

involve

a journal of mathematics

Nontraditional undergraduate research problems from sports
analytics and related fields

Carl R. Yerger



Nontraditional undergraduate research problems from sports analytics and related fields

Carl R. Yerger

(Communicated by Darren A. Narayan)

The purpose of this article is to encourage advisors to consider choosing a topic related to sports analytics for their next undergraduate research project. We discuss some of the advantages of working in problems related to sports analytics in an undergraduate research context. We also give a sense of the skills necessary to be successful in research, some ideas of what would make good problems, and avenues to present results. This article expands on the author's presentation at the 2012 Trends in Undergraduate Research in the Mathematical Sciences Conference.

1. Introduction

Over the last few years, analytic techniques to investigate problems in sports have grown in strength and popularity. In the 1990s and 2000s, the Oakland Athletics' "moneyball" approach was a novel way of examining the value of baseball players [Lewis 2003]. Analysts looked at more accurate baseball statistics than traditional stats such as batting average and earned run average. This provided an alternate way to find talented baseball players whose true skills were better than traditional baseball statistics indicated, and allowed the Athletics to find winning players with a limited payroll. One example of a less traditional statistic examined by analysts is on-base percentage. This statistic may give a more accurate representation of the run-scoring potential of a player than batting average. In the last few years, many professional sports teams have mimicked this approach by employing groups of researchers who analyze game and other team-related data searching for trends that will allow their teams to be more successful on the field. The improvements and suggestions are not necessarily complicated, nor do they require strong mathematics, and there are advantages for analysts who are novices to particular sports as this may allow fresh eyes to see an unnoticed trend.

Although we will focus on applications in sports, these types of analyses have become more prevalent in other fields. For instance, Nate Silver's election predictions

MSC2010: primary 62-07; secondary 91-02, 62P20, 62P30.

Keywords: sports, undergraduate research, analytics.

(fivethirtyeight.blogs.nytimes.com) are a well known application of data analysis to political science. Many industrial companies use analytic techniques for supply chain optimization. Other retail-based companies look for ways to save money or to better price products such as airfares, grocery items or insurance policies. There are also companies such as Coremetrics (now IBM Enterprise Marketing Management; see ibm.com/marketing-solutions), Omniture (now part of the Adobe Marketing Cloud; see adobe.com/solutions/digital-marketing.html) and Webtrends (webtrends.com) that specialize in analytics related to websites. In the future, there are likely to be many employment opportunities in analytics with large corporations or with analytics-focused companies. In this article, we will describe types of sports analytics problems found in the literature, useful skills for a successful project, ideas to help advisors find interesting problems and ways for students to present research results.

2. Useful skills

There is no prerequisite to know anything about sports before analyzing sports-related data. Sometimes strong knowledge about a sport can make it more challenging to develop new ideas because new ideas may involve unconventional approaches or statistics. One reason to use sports as a starting point is that there are mountains of available public data that can be analyzed in a variety of ways. In addition to new knowledge gained, one benefit for students who work in sports analytics is that they can learn skills directly applicable to many career paths. In subsequent sections, we describe important problems in the field, but first we will discuss the types of skills students will learn while conducting research in sports analytics.

One substantial skill that can be obtained by working on sports projects is data analysis. Students may need to learn theoretical aspects of different types of regressions and other statistical techniques. For example, a logistic regression may be helpful in making predictions about binary variables such as predicting a win or a loss. In addition students may need to test theories and so must learn about hypothesis testing to make mathematically rigorous statements about their results. A sports analytics focus could fit nicely with class projects in an upper-level mathematical statistics class. One source of ideas for projects in such a course is to look at problems similar to those listed in the essays for the 2010 Math Awareness Month (see mathaware.org/mam/2010/essays). Essays here discuss problems such as analyzing what makes a successful golfer, predicting baseball outcomes with sabermetric tools and understanding how wind, altitude and track geometries affect times in track and field. Another related source of ideas comes from the 2012 Math Awareness Month topic of mathematics, statistics and the data deluge (see mathaware.org/mam/2012/essays); included here are essays describing

the implementation of new analytic techniques. In addition, Amy Langville and Carl Meyer have recently published a new book [2012] on ranking appropriate for undergraduates.

