understand how often a particular player or team shoot from a particular position.

ESPN Shot Chart marks the court as a map of made and missed shots. It shows much more detail about the specific shots such as extra information about the shooter, the distance from the basket, and the time of the attempt. However, shots are often closely clustered together, making it difficult to single out individual shots.

NBA statistics typically focus on individual stats but do not recognize that a player's teammates will have either a positive or negative impact on his play. All 10 players on the court influence each other and this should be taken into account to create more accurate statistics. A new way of compiling stats needs to be implemented to overcome this problem.

For the shot chart to include player interactions it must include more information including who made the shot attempt, where they were on the court, and where the other nine players were when the shot was made. Tracking specific individual contributions as well as their interactions with the other 10
players creates a dataset that allow the user to look at team interactions in a new way. The analyst can use the statistics individually or combine them in order determine scoring averages, where players most frequently shoot from, and the position of the other players when they do make the shot. It makes it possible to look at how well a team performs when a certain player is on the floor.

This shot chart gives analysts a tool to evaluate NBA teams more in-depth, looking at the performance of specific lineups on the court. It is possible to see how team performance is affected by substituting players in and out of the lineup. Analysts can determine how players' field goal percentages vary between locations and the preferred shot location for different players.

Teams will be able to evaluate strengths and weaknesses of offenses by determining if the team spreads across the floor with its shots and figuring out the areas of high percentage shots when a certain player combination is on the court. It can also be used to evaluate team defense by looking at their points distribution and the most common locations where the opponent scores. Finally, teams can use this shot chart to change their lineups based on the opponent's players on the floor.

Analytics methods used in this research: shot chart framework encoding, expected value
Simulation

Simulation as a Method of Sports Analytics
In this chapter, we will look at another critical approach that is used in sport analytics, and that is simulation. Coaches, managers, or sport analysts employ simulation method to predict the outcome of events or performances of particular sportspeople or teams. This allows the managers to predict the chances of his team against opponent in order to enhance the performance of his team through training or formation lineup.

More often for sports analysts who enjoy predicting the chances of a particular team winning a match perhaps for the purpose of giving hints for a bet, simulation allows them to recreate the sport events through statistical models and to foretell the likelihood of the success of a team over a period of time.

This method starts with observation of the team performance over a particular period of time. This observation allows analysts to develop a model for making win-loss predictions, the progression of matches (such as changes in lineups of teams during a season, ball movements as the match progresses pass by pass, defensive and attacking strategy of the teams), and the margin of error during simulation.

Case Study of How Simulation Methods are Used
Let’s take a look at basketball match and see how coaches or analysts make use of simulation to predict various possibilities during the match, using the professional basketball player Kevin Durant of Oklahoma City Thunder as a case study.

First of all, we consider a game or season as an “experiment”. Then the actual results observed of a player or team over the course of a season will reflect the natural randomness of that player or team which forms the “data set” of the experiment.

For our case study, observing the performance of Oklahoma City Thunder over the period of six seasons, it is possible to build a statistics model which shows the position of the team in each of the six seasons, number of titles won in each of the six seasons, and the strength of the players across the six seasons. By building the model, sport analysts can recreate basketball events through the models built. This model is then simulated for unreal matches which help them predict what will happen in real life matches over the next one or two seasons subsequently after the six seasons which the team’s success is modeled upon.

Let’s a deeper look at how simulation helps make predictions based on the match progression of a team. It is very common to see changes during a match—like formation lineup, change of players during the match, or attack-defense strategy. This kind of changes is a very sensitive thing to do because it plays a key role in the success of a team. So coaches are caught in the habit of making models and simulation to help them determine the performance of their teams against opponent during a match by changing the pattern of the team.

For instance, the manager of Oklahoma City Thunder might discover that taking out Kevin Durant during a match would put the team out of possession, or switching the ball movements when the team is under-performing in a match can change the dynamism of the team and spur the team performance which will eventually increase the team’s chances of winning.
With the progression of a basketball match using a probabilistic graphical model, a coach can predict the best strategy that will produce winnings.

In conclusion, simulation as a method of sports analytics is widely used by various professionals. This is because it helps coaches and sport managers adopt the best tactics which will yield the most positive results. Also, in the area of sport betting, simulation helps developers build a virtual sport game based on the real-world performance of the team and players.

**Simulation in Sports Analytics**

More than 10 years ago, Honeywell's aerospace engineer, Barry Bixler, thought of joining aerospace simulation with sports and used Fluid Dynamics in his spare time to analyze the flow pattern around the arms of swimmers.

*Swimmers*

The swimming teams perceived the potential of the Bixler's studies and began to apply their recommendations to improve the performance of the athletes.

Since then, computational simulation has been gaining ground and is gradually applied in different ways in sports to improve athletes' performance, ensure comfort and reduce the risk of injury. This happens because the simulation allows predicting the behavior of a device, an athlete or a system that involves the athlete, devices and sports equipment under certain conditions.

*Simulation in Sport*

To be able to predict the behavior of these models, the simulation technology solves fundamental equations, such as conservation of mass, conservation of energy, Newton's second law or Hooke's law of elasticity to calculate magnitudes as speed, pressure, tension, deformation etc. Even the simplest models can provide interesting information about a system made up of the athlete, the equipment and the environment that surrounds them.

Through computer-based modeling, it is possible to determine and understand how parameters can impact sports performance and minimize or amplify an injury. Analyzing and foreseeing the consequences of these modifications means that sports team designers can better select the set of parameters to optimize performance and reduce the risk of injury. In addition to this, manufacturers can quickly launch improved products to the market and with a lower development cost.

*The Value of Simulation Engineering in Sport*

To better understand what is the importance and impact of these tools in sport, the report, "Dramatic Changes in Sports: The Contribution of Engineering Simulation" indicates how three different specialists focus on contributions that computer simulation made and can do in what refers to the most varied sports, whether high performance or not.

In order to make these computational models more and more realistic, there is a tendency to create and test more complex models and predict their behavior. New capacities are also added to interpret the
environment, increase the fidelity of the models and incorporate combinations based on a system composed of the athlete, the product and the environment.

The simulation is already applied - occasionally and sometimes systematically - in different sports and has brought significant achievements in either performance or comfort. It also allows analysis of isolated products as well as complex and complete systems, taking the simulation to a deeper and more detailed study. This is important in the case of cycling, where the two main performance parameters of the athlete are aerodynamics and weight.

The reduction of the weight of the bicycle can be achieved, without compromising aerodynamic resistance, by means of variations in the geometry of the different components and the use of new materials, such as composites. In addition, the position, the comfort of the athlete and the format of the components of the bicycle are decisive for a good performance. However, it is important to note that to improve overall performance, in addition to optimizing each subsystem, it is important to understand that the final result depends on the interaction of the entire system, therefore, the performance of the system as a whole is essential, instead of optimizing each component in isolation.
Social Network Analysis

Social Network Analysis in Sports Analytics
This chapter introduces sports social network analysis and how sports teams are leveraging social network analysis and the internet to increase team sales and marketability. And we’ll take a deeper dive into how social networks already play a fundamental role in many of the communication strategies of successful college and professional teams.

The constant evolution of the internet throughout history has highlighted the importance for players and teams to adapt to these changes, and learn how to manage social networks. In recent years, social networks have become a strong exponent of these changes, impacting the way we communicate in all areas.

Social networks have become a strong channel of dissemination in the sports world and have managed to become the perfect platform for brands to achieve visibility and direct contact with their followers. These channels have seen that sports marketing is now much more veritable. More and more big brands are betting on promoting their campaigns on social networks.

According to statistics, in recent years, there has been an increase of 38% in advertising campaigns on social networks related to sports. The activity of this sector usually brings large advertising investments, so it is not surprising that advertising in this medium is expected to represent 20% of all advertising on the Internet in 2019, according to Zenith Media, in its 2018 advertising forecast.

Thanks to social networks, not only brands can have greater access to customers, but the athletes themselves can show more through these channels. And fans can enjoy learning more about the day to day life of their favorite stars.

Knowing this reality, the sponsored tweet has become the widespread format of choice social networking. According to data from Opendorse, the consultancy specialized in personalities, sports and its dissemination, more than two thousand professional athletes signed more than five thousand publications sponsored on Twitter. Figures show how big companies related to sports bet on working with professional athletes in their advertising campaigns due to their influential role in these channels.

Thus, these sports influencers have managed to amass a large community of followers who follow their publications. According to Forbes, 85% of users trust more in the content generated by influential users.

This past year’s Super Bowl became one of the sporting events with the most impact on social networks. Facebook and Instagram were the platforms most used during the match and between them were recorded about 110 million mentions. On Twitter, the event’s hashtag was a global trend with more than 27 million tweets.

In these channels, athletes have an unparalleled influence on social networks and even have more engagement power (degree of connection with the audience measured in interaction with the ad that the brands themselves sponsor).

According to Opendorse, the average participation rate in social networks of these athletes is more than ten times that of the teams’ and 50 times more than the leagues’.
The numbers are clear and the combination of social networks and sports has become for brands a highly effective combination to engage digital sports fans.
Spatial and Visual Analytics

Visual analytics is an outgrowth of the fields of information visualization and scientific visualization that focuses on analytical reasoning facilitated by interactive visual interfaces.


Spatial analysis or spatial statistics includes any of the formal techniques which study entities using their topological, geometric, or geographic properties. Spatial analysis includes a variety of techniques, many still in their early development, using different analytic approaches. In a more restricted sense, spatial analysis is the technique applied to structures at the human scale, most notably in the analysis of geographic data.


CourtVision: New Visual and Spatial Analytics for the NBA

This is a review of the spatial and visual analytics research conducted by Kirk Goldsberry, Ph.D.

