Journal of Applied Sport Management

Volume 14
Issue 4 *The Sports Analytics Revolution*

Article 2

12-1-2022

Where Analytics Gets It Wrong

Nicholas Watanabe
University of South Carolina, nmwatana@mailbox.sc.edu

Johan Rewilak
University of South Carolina

Recommended Citation

Watanabe, Nicholas and Rewilak, Johan (2022) "Where Analytics Gets It Wrong," *Journal of Applied Sport Management*: Vol. 14 : Iss. 4.
https://doi.org/10.7290/jasm14hbcr
Available at: https://trace.tennessee.edu/jasm/vol14/iss4/2

Journal of Applied Sport Management **Vol. 14, No. 4**
https://doi.org/10.7290/jasm14hbcr **https://trace.tennessee.edu/jasm/vol14/iss4/2**

Where Analytics Gets It Wrong

Nicholas M. Watanabe
University of South Carolina
Johan Rewliak
University South Carolina

Please send correspondence to: Nicholas Watanabe, nmwatana@mailbox.sc.edu

---

Analytics, machine learning, sentiment analysis, forest learning models. This terminology was foreign to most practitioners or academics within sport management, merely a decade ago. Today, such terms are not only are en vogue, but have become tools through which many sport organizations make critical and wide-reaching decisions. Moreover, as analytics have gained traction within the sport industry, academic researchers have also followed suit in adopting these methods within their own work.

The emergence of the term "analytics" in sport management is most traced to Michael Lewis' book Moneyball (2004), which described the advanced statistical approaches utilized by the Oakland Athletics to try and identify overlooked talent in Major League Baseball (MLB). Following its publication, managers and executives of sport organizations began to realize the power of using analytics to improve organizational performance and sought to replicate the use of these approaches to varying success (Hakes & Sauer, 2006).

Similarly, within sport management there has been a proliferation of academic programs, majors, research papers, and even professors who have stylized themselves using this terminology (Berri et al, 2006; McHale & Relton, 2018; Kharrat et al., 2020). Although it is not necessarily problematic to emphasize analytical approaches to decision making, the current state of analytics in sport management does have certain problems. In the following commentary, we will discuss the emergence of analytics, and outline with examples, certain problems that plague both academics and practitioners in their approach to using sport analytics.

The Foundations of Analytics

Statistics have been abundantly used in sports, predominantly sports science, where biomechanics have been analysed for injury prevention and performance, as well as providing an evidence-based approach for injury rehabilitation. Therefore, a natural extension was to apply these techniques to sport management. While the Oakland Athletics are credited as the team that highlighted the importance of using statistical analysis to better understand performance and behaviour, there are earlier examples from college basketball, ice hockey, and even baseball where coaches, managers, and other stakeholders have attempted to accomplish the exact same thing (Watanabe et al., 2021). As such, the innovation behind the Moneyball revolution was how it popularised the use of analytics, leading to its wide-spread adoption and application in sport management. This is opposed to analytics themselves being the innovation.

From an academic perspective, analytics should likewise not be considered as a new construct, as this term really has just become a way to describe the use of various forms of statistical analysis. What sport management scholars call "analytics," other disciplines such as computer science, mathematics, physics, and economics call statistical methods. For example, forest learning models were first conceived of by researchers at AT&T Bell Laboratories (Ho, 1995) who were trying to find ways to use computers to analyze textual documents. However, at their very core, even machine learning algorithms often use regression analysis as a key part of their algorithms. As such, even the most advanced analytical methods are typically based on existing statistical approaches that have been employed within our discipline.

Although the use of analytics is often described as being "state-of-the-art" advances in how we understand behaviors and phenomena, we again emphasize these are not re-inventions of the approach to conducting social science research. Some have displayed even greater levels of hubris, indicating that analytics would be the solution to understanding all behavior, and would essentially eliminate the need for theory because of the speed at which new discoveries would be made (Mazzocchi, 2015). In reality, big data and analytics have actually allowed for the expansion of existing theory (George et al., 2014) and even the development of new ones (George et al., 2016). As such, we believe that it is vital to acknowledge and understand what analytics truly is based upon. If we do not do so, it is not possible for us as researchers and practitioners to be cognizant of what we are really doing.

The Foundations of Analytics

Considering the previous point, we argue that one of the biggest issues in the use of sport analytics is that many who are engaged with it do not fully comprehend what they are actually doing. This is not to say there are no properly trained practitioners or academics who are doing analytical research in sport. Nor is this an attempt to gatekeep analytics to a select few individuals. Rather, we argue that those who are stylizing themselves as experts in the field are transparent about their understanding of current analytical methods, as well as outlining if they possess a solid foundation of the statistical methods that underlie these new popular approaches. If they do not possess these traits then collaborating with those who do should avoid adverse scenarios. For example, many scholars now conduct sentiment analysis using Natural Language Processing (NLP) without understanding that this method uses probability distributions of the appearance of words in a sentence. This means these researchers only have a surface level appreciation of the methods and results, and as such are often more concerned with statistical significance than developing deeper meanings from their work. Collaboration with those more versed in such methods would allow for improvement and more sophistication in the analytics research in sport.