Tools from learning theory for the classification of objects into different sets may also be required for certain problems. For instance, a project may involve determining what factors related to the dynamics of a NASCAR race are conducive to a large number of caution flags. Another example might be tracking baseball pitching patterns to detect subtle signs of fatigue. Such techniques, including clustering and using singular value decompositions may require deep forays into numerical linear algebra.

A more practical issue in many studies is obtaining and working with large sets of data. Students may have to know rudimentary or possibly more substantial programming in order to acquire and process data efficiently. This programming may take the form of simple if/then statements in Microsoft Excel or another spreadsheet program. It could also involve writing a script in Perl or another more sophisticated programming language so that large amounts of data can be pulled from a website storing the results of a particular set of sporting events. How to scrape data from a website will be discussed in more detail later. The diversity of skills required to be successful in this area makes it an excellent candidate for interdisciplinary projects linking students in mathematics, economics, computer science and related fields. It will also force students to be able to communicate ideas to students in related fields. Further, students can see how the strengths of their majors can work well with others to make more interesting projects. For instance, at the 2012 TURMS Conference, Michael Dorff, professor of mathematics at Brigham Young University and director of the Center for Undergraduate Research in Mathematics, remarked in his talk about careers in mathematics that many engineering firms are interested in hiring mathematicians because their thinking skills, when combined with the engineering know-how of others, is synergistic. Dorff mentioned that an engineering firm told him that by applying different schools of thought to a problem, groups with mixed technical backgrounds often find a solution faster and more effectively than a group with, say, only engineers. Another more objective measure, the 2013 jobs rating by careercast.com cited in the Wall Street Journal [Weber 2013], lists jobs related to analytics projects as six of the top 20 jobs in their ranking system: mathematician, university professor, actuary, software engineer, computer systems analyst and statistician.

3. Research problems

Ranking. Ever since players and teams started playing against each other in competitions, a timeless question has been: “Who is the best?” In some sports, such as

college football or basketball, it may not be feasible for every team to play every other team. As a result, rating systems have been developed to help compare teams. These systems can be derived from complicated mathematics, such as the Colley [2013] and Massey rating systems (masseyratings.com/theory), which are based on ideas from linear algebra. Variants of each of these ranking systems are currently used as part of the computer ranking for the college football Bowl Championship Series (see bcsfootball.org/news/story?id=4765872). Another interesting ranking method is from Keener [1993] who works with eigenvectors of a set of scores to determine a ranking. One example of recent undergraduate and faculty collaboration in ranking is between Furman University faculty members John Harris and Kevin Hutson and their students. They ranked baseball players using linear algebra-based methods (the project, not yet published, is mentioned in [Chartier 2012b]). The technique they describe might help to find effective players who are “diamonds in the rough” and may not have a high salary. Another recent REU project at Duke University investigated different ways to rank basketball teams using the so-called BODGE model [Barrow et al. 2012]. Ranking techniques that have traditionally been used for baseball, figure skating, football and basketball have also been applied to other sports, such as golf [Minton 2010], and there is still plenty of room for improvement.

One challenge in the paper of Harris et al. is that the data used for their analysis was not readily available. Instead, the authors had to find ways to collect information off publicly available websites. One way to obtain data is by using an application programming interface (API) that allows users to search specific company data in a specific manner. For instance, the API allowing users to search within Twitter (see dev.twitter.com) is very popular. Another way to obtain data is to “scrape” it—that is, collect data via the actual HTML output of websites. This technique may take some additional knowledge, but it is sufficiently common that there are many programs and blog posts to help users, such as scrapy.org and [Brody 2012]. What makes scraping challenging is that often the scraper has to reverse engineer how the data of interest is stored on a particular website. Given the ever-increasing amount of data in today’s world, it may be worth a student’s time to gain familiarity with scraping as data collection has become a field in itself. In particular, for sports, companies such as STATS, Inc. and the Elias Sports Bureau specialize in providing statistical information to sports-related clients.