Basketball players need to have a strong spatial ability which is the ability to understand, reason and remember spatial relations among objects or space. This spatial ability is related to all aspects of the game including choosing where to shoot from, which defensive scheme to employ or even where to place yourself on the court. Every player and team has their own unique spatial abilities and behaviors.

"CourtVision" is a new model designed to analyze spatial information. It employs database science, spatial analysis and visualization. From this it is possible to see a player's or team's unique spatial style. These unique styles lead us to discover patterns in performance.

Spatial and visual analytics can provide insight into NBA shooting and differences in shooting skills around the league. It can also look at the value of space on the court. Visual analytics uses graphical representations to help improve the analytical reasoning process, in order to facilitate good decision making. Spatial analysis can be used to attempt to answer questions such as who is the best shooter in the NBA.

The concept of space is critical to the game of basketball. Shooting ability is dependent on space. Players need to understand the spatial relations between the ball and hoop in order to be able to put the ball into the net. Court space affects all aspects of the game and how players deal with their space can make the difference between being an average player and a star player.

The unique spatial ability of the players and teams has a strong influence on the competitiveness of the team and their ability to win games. CourtVision was designed to analyze and quantify this spatial ability. This will aide coaches and analysts in determining which spatial abilities are their strengths and in which areas they have weaknesses. They can also look at which spatial abilities make other teams and players so effective in what they do.
Spatial and visual analytics also answers the question of where. Where does a specific player prefer to shoot from? Where should a defender position themselves in order to be the most effective?

Coaches and teams can use spatial analysis to determine game plans for future games that can effectively defend their space on the court and effectively score offensively. It can be used to evaluate individual players when looking at making trades in order to determine who has the spatial ability that is most needed by the team. It also gives coaches insight into the weaknesses of their team allowing them to design a practice regime that will facilitate improvements across the team leading to increased competitiveness.

Analytics methods used in this research: spatial and visual analytics, metrics, visualization
Spatial Data Analysis

Characterizing the Spatial Structure of Defensive Skill in Professional Basketball
This is a review of the spatial data analysis research conducted by Alexander Franks, Andrew Miller, Luke Bornn and Kirk Goldsberry.

Basketball players must wear two hats - one for the offensive side of the game and one for the defensive side. However, most statistics deal only with their offensive skills, and not the defensive. Player tracking systems have the potential to change this.

Player tracking data from the 2013-2014 NBA season is used to develop a model that will allow analysis of both the offensive and defensive components of the game. First, the spatial data analysis model must identify which defender is guarding which offender at every point in the game. The results lead to the conclusion that defenders typically take up a position two-thirds of the way between the hoop and the offender they are guarding. When defending against the ball carrier the defender wants to be closer so they tend to take a position three-fourths of the way between the ball carrier and the hoop. Defense can be analyzed by determining how they affect the shot selection and shot efficiency of the offender they are guarding.

Next, shot types are added to the spatial data analysis model. Then the model is adapted to account for where different offenders prefer to shoot from on the court as well as the probability that the player is successful in making the shot. The probability that the player is successful is formulated using the offensive player's skill, who the defender is, the distance between the defender and the shooter, as well as where the ball carrier was on the court when he took the shot. Obviously, the further the defender is from the shooter, the less effective he is at stopping the shot.

Data analyzed determined that defenders can affect both the shot frequency and shot efficiency of the shooter. Some defenders are more skilled at affecting the shooter's shot frequency, while some are more skilled at affecting the shooter's shot efficiency. There are also those who are effective in both areas. The model also provides information regarding how shooters perform against specific defenders.

Coaches can use the model to analyze how much defensive attention each of their players receive. Players who receive more defensive attention would be those the defense considers to be the stronger threat. Coaches can also look at how to take advantage of this situation to identify which players are receiving less defensive attention. Also, looking at how individual shooters perform against specific shooters can help the coach decide which defensive matchups should be made.

Analytics methods used in this research: frequency, efficiency, expected outcome hidden Markov model, constrained generalized least squares, parameters
Spatio-Temporal Dynamics

Local interactions in space can give rise to large scale spatio temporal patterns (e.g. (spiral) waves, spatio-temporal chaos (turbulence), stationary (Turing-type) patterns and transitions between these modes). Their occurrence and properties are largely independent of the precise interaction structure. They are indeed seen to occur at many organizational levels of biotic systems. Space can be either 'real' space or a state space, e.g. 'phenotype space' in models of speciation or 'shape space' in immunological models of shape-based receptor interactions. We show that such spatio-temporal patterns have important consequences for fundamental bioinformatic processes.

Source - http://www-binf.bio.uu.nl/overview/node3.html

Wide Open Spaces: A Statistical Technique for Measuring Space Creation in Professional Soccer

This is a review of the spatio-temporal dynamics research conducted by Javier Fernandez and Luke Bornn.

Soccer statistics are heavily focused on what happens when a player has the ball. However, what about the rest of the players on the field? Statistics have shown that players only have the ball for an average of three minutes per game. What are they doing the rest of the time?

Professional soccer teams have been increasingly adapting their playbooks to incorporate strategies for those players who do not have the ball in order to generate better scoring chances. One aspect is space occupation or the ability of a player to create space for himself in highly valuable areas - areas from which he/she will be able to assist the team in creating a scoring opportunity. A second aspect is generating space. This involves a player getting an opponent to focus on him/her in order to open up space for a teammate.

A model using player tracking data was used, as tracking data is more efficient and detailed in gathering information on human behavior. The model was used during a first division Spanish league soccer match. It gathered data regarding where players were located on the field, their velocity and how far they were from the ball. The position of the players was analyzed to determine what influence a player had in that particular area. The importance of that area to the team was also analyzed. Areas close to the opponent's goal are always valuable but other areas become valuable depending on where the ball, teammates and opponents are located. The model evaluated the value each space had at a given point based on the location of the players on the field. Generally, most players will remain close to high value areas which is then what the model uses to determine importance.

Coaches spend a great deal of practice time focusing on the players' ability to create and occupy spaces. This model can be used to help coaches determine how well their players are using space on the field and what areas need to be improved upon. Players have strengths and weaknesses in occupying and generating space and may be skilled in certain areas of the field. Coaches can look at this data, using players' strengths as teaching tools for other athletes and using weaknesses to generate coaching plans to help the player improve their skills in that area.
In conclusion, the concept of space in a soccer game is extremely important given the minimal time each player has the ball within a 90-minute game. It is also a highly complex idea that requires further in-depth analysis in order to help coaches maximize their team's potential.

Analytics methods used in this research: parametric pitch control model, model for relative value, spatio-temporal dynamics, parameters, mean square error
Spatiotemporal Trajectory Clustering

Spatio-temporal trajectory clustering methods identify heterogeneous patterns and explore underlying mechanisms and are designed to consider both temporal and spatial information in trajectories. Applying the time-dependent shortest-path distance measurement and taking advantage of topological relations of a predefined network, this algorithm can define the shared sub-paths among trajectories and construct the clusters.

Quantifying the Value of Transitions in Soccer via Spatiotemporal Trajectory Clustering

This is a review of the spatiotemporal trajectory clustering soccer research conducted by Jennifer Hobbs, Paul Power, Long Sha, Hector Ruiz and Patrick Lucey.

In soccer, transitioning from defense to offense and vice versa is extremely important. However, there are no methods to quantify or rank the effectiveness of these transitions. In order to create a model to accomplish this a playbook of the most commonly used plays was put together. Data was collected from the 2016-17 English Premier League. This data was then used to create a playbook of those plays that were commonly used by the various teams throughout the season. Analyzing this playbook generates information regarding which plays were the most effective, and where players should position themselves on the field to be the most effective.

Spatiotemporal Trajectory Clustering model can also be used to determine which offensive plays lead to goal-scoring opportunities. Plays can be measured and ranked based on their probability of a shot being taken. Counter-attacks can also be evaluated. Reviewing the data for counter-attacks demonstrates that more shots are created from counter-attacks than other plays. In fact, shots are not only more likely to occur directly following a counter-attack but also at any point in the possession resulting from the counter-attack. The results also indicate that these shots are more likely to result in a goal than shots made at other times. Finally, the model also looks at how teams transition from offense to defense. It looks at the time required for a player to get into their correct position on the field during the transition period.

Spatiotemporal Trajectory Clustering can be very valuable for analysts and teams. Those plays that rank high in the probability of leading to a shot can be analyzed from both an offensive and defensive point of view. Offensively, teams can determine how these plays can be utilized most effectively by their players. Defensively, teams can work on how to best defend against these plays in order to minimize the chances of the opposing team scoring. As counter attacks have a greater probability of leading to a goal, these plays need to be analyzed carefully, again from both the offensive and defensive point of view. Teams can also look at the transitioning time period. Where are their own players located? How do they transition and how quickly are they able to do so? Is there a time of disorder on the field when players are out of position giving the opposition a better chance of obtaining a shot? Teams can incorporate all of their findings into their own playbook and training regime in order to maximize their offensive chances and minimize their defensive losses. It also enables the team to work as a cohesive whole and gain a better understanding of what is most effective for the team as a whole.
Analytics methods used in this research: hierarchal clustering, machine learning technique using supervised and unsupervised learning
Sports Injury Risk Model

The various measures of injury incidence are injury risk (proportion of athletes injured in a given period of training, playing, or other exposure time), injury rate (number of injuries per unit of exposure time), odds of injury (probability injury will happen divided by probability injury will not happen), injury hazard (instantaneous proportion injured per unit of time or mean injury count per unit of time), and mean time or mean number of playing exposures to injury. Effects of risk factors are estimated as values of effect statistics representing differences or ratios of one or more of these measures between groups defined by the risk factor. Values of some ratios and their sampling uncertainty (confidence limits) are estimated with specialized procedures: odds ratios with logistic regression, rate ratios with Poisson regression, and hazard ratios with proportional hazards (Cox) regression. Injury risks and mean time to injury in each group can also be estimated and can give a better sense of the effect of a risk factor.


