Beyond the Numbers: A Visual Display of Advanced NBA Defensive Analytics

Stephen Pelkofer
Georgia Institute of Technology

Richard Reasons
Georgia Institute of Technology

Matt Luppi
Georgia Institute of Technology

Hyun Jin lee
Georgia Institute of Technology

Tim Kim
Georgia Institute of Technology

1 THE PROPOSAL

A new wave of analytics have provided ways for fans of all sports to become more engaged with their favorite games. Unfortunately for fans, analytics have mainly focused on the offensive aspects of the game. Our team's aim is to provide an interactive application that will allow fans of the National Basketball Association (NBA) to explore, visualize, and assess the *defensive* performance of the league's teams and players.

2 SURVEY

As the burgeoning interest in sports analytics continues to grow, so too does the amount of new statistics, metrics, and models aimed at satisfying this collective appetite. However, as Franks et al. convincingly argue in their proposed set of "meta-metrics" [1], not all new sports metrics are constructive. Many lack, discrimination, stability or substantial independence from existing metrics. In our survey of the existing mainstream and academic analyses of defensive performance, we share in their criticisms and formalize a few of our own. These criticisms take two main forms. First, most conventional sources have over privileged offense-related metrics and have been slow to move beyond static, box-score style statistics. Secondly, early academic efforts have succeeded in incorporating modern computing and analysis techniques but have failed to reach or resonate with a general audience.

2.1 Offense-Oriented

The NBA has a very active, analytical following that pores over team performance, player statistics, and salary cap implications. Websites such as NBA Advanced Stats, Cleaning the Glass, and FiveThirtyEight are the major purveyors of basketball related statistics. Though these websites have begun to evolve and introduce more sophisticated analysis, much of what they provide are conventional, offense-oriented statistics in a timeworn tabular format. The historically heavy skew towards offensive statistics is a fact that has not gone unnoticed. Several early research papers cite this as a principal motivation for their work [2] [3]. Others have tried to expand the canon of traditional defensive statistics by exploring new metrics like rebounding efficiency [4], the three dimensions of rebounding [5] or the effectiveness of forcing weak side dribble[6]. We also reviewed research papers that eschewed defensive statics altogether in favor of defensive systems and strategy [7]. Despite these initial efforts, the reality has been slow to change. There is still a strong need to correct the offensive bias and to provide contemporary analysis that better reflects the true duality of the sport.

2.2 Academic Obscurity

One of the major recent developments propelling a new class of NBA statistics is the availability of play by play and player tracking data. Every NBA basketball arena now deploys advanced camera systems that track X, Y positioning coordinates of players and X, Y, Z positioning of the ball over 72,000 times per game [8]. The torrent of data enabled by this technology has led to a wave of new analyses that are able to incorporate critical but previously missing spatial dimensions. Some of the very first analyses to process player tracking data were able to enumerate previously inferred relationships, like the precise relationship between shot success and shot distance or shot success and opponent proximity [9].

Former Harvard researchers Kirk Goldsberry and Aledxander Franks were at the forefront of this new era of analysis. They published and contributed to several early papers that were among the first to incorporate and contextualize spatial data [10] [11] [12]. More recent analysis has taken several of the spatio-temporal techniques initially developed on basketball data sets and has begun applying them to team-based invasion sports in general [13]. While these papers have mined insight by applying some of the most sophisticated quantitative methods to these new spatially enriched datasets, such as expectation-maximization algorithms, Metropolis-Hastings, and singular-value decomposition (SVD), very little of this work has been able to cross-over into the mainstream.