If students or faculty do not have the time or expertise to obtain data via scraping, there are still many data sets ripe for exploration available online. One interesting project is outlined by Davidson College’s Tim Chartier [2012a], who essentially uses data from Ken Massey’s website (masseyratings.com/data.php) where data is stored in text file format for a variety of sports and seasons. A project like Chartier’s makes the analysis more approachable to students who may have minimal programming

experience. Another source of data that contains play-by-play and other forms of data for NBA games is publicly available for download from (basketballvalue.com). The NBA also has statistical information available on its website (nba.com). As time progresses, more and more data sets will likely be available in formats that are easy for students to manipulate.

Statistical modeling. Regression analysis is another tool that is widely used in many settings, including sports analytics. The basic premise is to develop a best-possible model for a particular situation based on a set of prior data. This technique is commonly employed in sports-related articles from economics journals, but practitioners in all fields have used this as an important modeling technique. For instance, in my mathematical statistics course at Davidson College, one of my students, Beau Reese, used regression analysis to determine how different factors (both on and off the field) affect the attendance of a Philadelphia Phillies baseball game. Among the characteristics he tested, factors that were significant were a Phillies' starting pitcher's earned run average, the number of "star" players starting in a given game, and whether the game in question was played on a Friday, Saturday or Sunday.

Game theory. Another area of research ripe with questions related to sports analytics is game theory. This field examines optimal strategies for players in structured situations. One example of an already studied area is penalty kicks in soccer. The penalty-kick situation has been modeled as a two-player zero-sum game. Recent empirical studies show that when examining some years of European soccer league [[Palacios-Huerta 2003](#)] penalty kicks, the average of the strategies is very close to the optimal minimax strategy predicted by game theory. Questions that are good for game theoretic analysis usually involve one or two players who have the option to make a small number of choices. In the penalty kick situation, the goalie can either decide to move left, move right, or stay in the center position. Similarly, the player may choose to kick the ball to the left, right or center of the goal. In the study, the author found that collectively soccer players play very close to a minimax strategy for penalty kick shooting.

Infographics. One additional area of interest in many settings is finding novel ways to present data. In an age where we are inundated with data, presenting results in an understandable way to potential customers, student-athletes, or donors is very valuable. *Infographics* is the study of data visualization. A simple example of an infographic is the USA Today series "Snapshots" (usatoday30.usatoday.com/news/snapshot.htm), where a graphic is presented to highlight a trend or describe survey results. Infographics may be useful for marketing, but more importantly, are a way to highlight patterns. For example, an illustrative computer

program might help tennis players find locations on the court where they are more than usually prone to make a mistake. It might help a baseball manager analyze a team's hitters quickly to understand their strengths and weaknesses, and be able to make the snap decision to bring in a particular pitcher. For some analytic tools to be useful, they need to be implementable in real time. For instance, a computer program may be helpful by informing a NASCAR crew chief the optimal pit stop strategy at any given point in the race. A further resource is David McCandless' TED talk, "The beauty of data visualization" [2010].

4. Ways to find problems

There are numerous methods for finding interesting research problems in this area. Consulting the academic literature, such as the *Journal of Quantitative Analysis in Sports* or other journals may be a good starting point. Interesting questions may also arise from local coaches who may have questions or coaching strategies that could form the basis of a paper. For example, a conversation with the basketball coach at Davidson College spurred the author to write an article [Britton and Yerger 2013] related to a coaching strategy that involved partitioning the game based on television time-outs. At Davidson College, we have also developed a relationship with Michael Waltrip Racing, and students are working on projects with applications to the racing teams there. Students may be more motivated when there are local people very interested in the outcome of their research. This also could help to develop relationships between colleges/universities and outside community organizations. pt