Low External Workloads Are Related to Higher Injury Risk in Professional Male Basketball Games

This is a review of the sports injury risk research conducted by Toni Caparros, Marti Casals, Alvaro Solana and Javier Pena.

Injuries are a fact of life within professional sports. Every season teams lose money due to injured players. Games are lost, fans are disappointed, and players are unable to play the game they love. This study was conducted to order to determine risk factors for injuries in professional basketball players in order to develop some preventative strategies to minimize injuries in the future.

A professional male basketball team was followed for three seasons. The data collected included 2613 observations, 246 games, and 33 different players. Specifically, the observers were looking for contact and non-contact injuries that lead to time loss. Time loss meant that the player, at a minimum, missed the next practice or game. Injuries included in the study occurred during practice or an actual game. Physiological, speed, distances, mechanical load, locomotor, and performance variables were all tracked.

The average age of the players involved was 24.9 years. Over the three seasons 11 of those players were injured. This included 29 time-loss injuries for a total of 244 missed games. The two variables that were highly related to time-loss injuries were fewer decelerations and less distance covered - in other words, players who had a lower workload were more likely to be injured. This specifically related to players who decelerated less than three times or covered less than 1.3 miles. The number of minutes a player is on the court would be the most obvious reason for a lower workload but it might also be related to readiness, performance, fatigue or even dependent on the particular of their opponents.

Teams need to look at how to best minimize time-loss injuries and sports injury risks. The players' workload needs to be closely managed, ensuring that all players work hard enough to decrease their chances of being injured. Part of this can be done through practices. Decelerations are related to the
players' ability to change direction. This would indicate that strength workouts are important. Teams may need to give some players more playing time at the beginning of the season in order to help prevent injuries later in the season when injuries can critically hurt a team's chances of making it into the playoffs and winning. Players can experience peaks in their workloads (especially during highly competitive games). When this occurs teams need to make sure the players have a sufficient recovery period. The amount of time needed varies by player so teams need to have individual plans in place to meet the needs of each player.

In conclusion, teams can use tracking systems during practices and games in order to monitor the workload each player is experiencing, especially in the areas of decelerations and distance covered. Players that are identified as having a lower than acceptable workload will need to have their routines altered. This will help lower the risk of injury to the players, a positive gain for everyone involved.

Analytics methods used in this research: observational respective cohort study, multivariate analysis
Statistical Prediction Model

Predictive modelling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect patterns, after the game has taken place. In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data.


Draft by Numbers: Using Data and Analytics to Improve National Hockey League (NHL) Player Selection

This is a review of the statistical prediction model to improve NHL player selection conducted by Michael E. Schuckers.

Every year NHL general managers need to decide who they will choose in the draft. A statistical model was built to rank players as to how they should be chosen in the draft. The model included height, weight, statistics regarding points per game and goals against average, and ranking from the Central Scouting Service. When looking at the future performance of the draftees it was determined that the model did a better job at ranking the draftees than the general managers did.

For this statistical prediction model, data was collected for two different groups. The first group consisted of 1398 players that were eligible for the draft from 1998 to 2002. The second group consisted of 1863 players eligible for the draft from 2003-2008. The first group was chosen so that statistics could be gathered from their professional careers. The second group was chosen to test the model built from the first group. Information was gathered from the year prior to their eligibility for the draft. Information gathered included height, weight, birthdate, league played in prior to their eligibility, number of games played, points scored, goals against average, and the CSS player ranking. The players' first 7 years in the NHL were then analyzed, based on the number of games played and their time on ice. Using games played and time on ice as indicators allows comparisons to be made across all positions (other than goaltenders). Therefore, there is not any bias towards any particular position.

Results from the actual draftees vs. those chosen by the model were compared from two different years, 2007 and 2008. Those drafted in 2007 by the NHL had a correlation of 0.547 in regards to time on ice and 0.547 regarding games played in the NHL. (The highest correlation possible is 1 so the higher the correlation the better.) Those drafted by the model had a correlation of 0.667 and 0.670 in those same areas. Players drafted by the NHL in 2008 had a correlation of 0.553 regarding time on ice in the NHL and 0.557 regarding games played. Those drafted by the model were again correlated more highly at 0.670 and 0.655. The model consistently demonstrates that it does a better job than NHL general managers of choosing which players to draft.
This model gives analysts data against which to compare their own choices for the drafts as well as those made by the NHL teams. It would provide a catalyst for talks regarding the need to change how these decisions are made.

Clearly, NHL general managers need to take a closer look at how they are choosing their draftees. More care needs to be taken that all factors that relate to future performance are taken into account. Managers may want to use a model like the one discussed here to help them in this process.
Stochastic Models

"Stochastic" means being or having a random variable. A stochastic model is a tool for estimating probability distributions of potential outcomes by allowing for random variation in one or more inputs over time. The random variation is usually based on fluctuations observed in historical data for a selected period using standard time-series techniques. Distributions of potential outcomes are derived from a large number of simulations (stochastic projections) which reflect the random variation in the input(s).


Analyzing the Pace of Play in Golf

This is a review of the approximation formula and stochastic model research for analyzing the pace of play in golf, conducted by Qi Fu and Ward Whitt.

Golf courses are always looking to maximize profit without affecting the golfers' enjoyment of the game. Stochastic models and computer simulations provide information regarding how to optimize the organization of a golf course from the best interval between tee times to how many groups should be scheduled to play each day.

An approximation model is developed using single-server queues, without precedence constraints, followed by an approximation formula for the expected value of the length of time it takes a group to play the entire course. The approximation is designed for use with a course that is in high demand and is well balanced with no bottlenecks.

On a course at any point in time, there will be golfers with widely varying skill levels. To handle this, stochastic models of group play are created for each of the four types of golf holes, par 3, par 3 with wave up, par 4 and par 5. Performance measures for each group for a hole are the waiting time, playing time, and total time. These models can be combined to correspond with a golf course's series of holes. The capacity of each hole is determined as well as the critical playing time, which is the time it takes a group to play the hole.

Using this information approximation formulas for the mean and standard deviation of the total time it takes a group to play a round of golf, relationships of intervals between tee times and the maximum players playing each hole to traffic intensity are generated. The formula assumes that the golf course is balanced with no bottlenecks. Assuming a balanced golf course lends itself to using only par 4 holes in the analysis.

The approximate performance formula for total playing time can be used to help design golf courses by creating an optimization problem with the purpose of maximizing the number of groups of players allotted time to play each day.

Simulations can be run to estimate the expected performance descriptors for any number of groups on any golf course regarding stage playing times and tee times. Courses can be tested to determine if they
are, in fact, balanced. This can be done by estimating the stage playing time and mean critical cycle for each hole. If they are approximately the same for each hole, the course is balanced. Another way to test if a course is balanced is to look at waiting times for each hole. Courses that are not balanced tend to have a few holes with longer waiting times, which then create longer waiting times at successive holes.

Tournament organizers can use this information to determine which golf course is the most suitable for their purposes. Course designers can use the model to help design a course that is balanced, allowing a maximum number of golfers to play the course on any given day.

Analytics methods used in this research: Stochastic Model, Computer Simulation, Estimation Formula
Strike Zone Plus Minus

Strike Zone Plus/Minus is a pitch framing methodology that “divide[s] the credit for whether a pitch is called a ball or strike among the catcher, the pitcher, the batter, and the umpire involved. It is different from other pitch framing methodologies in that it takes into consideration additional actors than the catcher.

Who Is Responsible for a Called Strike?

This is a review of the MLB strike zone research conducted by Joe Rosales and Scott Spratt.

Every baseball game creates some controversy regarding the calls of strikes and balls. No one ever completely agrees with the umpire and umpires, being human, are not 100% accurate. It seems that some players in the MLB have a touch when it comes to getting pitches called as balls. This is often attributed to the catcher. While catchers definitely do have an influence, so does the pitcher, the batter, and obviously the umpire. The catchers use their catching skills, pitchers use their ability to place a pitch, batters use body language, and umpires use their judgment.

Aside from the four individuals, other variables also have an effect on the call. In order to isolate the amount an individual contributes it must be determined what effect these other variables have on the call. Seven variables are looked at in this study: pitch location, batter handedness, pitcher handedness, ball/strike count, pitch type, command of the pitcher or how close he is to hitting the catcher's target, and if the pitcher is playing for the home or road team.

The seven variables were tested to determine which should be included within the model. Pitch location is the basis by which a ball or strike is called and therefore was the first variable included. Count and command showed the most significant variation, and consequently were also included. Command demonstrated a greater variation with horizontal distances than vertical distances in relation to where the catcher's mitt was set so pitches were grouped by their horizontal distance from the catcher's mitt. Batter handedness also shows a significant impact so it was the fourth variable added. The other variables had an insignificant impact and therefore were not included.

After each pitch was categorized by location, the other three variables were included in order to determine the percent likelihood that each pitch was called a strike. These strike percentages represent the Strike Zone Plus/Minus.

In order to maintain a balance between the requirement for both current data and legitimate sample size, a rolling four-year basis is used, meaning that the Strike Zone Plus/Minus will always be based on the pitch results going back four years from the current date. The Strike Zone Plus/Minus assigns a plus if a pitch is called a strike and a negative if the pitch is called a ball. From there it is determined how large or small the positive or negative value should be, which is based on how likely a pitch would be called a strike.
The process of categorizing an individual's plus/minus is run through twice. The first time through it is simply determined if a player has a positive or negative value. The second time through determines what percentage is due to the individual player and which is due to the environment. Any environmental percentage is evenly divided among the four individuals and added on to their individual contribution. The second process is run multiple times until the change for each player becomes minimal.