Advancements in digital technology have provided greater accessibility for researchers to utilize large data sets and conduct complex statistical analyses. However, the ease of use does present issues in terms of the work that is being done both by sport organizations and academics. One of the primary concerns is that researchers often undertake complex analytics projects without proper recognition of what they are doing, and as such, are unable to truly understand and reflect upon the results.

One example is the increasing use of word clouds to examine textual data. With advancements in technology, the simplicity of creating a word cloud makes them popular and as simple as feeding texts or datasets into a program. The trade-off is that researchers provide all their control and ownership to a pre-determined algorithm to produce this word cloud, and trusts it will produce results that are reliable and meaningful. Even worse, researchers will often present the findings from these word clouds in scientific or professional settings and pass them off as having important implications without the knowledge of precisely how their results were constructed.

In essence, the ease of access has fostered the ability for researchers to ignore and bypass the theoretical and empirical training that was previously needed to conduct complex statistical analysis. From this, it has created a situation where many researchers in analytics have become reliant on machines to do the work for them, and as such have lost control of the research process. It can certainly be argued that is not anything new for researchers to use methods that they do not completely understand. Indeed, such issues have plagued every academic field of inquiry. However, the issue that has now emerged is that too many scholars have become overly reliant on algorithms and programs created by others to do the work for them and are losing touch with the scientific method in conducting their work.

Additionally, many scholars have entirely detached themselves from the empirical side of the research process because of the services and software that are now available to them. Today, it is now possible for a researcher to have a service/company collect data for them, clean the data for them, and even conduct all of the analysis, providing the researcher with only the finalized results from which to create reports and publications. In essence, many researchers have not only skipped over the accepted practices of how research should be conducted but have handed over complete control of the empirical sections of their research. Although such practices are problematic in both academia and the industry side of sport management, the fact that scholars believe that such an approach is an acceptable research practice is striking. Not only does it run counter to the norms and expectations of the field, but entirely removes most of the human element and thought from the research process. As such, there is clear need for more thought and control in how we conduct analytics-based research in sport.

Poor Applications of Analytics in the Sport Industry

To keep ahead of rivals, or to maximize sales and showcase themselves as market leaders, is one potential reason why many scholars and practitioners have detached themselves somewhat from the empirical research process. This method of sub-contracting aspects of the research process does yield benefits in terms of speed. But without the necessary guidance from the head researcher, or enough time to outline precisely what the research is investigating can lead to large inaccuracies.

This approach is well demonstrated in professional sports as since the Moneyball revolution, many teams have tried to replicate the success of the Oakland Athletics who gained a competitive advantage via the use of analytics. However, with all teams pursuing similar strategies using the same data, an equilibrium will be reached as depicted in Table 1.

Table 1

Nash equilibrium from Moneyball between two teams (Red and Blue)

Team/Action	Red team participate in “Moneyball”	Red team do not participate in “Moneyball”
Blue team participate in “Moneyball”	No overall winner, both pay statistical costs	Blue team win with a cost
Blue team do not participate in “Moneyball”	Red team win with a cost	No overall winner, pay zero statistical costs

However, innovation, and finding new metrics that outperform the status quo, should naturally restore a team’s advantage. To achieve this, machine learning, and complex algorithms have been created, scraping all available data related to the problem. This has led to an explosion in the number of variables that are now examined in the pursuit of success.

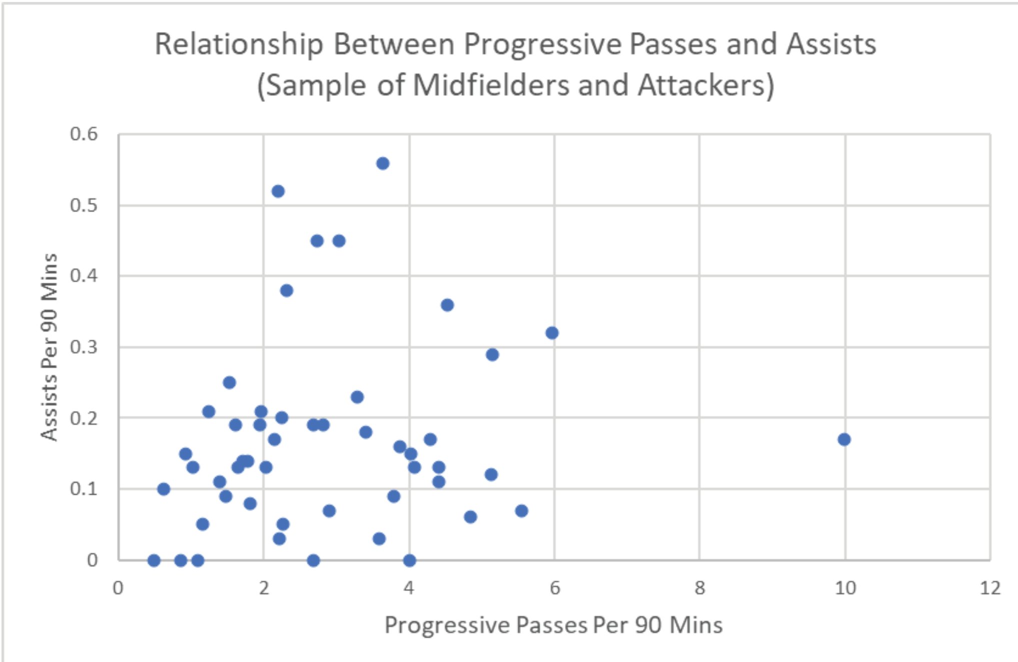
The increasing use of machine learning is understandable from a cost and calculative perspective. This is because typically, it is the researcher’s or the analyst’s time that is most expensive, therefore, permitting a machine or AI to take over this function is far more efficient. However, without enough guidance from a researcher, this will not yield the first-best outcome.