5. Presenting results

An advantage of working in sports analytics is that the diversity of skills needed to solve problems allows for a wide variety of venues where work can be presented. One natural place is in research journals related to sports analytics such as the *Journal of Quantitative Analysis in Sports*, the *Journal of Sports Economics* or the *Journal of Sports Sciences*. Undergraduate research journals such as *Involve* or the *Rose-Hulman Undergraduate Mathematics Journal* are also venues for presentation. Other more general purpose journals such as *The Statistician* or journals published by the MAA may also be appropriate, depending upon the problem. There are also many conferences where sports research would be of great interest. The best known conference in this area is the MIT Sloan Sports Analytics Conference. Other potential venues include the Wharton Sports Innovation Conference, the New England Symposium on Statistics in Sports, a Joint Mathematics Meetings special session, or an MAA section meeting. Internationally, the IMA sponsors an International Conference on Mathematics in Sport held biannually. From personal

experience, presentations involving sports tend to be well attended because conference participants from many specialties are attracted to sports research.

As a final note, I want to reemphasize that a deep knowledge of sports-related background information is not required to be successful in this area of research. Students have the ability to make contributions to a wide range of problems and develop and use a variety of skills that will be useful in their future endeavors. Another added bonus is that students can be the experts in sports-related topical knowledge and this can provide an enriching experience for both students and faculty.

Acknowledgement

I would like to thank the anonymous referees and Tim Chartier for their helpful comments and suggestions for expanding this article.

References

- [Barrow et al. 2012] D. Barrow, I. Drayer, P. Elliott, and G. Gaut, “Sports rankings REU final report 2012: an analysis of pairwise-comparison based sports ranking methods and a novel agent-based Markovian basketball simulation”, 2012, Available at <http://www.math.duke.edu/~idrayer/2012REU.pdf>.
- [Britton and Yerger 2013] P. Britton and C. Yerger, “Boxing in basketball: a round-by-round analysis of the college game”, 2013. unpublished manuscript.
- [Brody 2012] H. Brody, “I don’t need no stinking API: Web scraping for fun and profit”, blog page, 2012, Available at <http://blog.hartleybrody.com/web-scraping>.
- [Chartier 2012a] T. Chartier, “Got March Madness? Try math!”, The Huffington Post, 2012, Available at http://www.huffingtonpost.com/tim-chartier/march-madness-math_b_1341829.html.
- [Chartier 2012b] T. Chartier, “Mining the ball field”, The Huffington Post, 2012, Available at http://www.huffingtonpost.com/tim-chartier/mining-the-ball-field_b_1400696.html.
- [Colley 2013] W. Colley, “Colley’s bias free college football ranking method”, webpage, 2013, Available at <http://www.colleyrankings.com/matrate.pdf>.
- [Keener 1993] J. P. Keener, “The Perron–Frobenius theorem and the ranking of football teams”, *SIAM Rev.* **35**:1 (1993), 80–93. MR 94a:15012
- [Langville and Meyer 2012] A. N. Langville and C. Meyer, *Who’s #1? The science of rating and ranking*, Princeton University Press, 2012.
- [Lewis 2003] M. Lewis, *Moneyball: the art of winning an unfair game*, W. W. Norton and Company, 2003.
- [McCandless 2010] D. McCandless, “The beauty of data visualization”, TED talk, 2010, Available at http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html.
- [Minton 2010] R. Minton, “Tigermetrics”, 2010, Available at <http://www.mathaware.org/mam/2010/essays/MintonTigermetrics.pdf>.
- [Palacios-Huerta 2003] I. Palacios-Huerta, “Professionals play minimax”, *Review of Economic Studies* **70** (2003), 395–415.

[Weber 2013] L. Weber, “[Dust off your math skills: actuary is best job of 2013](#)”, blog post, 2013, Available at <http://goo.gl/MqUGx>.