Coaches can use this information to determine which pitchers, catchers, and batters have the greatest skill in influencing a call and analyze those players to isolate the skills used. These skills can then be taught to other players in order to increase their ability in this area. Coaches and general managers can use this model to determine the effectiveness of their players, especially catchers, in getting called balls rather than strikes. This information can then be used during salary negotiations and when looking at making trades. As not all teams place an importance on players with this skill, it might be possible for teams to pick up players with this skill at a lower cost.

An important note to remember is that while a pitcher wants a positive plus/minus rating, batters want a negative plus/minus rating.

Analytics methods used in this research: Strike Zone Plus/Minus
Subjects and Variables

Subjects and Variables in Sports Analytics

Analytic methods organize and analyze data in order to look for patterns that help in the decision making process. Statistical models are capable of sifting through vast quantities of information in a very short period of time. This is extremely useful in the world of sports as the amount of data collected by the various leagues continues to grow each year.

When we talk about data, we are referring to subjects and variables. Data is collected on a variety of subjects. Subjects can be individual players, teams, games, seasons, plays, coaches, etc. The same subject can be used in different ways depending on exactly what analysis is required. If we are comparing pitchers, we could compare Justin Verlander and Clayton Kershaw regarding their runs allowed per game in a season. In this case, both Verlander and Kershaw are the subjects. Alternatively, we could compare Justin Verlander in the 2016-2017 season to Justin Verlander in the 2017-2018 season. In this case, we are using Verlander as two different subjects in the same analysis.

A variable is the specific characteristic of the subjects that is being measured. If we look at National Basketball League players for a season some possible variables would be points scored, shots taken, and games played. The list goes on and on. After deciding on the subject, it is necessary to decide on which variable or variables are going to be included in the analysis. From there the type of analysis needs to be chosen, which is dependent on the properties of the variables and the patterns being studied.

Variables can be either qualitative or quantitative. Qualitative variables are those that do not have a number associated with them. If you are comparing left-handed shooters to right-handed shooters, you are looking at qualitative or categorical variables. Data analysis is more limited with this type of variable. Quantitative variables, on the other hand, do have numbers associated with them. Examples are number of shots made, number of goals or assists. Data analysis is infinitely diverse when dealing with numerical data.

The infinitely wide range of variables and subjects allows analysts to continually come up with new ideas for areas to study. The research never ends, with new ideas being presented all the time. Analysts are always looking for new ideas and patterns that can help their team gain an advantage over the competition.

Teams can study any combination of subjects and variables. When they are investigating possible trades or draft picks, they can examine a broad range of subjects and variables to narrow down their choices. Then they can decide to examine a few very specific variables to help in the decision between their top choices. This gives coaches the flexibility to look at the overall player or at one very specific aspect of the player. This idea is also applicable when focusing on the entire team. They can look at how the team is working together as a whole or look at only the offensive players or narrow it down even further to only offensive linemen. This allows coaches to develop strategies individualized for each player and, at the same time, strategies for the team as a whole. Thankfully, statistical models are able to take in large amounts of data and analyze the information in a very short period of time. This was simply not feasible before computers became an everyday object in our society.
Supervised and Unsupervised Learning

Supervised and Unsupervised Learning as a Method of Sports Analytics

Within machine learning, there are two main types of tasks: supervised and unsupervised. This chapter explains what supervised learning is and how it relates to unsupervised learning.

Coaches and sports analysts use the data mining process of supervised and unsupervised learning techniques to group players or athletes in a team or a team in a sport with specific goals, such as finding structure in sports training data. With this method, coaches hope to design better structured training sessions and game plans, which will eventually increase the performance of the team or the athlete under tutelage.

Supervised Learning as a Method of Sport Analytics

Supervised learning is an algorithm that is used to analyze the training data and provides a function that can then be used for mapping out new examples other than the initial input. This function reveals the scenario with the best result that can be obtained from an unseen event. Let’s look at the following example:

Suppose you’re a coach that trains different players or athletes. You can group your players according to a particular theme or category such as height range, position, power, and so on. You can then teach the machine using the factors of categorizing based on their specifications. These factors become the data you feed the machine with. In this case, because of your presence to instruct the machine, it is called “supervised learning”.

Now, the machine learns from those factors you input. Each time you specify the characteristics of another player, the machine immediately categorizes the player. This may be done for several purposes including enhancing the training drills you give your players.

Suppose you instruct the machine to keep the data according to drills. The performance of each player across these drills can be averaged and stored. This becomes the data set of which the amount reflects in the number of observations. For instance, if the average drills stored gives a data set of 10, the expected number of observations should correspond to each of the players. These observations can help a coach determine which drills work best for which players.

There are many supervised learning procedures such as linear regression, decision trees, logistic regression, random forest, etc.

Unsupervised Learning

Now, remember that in the case of supervised learning, the coach or analyst is there to instruct the machine with certain factors (or classifications) which are converted into data. But under the unsupervised learning technique, there are no classifications.

In unsupervised learning methods, the coach or sports analyst trains the machine by inputting information, which is not categorized or classified. In this case, the machine algorithm acts on its own by
searching for similarities, patterns, differences without any prior knowledge of the players, athletes or teams.

Coaches and sports analysts mostly use unsupervised learning techniques to seek out that which is hidden in the structure of drills or performances of the team, players or athletes.

The two basic methods the machine uses under unsupervised learning for sport analytics is clustering and association.

Conclusively, supervised and unsupervised learning methods of sports analytics are advantageous in that they do not require much brainstorming from the coaches or analysts. And yet they give optimum results, which assists coaches in analyzing player performance.
Support Vector Machine

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.


A Team Recommendation System and Outcome Prediction for the Game of Cricket

This is a review of the cricket research conducted by Sandesh Bananki Jayanth, Akas Anthony, Gududuru Abhilasha, Noorni Shaik and Gowri Srinivasa applying support vector machines.

In Asian countries cricket is extremely popular with a variety of series taking place across several countries and a World Cup every four years. For each tournament a team's coaches and captain choose 11 players and 4 extras to participate. These players are selected from a pool of players dependent on the type of tournament being played as well as the opponents. Players are chosen based on performance measures and player statistics and is highly reliant on the experience and analytic skills of the coaches and captain. This often results in biased choices rather than choices that will ultimately result in the highest performing team. A more effective process would be to consider which group of players has the greatest chance of winning.

The model first consists of processing and storing raw match data. To do this information from the 2011 cricket World Cup and scorecards from Howstat.com was entered into the database. Aggregate data is compiled regarding performance statistics for the entire tournament including performance measures of the batsmen, bowlers, and overall tournament.

This information is then used quantify and rank the players. Batsmen are ranked based on calculations of batting average, batting strike rate, milestone reaching ability, out rate, and boundary runs per innings. Next, statistics regarding the overall tournament are determined including batting general average, batting general out rate, and batting general strike rate. Both sets of data are used to find the above generic average runs (AGR) for each player. Finally, the ranking index for each batsman is determined using the AGR and team generic batting average. A similar process is used to rank the bowlers with their ranking based on number of runs conceded, number of wickets taken and number of overs bowled.

The analysis of predicting wins concerning the team chosen is conducted using the previous information and a support vector machine (SVM) which separates the samples into positive (wins) and negative (losses). Coaches and captains are able to input different possible variations of players to predict
whether they would win or lose when playing a specific opponent. The SVM also creates projections regarding strengths and weaknesses of the proposed team.

Finally, clustering and k-nearest neighbor methods are employed to determine the preferential role for each player using the player preferred role recommendation system, which is able to suggest players who had not played in the tournament. In order to do this the statistics and performance measures determined previously are clustered using k-means to determine similarities. From this, five players similar to each player who participated in the tournament are found and entered into the database.

This process provides coaches and analysts alike the ability to determine which combination of players has the greatest chance of winning against a specific opponent. It allows for experimentation with any variety of players. It would be useful when looking at making trades or drafting players. Possible recruits could be included in the SVM in order to determine their effectiveness in increasing the team's probability of winning any given match.

A model such as this takes the bias out of the player selection process, leading to a statistically greater probability of winning.

Analytics methods used in this research: Performance Measures, Support Vector Machine, Clustering, and K-Nearest Neighbor
Survival Analysis Model

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems. This topic is called reliability theory or reliability analysis in engineering, duration analysis or duration modelling in economics, and event history analysis in sociology. Survival analysis attempts to answer questions such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?


In this chapter, we take a look at how survival analysis can be applied to examining the ‘lifetime’ of intercollegiate athletic donors.

New Applications of Survival Analysis Modeling: Examining Intercollegiate Athletic Donor Relationship Dissolution

This is a review of the survival analysis research conducted by Liz Wanless, Jonathan A. Jensen, and Parker Poliakoff.

Retaining customers is less expensive than gaining new ones. Colleges rely on intercollegiate athletic donors to help combat declining football attendance and increasing expenses. This puts immense pressure on university athletic development offices to maximize, retain, and upgrade donations.

To date, research has used a philanthropic approach, appealing to the donor’s motivations. This approach is not effective in determining the best methods for gaining and retaining donors as the research is based on self-reports from donors, depends on donor participation rates, and does not take into account the relationship between the donor and the athletic department.

In order to deal with these limitations a survival analysis model is applied to 10 years of data regarding donor behavior for a medium sized NCAA Division I Football Bowl Subdivision institution. This does not rely on self-reports by the donors but analyzes past data regarding all donors. It also looks beyond the psychological, sociological, and demographic information to the relationship between the donor and athletic department, specifically the number of times the department contacts the donor. To better understand when a donor is likely to stop their donations, variables related to economic, demographic, marketing and athletic success dimensions are analyzed in order to try and determine what factors impact the length of a relationship between donor and department.