In addition, the explosion in data collection is consistent with the statistical mantra “you can never have enough data.” Yet, it has forgotten the essential aspect of this, that you can never have enough “useful” data. This can lead to irrelevant inputs being used to draw out equally irrelevant conclusions, resonating with the other statistical saying “if the salami is spoilt, so is the meal.” For sports analysts, this urges them to be carefully think about their models and variable selection to ensure accurate findings and not fall into a trap outlined in the example below.

In association football, progressive passes are now a common statistic. Tactic podcasts highlight their importance, scouts are being told to keep tabs on this metric when observing games, and player trading decisions may be determined by whose metric is greater. In Figure 1 we show the correlation between progressive passes and assists for a random sample of 50 Premier League midfielders and attackers.

To the naked eye, it appears that a player who produces more forward passes is rewarded with more assists. The correlation is weakly positive (0.15) but statistically insignificant. By adding more data - 30 randomly selected defenders - into the sample, the correlation rises (0.20) and the relationship becomes statistically significant. When examining the 30 defenders exclusively, the statistical significance increases, as does the correlation (0.63).

Figure 1
A Scatter Plot of Premier League Footballer Data



Note: Authors calculation using data from fbref.com

This demonstrates the plight of laissez-faire sports analytics, without the proper guidance or solid foundations. The obvious policy advice from this exercise is to hire defenders who pass progressively to increase the number of assists, but we know this is a bad recommendation and the relationship is far more complex. However, if sports analysts over-rely on machine learning and AI, these spurious correlations will keep appearing which is more detrimental than beneficial within the discipline.

Yet, ignoring machine learning and other novel techniques may also result in disastrous consequences, implying that sophisticated methods, which typically require complex computer algorithms still have a place in sport management. A Premier League club, that shall remain nameless, adopted an overly simplified approach to replace a striker who roughly scored a goal every other game for the club. The recruitment approach was to rank players based on their goals per game ratio from highest to lowest across Europe, and selection was based upon the highest ranked player who was attainable. Needless to say, the new acquisition didn't perform nearly as well as the player they replaced.

Thus, the overarching aspect is that humans need to think about the modelling, before allowing machines to overcome the complexities in calculations that the human mind cannot compute. As part of this, rather than exclusively favouring statisticians, those with a background in sports should be as highly sought after in analytics teams. Sports analytics is as much of an art than it is science. Still, when job adverts are provided for analysts to join sport analytical teams, the essential criteria often list a computer science or statistical major, and experience in statistical software and coding. Secondary is the desirable characteristic of an interest in the sport. Reweighting these priorities would be the first step in a more holistic and better sport analytics set-up.

Conclusion

Within this commentary, we have noted the rise in popularity of analytics within the management of sport organizations, as well as in sport research. While analytical approaches using statistical methods can certainly yield important contributions to the literature and management of sport organizations, we discussed numerous issues that exist in their application providing an example. In general, the issue is not in the statistical methods that compose what is now described as "analytics," but is instead in how we have been using them.

References

Berri, D., Schmidt, M., & Brook, S. (2006). *The wages of wins: Taking measure of the many myths in modern sport*. Stanford University Press.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, *57*(2), 321-326.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal*, *59*(5), 1493-1507.

Hakes, J. K., & Sauer, R. D. (2006). An economic evaluation of the Moneyball hypothesis. *Journal of Economic Perspectives*, *20*(3), 173-186.

Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

Kharrat, T., McHale, I. G., & Peña, J. L. (2020). Plus–minus player ratings for soccer. *European Journal of Operational Research*, *283*(2), 726-736.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Mazzocchi, F. (2015). Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Reports*, *16*(10), 1250-1255.

McHale, I. G., & Relton, S. D. (2018). Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research*, *268*(1), 339-347.

Watanabe, N. M., Shapiro, S., & Drayer, J. (2021). Big data and analytics in sport management. *Journal of Sport Management*, *35*(3), 197-202.