Received: 2013-01-15

Revised: 2013-05-28

Accepted: 2013-10-11

cayerger@davidson.edu

Department of Mathematics and Computer Science, Davidson College, Box 7059, Davidson, NC 28035, United States

EDITORS

MANAGING EDITOR

Kenneth S. Berenhaut, Wake Forest University, USA, berenhks@wfu.edu

BOARD OF EDITORS

Colin Adams	Williams College, USA colin.c.adams@williams.edu	David Larson	Texas A&M University, USA larson@math.tamu.edu
John V. Baxley	Wake Forest University, NC, USA baxley@wfu.edu	Suzanne Lenhart	University of Tennessee, USA lenhart@math.utk.edu
Arthur T. Benjamin	Harvey Mudd College, USA benjamin@hmc.edu	Chi-Kwong Li	College of William and Mary, USA ckli@math.wm.edu
Martin Bohner	Missouri U of Science and Technology, USA bohner@mst.edu	Robert B. Lund	Clemson University, USA lund@clemson.edu
Nigel Boston	University of Wisconsin, USA boston@math.wisc.edu	Gaven J. Martin	Massey University, New Zealand g.j.martin@massey.ac.nz
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA budhiraj@email.unc.edu	Mary Meyer	Colorado State University, USA meyer@stat.colostate.edu
Pietro Cerone	La Trobe University, Australia P.Cerone@latrobe.edu.au	Emil Minchev	Ruse, Bulgaria eminchev@hotmail.com
Scott Chapman	Sam Houston State University, USA scott.chapman@shsu.edu	Frank Morgan	Williams College, USA frank.morgan@williams.edu
Joshua N. Cooper	University of South Carolina, USA cooper@math.sc.edu	Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran moslehian@ferdowsi.um.ac.ir
Jem N. Corcoran	University of Colorado, USA corcoran@colorado.edu	Zuhair Nashed	University of Central Florida, USA znashed@mail.ucf.edu
Toka Diagana	Howard University, USA tdiagana@howard.edu	Ken Ono	Emory University, USA ono@mathcs.emory.edu
Michael Dorff	Brigham Young University, USA mdorff@math.byu.edu	Timothy E. O'Brien	Loyola University Chicago, USA tbriell@luc.edu
Sever S. Dragomir	Victoria University, Australia sever@matilda.vu.edu.au	Joseph O'Rourke	Smith College, USA orourke@cs.smith.edu
Behrouz Emamizadeh	The Petroleum Institute, UAE bemamizadeh@pi.ac.ae	Yuval Peres	Microsoft Research, USA peres@microsoft.com
Joel Foisy	SUNY Potsdam foisyjs@potsdam.edu	Y.-F. S. Pétermann	Université de Genève, Switzerland petermann@math.unige.ch
Errin W. Fulp	Wake Forest University, USA fulp@wfu.edu	Robert J. Plemmons	Wake Forest University, USA rplemmons@wfu.edu
Joseph Gallian	University of Minnesota Duluth, USA kgallian@d.umn.edu	Carl B. Pomerance	Dartmouth College, USA carl.pomerance@dartmouth.edu
Stephan R. Garcia	Pomona College, USA stephan.garcia@pomona.edu	Vadim Ponomarenko	San Diego State University, USA vadim@sciences.sdsu.edu
Anant Godbole	East Tennessee State University, USA godbole@etsu.edu	Bjorn Poonen	UC Berkeley, USA poonen@math.berkeley.edu
Ron Gould	Emory University, USA rg@mathcs.emory.edu	James Propp	U Mass Lowell, USA jpropp@cs.uml.edu
Andrew Granville	Université Montréal, Canada andrew.andrew@dms.umontreal.ca	József H. Przytycki	George Washington University, USA przytyck@gwu.edu
Jerrold Griggs	University of South Carolina, USA griggs@math.sc.edu	Richard Rebarber	University of Nebraska, USA rrebarbe@math.unl.edu
Sat Gupta	U of North Carolina, Greensboro, USA sgupta@uncg.edu	Robert W. Robinson	University of Georgia, USA rwr@cs.uga.edu
Jim Haglund	University of Pennsylvania, USA jhaglund@math.upenn.edu	Filip Saidak	U of North Carolina, Greensboro, USA f_saidak@uncg.edu
Johnny Henderson	Baylor University, USA johnny_henderson@baylor.edu	James A. Sellers	Penn State University, USA sellersj@math.psu.edu
Jim Hoste	Pitzer College jhoste@pitzer.edu	Andrew J. Sterge	Honorary Editor andy@ajsterge.com
Natalia Hritonenko	Prairie View A&M University, USA nahritonenko@pvamu.edu	Ann Trenk	Wellesley College, USA atrenk@wellesley.edu
Glenn H. Hurlbert	Arizona State University, USA hurlbert@asu.edu	Ravi Vakil	Stanford University, USA vakill@math.stanford.edu
Charles R. Johnson	College of William and Mary, USA crjohnso@math.wm.edu	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy antonia.vecchio@cnr.it
K. B. Kulasekera	Clemson University, USA kk@ces.clemson.edu	Ram U. Verma	University of Toledo, USA verma99@msn.com
Gerry Ladas	University of Rhode Island, USA gladas@math.uri.edu	John C. Wierman	Johns Hopkins University, USA wierman@jhu.edu
		Michael E. Zieve	University of Michigan, USA zieve@umich.edu