Survival analysis is used as it capable of taking into account staggered entries and censored observations and unknown duration times of events. It is also capable of dealing with the dual nature of the variable, in this case whether the event occurred and the duration of the event. Using survival analysis also provides the ability to look for any variables that change values over time. Further, it is able to look at
how the relationship between the donor and university affects the length of the relationship. The probability that a donor will end the relationship and underlying factors are calculated.

Data from close to 3000 donors was collected including variables related to the donor's relationship with the athletic department, donation history, and location. These variables were organized into four groups or factors: economic, demographic, marketing, and athletic success. Results indicated that donors are typically lost within the first two years of the relationship. In fact, the longer the relationship the less likely the donor is to leave.

After analyzing the output from the survival analysis, a Cox proportional hazards model was utilized in order to gain a clearer understanding of how the various characteristics related to the donors and athletic departments affect the probability of the relationship discontinuing. Results indicate that donors living in the same state as the university are more likely to remain donors, as are those who are retired.

Athletic departments can use this information in order to determine the optimal marketing strategies for retaining donors. As donors are most likely to leave within the first two years more emphasis should be put into strengthening the relationship with these new donors in order to increase the probability of retaining their relationship with the team.

Analytics methods used in this research: Survival Analysis Model, Cox Proportional Hazards Model
Team Strategy Model Framework
Invasion Team Sports: Strategy and Match Modeling

This is a review of the team strategy model and match model research conducted by Leonardo Lamas, Junior Barrera, Guilherme Otranto and Carlos Ugrinowitsch.

Invasion games refer to team games that require one team to invade another team's defense to score. In order for these teams to be successful, the players must work together as a whole. Team strategies are more complex than individual strategies and lead to a greater chance of winning.

A model is created that looks at both strategic and tactical alternatives. Strategic alternatives are those when a player's actions are in agreement with the team's strategies, in other words the player is working with his teammates. Tactical alternatives are those when a player does not follow the team's strategy, but is relying on their own experience. In both cases the player is making the decisions, he is just handling the information he sees unravelling in the game differently.

One important element of the model is the action rule. It determines the logical choices a player has in any given situation. Action rules are divided into two categories - high-level and low-level rules. High-level rules consist of a motor skill and low-level rules consist of the individual components of that skill. Therefore, high-level rules include a sequence of low-level rules. When a player is following team strategy, they are using strategic action rule; when they are relying on their past experiences, they are using tactical action rules. In the midst of playing a game, a player must respond to the opposition and often they realize that the team's strategy will not work in that situation and therefore they will then rely on their own experiences.

A team strategy model framework is constructed. This model looks at whether an action is strategic or tactical, the position of the players, and the intensity of the action. These actions are looked at from both the offensive and defensive perspective. The model then analyzes each play - it looks at how a player's decision affects how his teammates react. If it is a strategic move then his teammates should move according to the team strategy. The model is then able to lay out team strategy on a map of the playing field.

A match model can now be created that looks at how two teams would interact during a game. Analysts can use this to determine which team has a greater chance of winning a game. They can also look at the probability of how individual players will perform against their opponent. Coaches can use this tool to refine their team strategies in order to react more efficiently to their opponents. Coaches can model possible outcomes related to a player's decisions, teaching their team how to make better decisions within the circumstances. By following a map of an individual play, players will be able to see how their actions affect their teammates giving them the opportunity to determine how to maximize the efficiency of their own actions in order to maximize the team's potential of winning the game.

Analytics methods used in this research: team strategy model framework, match model
Text and Sentiment Analysis

Text and Sentiment Analysis as a Method of Sports Analytics

Sentiment analysis is a relatively new technique which is now being used in sports analytics. This chapter discusses what text and sentiment analysis entails and how it is applied in sports.

What is sentiment analysis?

When we talk about Sentiment Analysis (also called Mining of Opinions or Emotional Artificial Intelligence), we are referring to a series of applications of natural language processing techniques, computational linguistics and text mining, which aim to extract subjective information from generated content by users such as comments on blogs, social media, etc.

With this type of technology, we can extract a tangible and direct value, such as "positive"/"negative", from a textual comment.

Depth of analysis

To speak of Mining of Opinions is to speak of an increasingly extensive field related to the analysis of the subjective components that are implicit in the contents generated by the users. Within this field, there are applications that perform a more or less profound analysis of the textual content, depending on the task or problem that you want to solve. In general, we find two types of tasks related to Mining Opinions:

Polarity detection: This refers to being able to determine if an opinion is positive or negative. Beyond a basic polarity, you can also obtain a numerical value within a certain range, which in a certain way tries to obtain an objective rating associated with a certain opinion.

Analysis of sentiment based on characteristics: This refers to determining the different characteristics of the product, personality or team treated in the opinion or review written by the user, and for each of those characteristics mentioned in the opinion, be able to extract a polarity. This type of approach is much more complex and much finer than the detection of polarity.

Relaying it to sports

Relating sentiment or emotional AI to sports, for example, we live in the era of the "experts" of football, of deep analysis and of special guests who are ex-technicians and ex-football players who manage and weave countless possibilities. This generation of wise men analyze corner shots, make sketches on blackboards and between questions and answers assemble alignments and predict results. It is the intellectual age of football that has caused a notable influence on managers, coaches and footballers. Nobody escapes excessive criticism.

From the show to the commercialization

For those who have the opportunity to see soccer players and national teams, such as Brazil and Germany, get an exciting match every tie. In the past, journalistic specials were unavoidable without the topic of "Pelé" being touched upon. After the Argentina World Cup 78, soccer as a show and as a business began to take commercial roots. From the quality and brilliance of the Brasileirao, we passed to
the skill and "malice" of the Argentineans without stopping recognizing the birth of new stars such as Diego Armando Maradona that led to the second albiceleste title in Mexico 86. There began sentiments.

The soccer world relives the golden stage of a rivalry created at the beginning of sports analysis, between Argentines and Brazilians. To be more concrete, between Pelé and Maradona. Who's the best? This first controversy flooded the South American soccer world of an absurd rivalry that was well-taken advantage of by the merchants of the best spectacle of the world. From there, fierce rivalry and sentiments took over the football world.
**Third-Order Polynomial Regression**

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable $x$ and the dependent variable $y$ is modelled as an $n$th degree polynomial in $x$.

The model is simply a general linear regression model with $k$ predictors raised to the power of $i$ where $i=1$ to $k$. A second order ($k=2$) polynomial forms a quadratic expression (parabolic curve), a third order ($k=3$) polynomial forms a cubic expression and a fourth order ($k=4$) polynomial forms a quartic expression.


**Assessing the Offensive Productivity of NHL Players using In-Game Win Probabilities**

This is a review of the NHL research conducted by Stephen Pettigrew, applying third order polynomial regression.

Hockey statistics have come a long way over the years, yet they still lag behind other sports like baseball and football. To close this lag two new statistics are created, In-Game Win Probability and Added Goal Value. The Game Win Probability provides second-by-second probabilities that a team will win the game. It takes into account the current score, penalty situation, and home-ice advantage. Added Goal Value looks at the contribution individual players make to the team's probability of winning a game.

An algorithm is developed to deal with penalty situations, which looks at remaining time based on unexpired penalty time, goals that may erase penalty time, and league rules that dictate how penalties are handled. From there, win probabilities for each goal differential are determined. These probabilities are regressed with a third order polynomial for the time remaining in the game. The resulting coefficients are used to calculate expected win probabilities at each point in regulation time. The probabilities of short-handed and power play goals are determined using a Poisson distribution.

Not all goals have the same impact on a game or the same importance level. A goal that breaks a tie is more valuable than a goal that puts a team up 8-2. To deal with this the Added Goal Value statistic is created. Added Goal Value looks at the impact each goal has on the Game Win Probability to determine its value.

This information provides analysts with information regarding the importance of goals and killing penalties and the impact they have on win probabilities. It also provides analysts with a tool to watch for patterns throughout a game in order to determine what actions have the greatest impact on the ongoing win probability. From there player skills can be evaluated as to which add the most benefit to the win probability. It can also be used to predict who will win a playoff series by looking at the in-game win probabilities for each game.

Players can be compared and contrasted based on their Added Goal Value, especially useful when teams are deciding between two similar players in a trade situation. Players who score approximately the same number of goals in a season can be ranked based on the goal added value. This statistic is especially
useful when looking at young players. Those with a higher goal added value will add greater value to a
team and should be looked at closely in order to add more depth to a team.

A player's Added Goal Value tends to remain fairly constant across seasons. This makes it an extremely
useful statistic when making trades. Teams can be fairly assured that what a player contributes to their
current team will be carried over to a new team.

It must be noted that the Added Goal Value is most useful when comparing players with similar goal
scoring ability.

Analytics methods used in this research: In-Game Win Probability, Added Goal Value, Third-Order
Polynomial Regression, Poisson Distribution
Three-Dimensional Markov Model

In probability theory, a Markov model is a stochastic model used to model randomly changing systems. It is assumed that future states depend only on the current state, not on the events that occurred before. Generally, this assumption enables reasoning and computation with the model that would otherwise be intractable.

https://www.agilesportsanalytics.com/three-dimensional-markov-model/

In sports, three-dimensional Markov models can be used for behavioral recognition to categorize player actions in various scenarios and also for predicting win probabilities during different game time situations.

An Analysis of Curling Using a Three-Dimensional Markov Model

This is a review of the curling research conducted by Paul Brenzel, William Shock, and Harvey Yang, applying a three-dimensional Markov model.

Curling enjoys great popularity in Canada and is on the rise in the United States. This study models curling as a Markov process to estimate win probabilities of different states during a curling match.

In this research, score information was entered into the model from the matches played from 1998 to 2014 in the Canadian Men's Curling Championships, including the year of the match, round of the tournament, match location, teams competing, score in each end, final score, time remaining for each team, and which team started with the hammer in the first end.

In order to use the Markov method all possible scenarios that can occur during a game must be defined. All possible state transitions and their associated probabilities must also be known. This data is used to create a three-dimensional Markov model using three states - the end being played, hammer state, and score differential. The purpose of the three-dimensional Markov model is to determine the expected win probability for any team based on the current state of the game and taking into account all possible future transition states and their associated probabilities.

Two models were created, a homogeneous one that assumed that state transitions were strictly a function of hammer possession and independent of any other parameters and a heterogeneous model, which assumed state transitions were dependent on other parameters. The results from both models were very similar with the exception of increased accuracy of the heterogeneous model in predictions of state transition towards the end of the game, especially the tenth end. Strategy at that point of the game is typically different from that employed earlier in the game. A team that is down by two points will choose very different strategies from a team that is up one point.

Teams can use the information to determine when to score one point and give up the hammer or blank the end to maintain possession of the hammer. The Markov process takes into account not only the
team's probability of scoring a point but also the effects of lost opportunities. The model will also aide teams in deciding when they should concede a game. The analysis actually determines that teams should concede less often than they currently do. Current decisions appear to be at least partially based on psychological conditions rather than statistical analysis.

Analysts can use the Markov model to graph the expected win probabilities for each team over the course of a game. This provides the ability to relay information to viewers in a clearly understandable manner. It also gives them opportunities to more deeply delve into how effective a team's choices are, or if alternative choices should be made in any given situation.

This Markov model provides the ability to analyze the game of curling at a deeper level than was previously possible.

Analytics methods used in this research: Three-Dimensional Markov Model
Transformations

Functions which map points of a pre-image onto its image is called transformation. The dimensions of three-dimensional figures are length, width, and height.


Using Transformations to Improve Measure of Team and Player Performance

Some measurements can be difficult to work with and understand. In these cases, it can be useful to transform them to another function. Linear or logarithm transformations are two possible ways of accomplishing this. The transformation does not change the information, but rather displays it in a way that makes it easier to analyze and interpret the results.

Linear functions have limited uses. However, they are useful in changing the range of the data to one that it easier to work with and understand. One instance would be if you were looking at the number of short-handed goals scored by a team during one season. Suppose a team scores 238 goals in a season with 12 of them being short-handed goals. Dividing 238 by 12 gives us the proportion of 0.0504. Small numbers like this are trickier to work with, making accurate analysis more difficult. However, if a linear relation is used to transform the data to a percentage the number becomes 5.04, which is definitely easier to understand and compare to other teams' numbers of short-handed goals.

A logarithm, more commonly known as logs, is another frequently used transformation. Logs are used when you want to look at the proportional differences between the variables. This is often useful when looking at the earnings of players. If Player A is making 8.4 million, player B is making 5.4 million and player C is making 2.4 million it would appear that player B's performance is halfway between that of players A and C as their earnings are all 3 million dollars apart. However, if you want to look at the difference in proportion between the players it would be better to use a log transformation. The result would be 15.94 for player A, 15.50 for player B and 14.69 for player C. It is now easier to see that player B's performance is actually closer to that of player A than to player C. One caution to note when using log-transformed values; as most people are not familiar with log data it is often interpreted incorrectly. If this were the case, it would be more appropriate to use the original data.

In order to determine which type of transformation may be appropriate for a set of variables you need to look at the distribution of the transformed data. If the resulting data forms close to a normal distribution, in other words a bell curve, the transformation is a useful tool. If the resulting histogram is skewed with the peak towards the left or right side of the graph, then the transformation is not appropriate for use with this data.

Using a linear, log or other type of transformation allows analysts and coaches to easily detect patterns within a data set. Predictions will be more accurate, which would result in better responses by the teams involved.
In this chapter, we will look at another method of analytics that is widely used in sports—and that is uplift and persuasion modeling. Uplift and persuasion modeling is an aspect of data science which uses data aggregation to create predictive models for performance analysis.

Sports analysts build predictive models, and test and refine them based on factors such as demographics, geography, leagues, teams, player skills, and so forth in an effort to perform statistical analysis, which help identify the characteristics of the data in other to draw certain inferences and decisions.

So, what is uplift and persuasion modeling? Uplift modeling deals with the increase in likelihood of the outcomes of an event with the treatment when compared to the outcomes of an event without the treatment. That is why uplift modeling is also called “treatment effects modeling”. There is no way you can tell the different between treated and untreated events in a direct way or using a direct technique, but rather by inferring answers from an experiment.

Now, let’s briefly discuss persuasion modeling. There is only a slight difference between uplift modeling and persuasion modeling. Suppose you have business products you want to sell, so you create and launch advertisements. Once you’ve advertised your product, some consumers have already decided to buy from you, while some others have already made up their minds not to buy. By focusing future advertising on those who have decided not to buy, you are likely to anger these consumers and not get a return on your investment.

Then there is the set of potential customers. We’ll call them potential, because they don’t know if they want to buy your product or not. This group of people need to be convinced. But the challenge is isolating this group of potential customers, while avoiding the first two groups who have made firm decisions. This is where persuasion modeling is effective.

How is this applicable to sport analytics? It’s mostly applicable to the business side of sports. It helps club owners and managers make persuasive decisions about the business that surrounds their sport.

Now, after the persuasion modeling has been used to identify targets. The next question is how do you turn this identified group into customers? This is achieved by uplift modeling. Uplift, simply put, means to raise something higher. That means you want to increase your chances of winning those potential customers to your side.

There are two main areas of uplift modeling. They are (1) Predictive models, and (2) Prescriptive models.

While predictive models focus on the use of statistical tools and models to provide insights into future events with the aim of making predictions, prescriptive models employ the use of algorithms to optimize and simulate data and/or events with the goal of giving possible outcomes of an event.

In conclusion, coaches and analysts utilize uplift and persuasion modeling to analyze subjects, categorize them to determine their profitable aspects, and then treat those aspects with the aim of optimizing the performance of their team.
Variables

A variable is any characteristic, number, or quantity that can be measured or counted. In sports, variables can range anywhere from teams, players, height, age and weather to possessions, plays and in-game decisions.

A Starting Point for Analyzing Basketball Statistics

This is a review of the basketball sports analytics research by Justin Kubatko, Dean Oliver, Kevin Pelton, and Dan T. Rosenbaum

Within standard basketball statistics there are a variety of basic variables used. Understanding these variables helps everyone, the knowledgeable and the amateur, better understand the meaning behind the statistics.

The first variable is possession. In basketball a possession starts when a team gains control of the ball and ends when they lose control of the ball, either by scoring or losing possession of the ball to the other team. Analysts look at the number of possessions each team has during the course of the game. Typically, both teams have a fairly equal number of possessions throughout a game.

Possessions are used to evaluate the efficiency of teams and players by looking at the number of points scored per possession. Points scored per 100 possessions is the offensive rating and points allowed per 100 possessions is the defensive rating. A higher offensive rating and lower defensive rating is how teams win games. Offensive and defensive ratings are not related to each other. Teams with high offensive ratings are not necessarily more likely to have a better defensive rating.

Another variable is plays. Plays are similar to possessions except that now rebounds are taken into account. A team can shoot, miss, and rebound the shot many times without losing possession of the ball. The result is multiple plays within a single possession. So, while possessions are fairly equal between teams during a game, the number of plays is not.

When statistics are calculated on a per-minute basis they tend to be consistent across players even if those players play different amounts of time. These statistics allow players at different levels to be compared with each other.

There are three types of field goal percentages. Field goal percentage (FG%) does not include points gained by three pointers or free throws. Effective field goal percentage (eFG%) includes three pointers and true shooting percentage (TS%) includes both free throws and three pointers.

Rebound rate is determined by looking at the number of shots a player rebounds while they are playing. This can divided into offensive rebounding percentage and defensive rebounding percentage. These are useful tools as players are typically not equally skillful in both offense and defense.

Plus/minus statistics are determined by subtracting defensive points from offensive points while that particular player is in the game. They are determined on a per-minute or per-possession basis. The net
plus/minus is found by taking the plus/minus statistic for a given player and subtracting the plus/minus statistic for the team when that player is not in the game. Adjusted plus/minus statistics take into account the skill levels of the player's teammates and their opponents.

Individual possession rate looks at how many field goal attempts, free throw attempts, assists and turnovers a player has per possession.

Linear weights involve putting weight or importance on different statistics. Linear weights tend to be subjective and consequently are typically not as reliable as other statistics.

These are just a few of the basic statistical ideas used in basketball. Further analysis can be done using the Pythagorean winning percentage which is based on the idea that team winning percentages, points scored, and points allowed are all closely related or the bell curve method which is based on the idea that points scored for and against are normally distributed.

Analysts, teams and fans alike can use these statistics to compare players and teams, determining strengths and weaknesses.

Analytics methods used in this research: variables, field goal percentages, plus/minus statistics
Variation

Measuring the Variation in Sports Data
Variation is the name of the game in sports. No two players are the same, no two teams are the same, how a team plays in one game is different from they perform in another. Understanding these differences and the cause behind them is a main reason for sports statistics. Analysts and coaches are always looking for reasons why their team does not fare well against another particular team or why a player is faltering. They want to determine the best strategies to improve their team and probability of defeating their opponents.

On method used to analyze variation is setting a standard value and comparing all other values to that standard. Using the median is one way to do this. The median is the middle number in the data and others can be compared as being higher or lower than the median. The mean is another tool for measuring variation. The mean number of goals per game is determined for the league and then the team's goals per game is compared to that average.

Standard deviation is another method used to analyze variation. Standard deviation is the measure of how spread out the numbers in the data set are. A low standard deviation indicates the numbers are all clustered close to the average. A high standard deviation indicates the numbers are spread out further from the mean, indicating a larger variance in the data. Standard deviations cannot be compared directly with each other, as they are dependent on the variables being used. Football games typically have a wider variety of final scores than do hockey games. Therefore, the standard deviation of points in a football league would be larger. This does not give an accurate representation of the individual situation. One way to deal with this dependence is using the coefficient of variation where the standard deviation is divided by the mean of the data. This eliminates the dependency; however, the coefficient of variation is only used for values that are greater than zero.