PRODUCTION

Silvio Levy, Scientific Editor

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2014 is US \$120/year for the electronic version, and \$165/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW[®] from Mathematical Sciences Publishers.

PUBLISHED BY

 **mathematical sciences publishers**
nonprofit scientific publishing

<http://msp.org/>

© 2014 Mathematical Sciences Publishers

involve

2014

vol. 7

no. 3

Preface	245
DARREN A. NARAYAN	
Undergraduate research in mathematics with deaf and hard-of-hearing students: four perspectives	247
HENRY ADLER, BONNIE JACOB, KIM KURZ AND RAJA KUSHALNAGAR	
Challenges in promoting undergraduate research in the mathematical sciences	265
FERYAL ALAYONT, YULIYA BABENKO, CRAIG JACKSON AND ZSUZSANNA SZANISZLO	
Undergraduate research as a capstone requirement	273
HANNAH L. CALLENDER, JAMES P. SOLAZZO AND ELIZABETH WILCOX	
A decade of undergraduate research for all East Tennessee State University mathematics majors	281
ARIEL CINTRÓN-ARIAS AND ANANT GODBOLE	
The MAA undergraduate poster session 1991–2013	295
JOYATI DEBNATH AND JOSEPH A. GALLIAN	
Nonacademic careers, internships, and undergraduate research	303
MICHAEL DORFF	
REU design: broadening participation and promoting success	315
REBECCA GARCIA AND CINDY WYELS	
Papers, posters, and presentations as outlets for undergraduate research	327
APARNA HIGGINS, LEWIS LUDWIG AND BRIGITTE SERVATIUS	
ISU REU: diverse, research-intensive, team-based	335
LESLIE HOGBEN	
AIM's Research Experiences for Undergraduate Faculty program	343
LESLIE HOGBEN AND ULRICA WILSON	
Institutional support for undergraduate research	355
KATHY HOKE, ALESSANDRA PANTANO, MAZEN ZARROUK AND AKLILU ZELEKE	
Experiences of working with undergraduate students on research during an academic year	363
JOBBY JACOB	
The role of graduate students in research experience for undergraduates programs	369
MICHAEL A. KARLS, DAVID MCCUNE, LARA PUDWELL AND AZADEH RAFIZADEH	
An unexpected discovery	373
ERIKA L. C. KING	
Alternative resources for funding and supporting undergraduate research	377
ZACHARY KUDLAK, ZEYNEP TEYMUROGLU AND CARL YERGER	
Academic year undergraduate research: the CURM model	383
TOR A. KWEMBE, KATHRYN LEONARD AND ANGEL R. PINEDA	
Information for faculty new to undergraduate research	395
CAYLA MCBEE AND VIOLETA VASILEVSKA	
Promoting REU participation from students in underrepresented groups	403
HEATHER M. RUSSELL AND HEATHER A. DYE	
The Center for Industrial Mathematics and Statistics at Worcester Polytechnic Institute	413
SUZANNE L. WEEKES	
Nontraditional undergraduate research problems from sports analytics and related fields	423
CARL R. YERGER	