Another measure of variation is the range. The range is simply the highest value minus the lowest value. However, the range is not always useful as outliers have a large influence on the outcome. Outliers are results (either extremely high or extremely low) that rarely happen. They inflate the range, making it appear as though the variation is much larger on a regular basis than it actually is. One method to eliminate the outlier effect on the range is to use the interquartile range. The interquartile range does not include the bottom 25% of the results nor the top 25% of the results. This allows it to present a more accurate picture.

All of these measurements have their uses within sports statistics. Analysts and coaches alike use them when comparing teams and players. It aides in determining the consistency of teams and players from game to game, and even quarter to quarter.

Comparing Between-Team and Within-Team Variation
Variation can be determined using different sources of data. Three typical sources that are often compared are overall variation, between-team variation and within-team variation.
You can use all three types of variation when analyzing points scored in the National Football League. Overall variation is determined using all of the points scored by all of the teams during one season. This overall variation is comprised of both between-team and within-team variations. Each team has its own unique make-up of players and scoring ability, which creates a variation between teams. However, teams themselves are not always consistent. They play better in one game than they do in another, which can be due to either their own offensive ability or the defensive ability of their opponent, or, most likely, a combination of the two. This is the within-team variation.

When the standard deviations of all three measurements are determined for a single season it is evident that the three measurements are related. Adding together the between-team standard deviation and the within-team standard deviation equals approximately the overall standard deviation.

With these three stats, the effect of the between-team and within-team variations on the overall variation can be analyzed. This allows analysts to focus on whether the difference in scoring ability is more predominant between different teams or between different games played by the same team. This gives the ability to discuss the consistency across the league as well as the consistency of each team. Are the rules of the league fair to all teams or does there seem to be a bias towards certain types of teams? If there is a bias, how can this be alleviated, either by the league or by coaches reconfiguring their teams. On the other hand, is the variance based on the player talent on each team, coaching talent, team cohesiveness, or any other of a myriad of factors? Looking at these numbers helps analysts determine which trades or draft picks might best be able to help improve the consistency of the team in their scoring ability from game to game.

Coaches can determine whether their standings in the league are caused more by between-team or within-team differences. Do they need to improve the overall ability of their team in order to compete more effectively with the competition? Alternatively, is the problem not with the players they currently have, but with the consistency of the team? Does their own inconsistency have a larger effect on the standings than how they compare to other teams? If so, decisions must be made regarding how to improve the team's consistency, as well as the consistency of the individual players. Is it due to a lack of training, difficulty carrying out the plays or some other factor? Strategies can then be put into place to help alleviate these issues.
Weighted Least Squares Regression Model

Weighted least squares (WLS), also known as weighted linear regression, is a generalization of ordinary least squares and linear regression in which the errors covariance matrix is allowed to be different to an identity matrix. WLS is also a specialization of generalized least squares in which the above matrix is diagonal.

Source - https://en.wikipedia.org/wiki/Weighted_least_squares

An Improved Adjusted Plus-Minus Statistic for NHL Players

This is a review of the Weighted Least Squares Regression Model research with NHL data conducted by Brian Macdonald.

NBA analysts and teams use the adjusted plus-minus (APM) stat to determine players' contributions to the offense and defense. One strength of the APM is that each player's score is not dependent on his teammates' scores which is an improvement over the traditional plus/minus stat in which players' scores were dependent on their teammates. This APM model can be adjusted for use in the NHL. The issue becomes that hockey games are not always played at even strength. Teams can be down one, or even two players during a penalty. Including these situations in the APM would be unfair to those players on the ice as it would lower their ratings, in some cases quite drastically.

So, a new model needed to be developed. In fact, two models were necessary - one for special team situations and one for even strength situations.

The model for even strength situations includes both an offensive and defensive statistic with the exception of the goalie who will have only a defensive statistic. The information will be inputted on a shift by shift basis with a shift being a period of time when no player substitutions are made. The offensive stat is based on goals scored by the team and the defensive stat is based on goals scored against the team. Then the model is adjusted to take into account which zone the faceoff takes place in.

The model for special team situations will account for both power play and shorthanded situations. Therefore, four stats will be needed - goals scored for and against during power plays, and goals scored for and against while shorthanded.

To get the total APM the scores from the two models are added together.

Using this model would give a more accurate picture of the contributions each player makes to their team. Analysts can compare players across the league in terms of total plus/minus, offensive plus/minus, defensive plus/minus, even strength plus/minus and special team plus/minus. This will allow analysts to gain a more rounded picture of how players compare with others in the league.

Teams can look at the individual stats when they are looking at trading players. If they know that their team is weak in one area, they will be able to look at the stats of the available players to determine
which player would be most likely to improve their team in that area. It would also help coaches work out training plans for improvement in these specific areas.

Analytics methods used in this research: adjusted plus/minus, weighted least squares regression model
Weighted Plus/Minus

Weighted plus/minus statistic provides coaches and analysts with a quantitative evaluation of each player's performance allowing for a comparison between players.

A Weighted Plus/Minus Metric for Individual Soccer Player Performance

This is a review of the weighted plus/minus soccer metric research conducted by Steven R. Schultze and Christian-Mathias Wellbrock.

While soccer is the most popular sport in the world, its' analytics have not reached the same level as those for other sports. However, the field is continuing to grow. A plus/minus metric has traditionally been considered too unwieldy and difficult for soccer due to the complex nature of the game. The plus/minus metric discussed here would be used to evaluate player performance and show whether a player's presence on the field has a positive or negative effect. It also allows for a comparison regarding which players provide the greatest positive influence on the team.

The weighted plus/minus metric proposed by this research actually extends the data provided by the statistic in other sports. It takes into account the importance of the goal - a goal that wins the game is significantly more valuable than a goal that puts a team up 4-0. It also takes account the strength of the opponents, as defeating a strong opponent should have a higher weighting than defeating a weaker opponent. Variables taken into account in the calculations are game number, opponent, and time, shot taker, shot result, and shot type. Each player's time spent on the field was standardized to 90 minutes. This helps level out the playing field and eliminates bias between players who play many minutes every game versus those who play fewer minutes per game.

This weighted plus/minus statistic provides coaches and analysts with a quantitative evaluation of each player's performance allowing for a comparison between players. It also provides coaches with insights in how training should be adapted for each individual player. During the scouting process, this metric would enable the comparison of players to help determine which player would be best suited for the team.

Another strength of the weighted plus/minus is its objectivity, allowing coaches to see which players are positively affecting team results and which are not. It also provides an objective basis for comparing players. This aides coaches when decisions need to be made regarding which players will play in an upcoming game given the strengths of the opponents. Coaches can use the weighted plus/minus metric to determine their opponent's strengths and weaknesses in order to formulate the best defensive and offensive strategies that optimizes their chances of winning.

The weighted plus/minus statistic provides coaches and analysts with a more efficient method for evaluating players' performance compared to traditional statistics and analyzing video.

Analytics methods used in this research: Weighted Plus/Minus
**Weighted Voronoi Diagram**

In mathematics, a weighted Voronoi diagram in n dimensions is a special case of a Voronoi diagram. The Voronoi cells in a weighted Voronoi diagram are defined in terms of a distance function. The distance function may specify the usual Euclidean distance, or may be some other, special distance function. Usually, the distance function is a function of the generator points' weights.


**NBA Court Realty**

This is a review of the NBA research conducted by Dan Cervone, Luke Bornn and Kirk Goldsberry, applying a weighted Voronoi diagram.

The regions on a basketball court can be looked at as if they are realty. How valuable is the realty that a player has control over? Which region does the player need to take control of next in order to help facilitate a chance of scoring? Do players always make the best decision when choosing which space to occupy?

A Weighted Voronoi diagram is built to determine which regions of the court are the most valuable and to quantify decisions made by the players regarding the space they occupy or want to occupy. The first step is to divide the court up into spaces that measure approximately 2' x 2'. The next step is to determine the value of each space. This is done by looking at which spaces seem to be preferred by the players. Spaces that contain many players are obviously considered more valuable than spaces occupied by a lone player. One way to evaluate the spaces is to analyze the passes between players. When a player passes the ball to a teammate he obviously feels that the space his teammate occupies is more valuable than the one he occupies himself. By looking at the passes each team makes it is possible to determine which spaces are valued the most by teams and any variation that exists, as not every team places the same value on each space.

The Weighted Voronoi diagram can be altered to look at the individual players and which spaces they consider to have the highest value.

Once each space on the court has been given a value, possession can be analyzed to determine if the players are maximizing that value. This allows possessions to be compared by their "value". Are players maximizing the value? If not, this is something that can be practiced in order to improve the players' decision making in future games.

The value of space occupied by the offense and defense can be compared. This would help evaluate which team is making the better choices. Coaches could analyze this data in order to improve the decisions their players make regarding which space they want to occupy. Coaches would be able to see which spaces are preferred by individual players and help determine if players are maximizing their space. Coaches could also look at which spaces their opponents favor. This would aide in determining the best defense strategy to limit the opponents' points.
Analytics methods used in this research: weighted Voronoi, logistic regression model
Wilcoxon Signed Ranks and Kruskal Wallis H Tests

Wilcoxon Signed Ranks Tests and Kruskal Wallis H Tests as a Method of Sports Analytics

This chapter discusses two similar sports analytics methods called the Kruskal Wallis H Test and the Wilcoxon Signed Rank Test.

In the application of Wilcoxon Signed Rank Test and Kruskal Wallis H Test, coaches and sport analysts employ tests to compare two or more independent samples that are of different or the same sizes, and two dependent samples, respectively. This comparison allows analysts to determine whether or not the samples originate from the same distribution or are normally distributed. Wilcoxon Signed Rank Test and Kruskal Wallis H Test are typical tests that are used in Grand slam tennis matches to determine the percentage of points won based on serve.

Wilcoxon Signed Rank Test

Wilcoxon Signed Rank Test is a non-parametric statistical test. In plain terms, the test is used to compare two samples that are related. And because it is non-parametric means that the population from which you draw your samples is not normally distributed.

Sport analysts employ this test when the differences between the sample data do not have a normal distribution. The analysts generate a null hypothesis for this test, and the null hypothesis is that the two samples to be used must have equal median.

Let’s see how tennis analysts use Wilcoxon Signed Rank test to analyze a grand slam tennis match. In a tennis study of 252 matches, it was discovered that players won a higher percentage of points in 15 matches on second serve, while a higher percentage of points was won by players in the other 237 matches on first serve. These results are then used with the median values to obtain the median percentage points. It is from this median percentage point that all other variables are obtained.

Kruskal Wallis H Test

Kruskal Wallis H Test is also a non-parametric statistical test. It is an alternative to the common one-way Analysis of Variance (ANOVA). This test is used when the assumptions that are normally used for ANOVA cannot be met and it is used to compare three or more independent samples. Sports analysts use this test to determine if the medians of two or more statistical groups are different.

Let’s see how Kruskal Wallis H Test is used for the same grand slam tennis matches mentioned above. Supposing a tennis analyst wants to carry out analysis of 252 matches in four different tennis tournaments, for instance, where the mean rally duration is compared, there are two important variables. They are (1) Test Variable List and (2) Grouping Variable. The length of the rally is used in the Test Variable List while the surface is used in the Grouping Variable. Now, in a tennis study of 252 matches, the expected mean rank should be about 126.5 provided there is no large difference between the surfaces.

This Kruskal Wallis H test results in a value that is used for a normal distribution. This value is called “H value” and the common normal distribution test used here is the “Chi-square distribution”. Now, from
the result given by the H value and Chi-square distribution, the analyst can tell if there is a significant difference between the groups. However, it won’t be revealed yet the exact group that is different from the others. It is then the analyst will now carry out an “H test” or a “Post Hoc” test.

It is this H test that reveals the group that is different from others causing the significant difference in the distribution.

Conclusively, Wilcoxon Signed Rank Test and Kruskal Wallis H Test help coaches and sports analysts to determine variation in performances of different tennis players.
Win Probability

Win probability is a statistical tool which suggests a sports team's chances of winning at any given point in a game, based on the performance of historical teams in the same situation. The art of estimating win probability involves choosing which pieces of context matter. In baseball for example, win probability estimates often include whether a team is home or away, the inning, number of outs, which bases are occupied, and the score difference. Because baseball proceeds batter by batter, each new batter introduces a discrete state. There are a limited number of possible states, and so baseball win probability tools usually have enough data to make an informed estimate.


The Problem with Win Probability

This is a review of the NBA win probability research conducted by Sujoy Ganguly and Nathan Frank.

In the NBA the win probability statistic is based on game time, possession, and point differential in order to predict who the winner of the game will be. There are two main issues with this statistic. One, it does not contain enough information such as injuries and fouls, which can have a major impact on the outcome of a game. Second, the chance of winning or losing is not an absolute - no team is guaranteed to win or lose - there is always an element of uncertainty involved which this model does not take into account. There are two possible methods to deal with these issues. The first is to include the lineup of players for the game in the model and the second is to include the score difference distribution - in other words the difference in the points scored by the two teams.

In order to build this model data from the 2002-2017 seasons were used. This data included over 8.7 million play-by-play events. Each event included game time, ball possession, and score difference. In order to improve this model information regarding the teams playing was added to the data set. This included lineup information including player identity, number of games played, plus/minus per game, and minutes played. Putting this together means that for each play-by-play event there are 352 features taken into account. Then, rather than predicting who will win the game, the score difference is predicted in the form of home team score to away team score.

When this model was tested it was found that it was 88% accurate while the previous model was only 75% accurate - a 13% improvement.

With this improved model analysts will be better able to predict the outcome of upcoming games. The predictions will change as lineups for the teams change, making it a more fluid and dynamic analysis that responds to the ever changing compositions of the teams. Predicting a final score difference rather than simply who will win gives the analysts the flexibility to explore different alternatives for that differential. Such as what if a certain player is injured or teams make a major trade?

As this model now takes into account the lineup of each team, teams will be able to look at the expected outcome of their games and test how the expected outcome will change with different players in the
lineup. This will allow them to be more proactive, rather than reactive in working to improve their team's performance.

Analytics methods used in this research: win probability, measure of uncertainty, datasets, encoding, explicit prediction, Gaussian distributions, mixture density network
Win Probability Framework

Estimating an NBA Player's Impact on His Team's Chances of Winning

This is a review of the research conducted by Sameer K. Deshpande and Shane T. Jensen, applying the win probability framework.

Fans are always arguing about which player contributes the most to their team. The argument arises in all sports, including the National Basketball Association. Traditionally, scoring statistics like point-per-game, shooting percentage, Adjusted Plus-Minus, player efficiency rating, or one of the other myriad of statistics are used to compare players. The problem with these statistics is that they do not take into account any context. As a result, player's statistics are inflated for points obtained in less than important conditions such as increasing a lead from 90-60 to 93-60, or are underinflated in key situations such as increasing a lead from 90-89 to 93-89 in the dying seconds of a game. Clearly, the points scored by the player in the first scenario should not be weighted as heavily as the points scored in the second scenario. To deal with this issue a win probability framework and linear regression model to estimate each player's contribution to his team's overall chance of winning games is developed.

The win probability framework is based on the win probability added statistic. Computing the win probability added starts with determining the team's probability of winning a game at each point in the game. The player's WPA is then determined as the change to the team's win probability from the time the player enters the game to the time he leaves the game and these changes from over the season are added together. The change in the team's win probability during a shift is regressed onto signed indicators corresponding to the ten players on the court to estimate each player's partial effect on his team's chances of winning.

Comparing players regarding their partial effect is complicated by the fact that players perform in different contexts. To overcome this, the total number of shifts each player plays is calculated, along with the team's average win probability at the start of his shifts, average duration of the shifts and average length of each shift. This creates the player's leverage profile. The Mahalanobis distance between the leverage profiles of each pair of players is calculated. This also helps ensure that player's statistics are not overly inflated if they are playing for a weak team or overly deflated if they are playing for a strong team.

A player's partial effect is an indication of his value to the team so players on each team can be ranked based on their partial effects. These player rankings are averaged over 1000 posterior samples to obtain their Impact Ranking.

In order to rank all players in the league their Impact Score is determined as the ratio between the posterior mean and the posterior standard deviation of a player's partial effect. The Impact Score is the balance between the uncertainty of his partial effect and his average partial effect.

These statistics allow coaches to determine if they are dividing playing time among their players in the most effective manner. Various five man lineups can be compared against each other, helping coaches put together the best combinations and also decide which lineups will likely be the most effective.
against the lineups of their opponent. They also help determine what effect an individual player has on the play of his team. This would be especially important when it is time to negotiate a contract.

On qualification regarding this statistic to remember is that as the statistic is based on the context in which the players perform it is retrospective in nature and will not be stable from year to year.

Analytics methods used in this research: Win Probability Framework, Linear Regression Model, Mahalanobis Distance, Impact Ranking, Impact Score
Win Probability Matrix

Using Cumulative Win Probabilities to Predict NCAA Basketball Performance

This is a review of the NCAA basketball performance research conducted by Mark Bashuk, applying a win probability matrix.

Basketball games are often judged on how they end. If our team scores a last second point to win, we feel it was an exciting game. If we were cheering for the other team, we feel the game was depressing. However, the final score does not tell the complete story. In order to combat this, a method is developed that uses cumulative win probabilities to quantify each game, combining a team’s average cumulative win probability with strength of schedule to rank teams and predict future game performance.

Rankings are calculated using a SQL stored procedure. The first step is creating a table, which combines play-by-play data and a Win Probability Matrix. The second step then combines the Cumulative Win Probability of each game with each team’s strength of schedule, which produces the team ranking statistic.

In order to determine the optimal value for each of the three variables used in the calculations, three simulations are run. The first simulation tests 100 combinations of each team’s Cumulative Win Probability and its Strength of Schedule. The second simulation looks at the idea of home court advantage by giving the home team an increased percentage and removing that percentage from the visiting team. The third simulation is designed to determine if the performance of a team at the end of a game is more predictive than their performance at the beginning of the game. Games are split into eights segments of three values. The results are charted in a histogram, which clearly indicate that the model is more accurate as later game segments are weighted more heavily than earlier segments.

In order to predict the margin of outcome in a game the rankings of the home team, rankings of the visiting team, and home court advantage are incorporated into the equation.

Analysts, when looking to make predictions regarding upcoming games or predicting which teams will make it to the playoffs, can use this information. This statistic judges team and player performance more accurately as it takes into account more than just the final score. This provides coaches and analysts with an improved ability to judge the ability of their players and the contribution they are making to the team. This will be useful when the time comes to look at making trades. Coaches can also help their players become better all-around athletes by looking at the skill sets of the top ranked players. As Strength of Schedule is incorporated, coaches and analysts can determine what effect this has on their team and their winning probability. Leagues can incorporate this information when designing the schedule for a new season, working to minimize the impact strength of schedule has on a team's performance.

While fans can use emotions to rank a game, analysts and coaches need to put the emotion aside and look at quantifiable aspects of the game in their determination of the ranking.
Analytics methods used in this research: SQL Stored Procedure, Win Probability Matrix, Cumulative Win Probability, Strength of Schedule, Simulations