2.3 Moving Beyond The Box-Score

To date most of the advances in defensive analytics on popularized media platforms, such as FiveThirtyEight, are composite, all-in-one type scoring systems. For example, FiveThirtyEight developed a proprietary system called Defensive Rating Accounting for Yielding Minimal Openness by Nearest Defender (DRAYMOND) [14] while ESPN utilizes it's own Defensive Rating Plus Minus (DRPM) [15] in order to distill defensive performance into a singular score. Though these metrics show an appreciation for defense, there are several serious drawbacks to this type of approach. First, it's very likely that the process of compressing the breadth of defensive performance into a singular statistic results in high amounts of distortion. Secondly, it obscures, or perhaps fails to reveal, the different underlying dimensions and relationships that encompass good defense. Lastly, despite their aims these metrics are ultimately another box-score style statistic best situated in a table and lacking the expressiveness of a strong visual presentation

However, in our review of the existing literature, we have come across several examples that strike a powerful balance between analytical rigor and accessible results. While the Sloan research paper Quantifying Shot Quality in the NBA [16] primarily focuses on offense, it contains many effective examples of visualizations that are intuitive yet informationally dense. Similarly, CourtVision: New Visual and Spatial Analytics [17] and Counterpoints: Advanced Defensive Metrics for NBA Basketball [18] expose general tendencies and spatial behaviors that can only be conveyed and interpreted visually. Since sports research and analytics remains a growing but nascent field, we also expanded our survey to include general research on topics such as visualization techniques for high-dimensional data. This search has helped surface several approaches we intend to implement including UMAP and t-SNE dimensionality reduction techniques for enhanced cluster separation and visualization [19] [20].

Ultimately, we find ourselves very supportive of the arguments Goldsberry puts forward in his recent book Sprawlball [21]. We believe that the future of defensive analytics in the NBA will increasingly rely on spatial data for quantification and visualizations for communication. It is the goal of our project to advance this paradigm forward, however slightly, by blending advanced quantitative analysis with engaging and interactive visualizations.

3 PROPOSED METHOD

3.1 Innovation

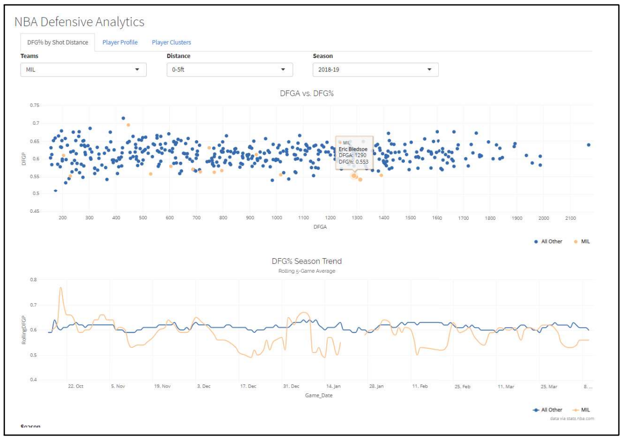
Our analysis benefits from several key innovations, the first of which is our custom feature set that was manually scraped and compiled from multiple sources. Our application benefits from the diversity of this data set which takes into account hundreds of features, including shot contests, rebounds, box-outs, and adjustments based on time played as well as the pace of individual games played. This data set appears to be significantly more exhaustive than those referenced in any of the analyses we reviewed.

The breadth of our data set is also a critical enabler of other innovations in our approach. As noted in our survey, certain sports metrics can have a high degree of absolute correlation. This interdependence occurs when several different measures are all highly connected to the same underlying skill or physical trait. To prevent inadvertently overweighting facets of defensive that are highly correlated or over-represented in box-scores, we performed principal component analysis on our feature set. By performing this dimensionality reduction we can ensure we're maximizing the amount of independent information expressed by our data set.

Our research also revealed that nearly all the popularized sites have focused on developing scoring systems that result in an ordinal rating of players. These systems rely on weightings that are preset, opaque, and unchanging despite a reasonably wide-range of defensive preferences and strengths that fans, players, and teams may value. So instead, a large part of our analysis focuses on leveraging machine learning techniques to cluster similar defensive performances and styles without imposing our own biases. The end-user can therefore select players whose defensive-play they admire and discover the players who exhibit the most similarity. This approach respects and incorporates the subjective preferences that enrich fandom, while still providing objective, data generated results.

3.2 Our Approach

Through various web-scraping programs, we have compiled a large dataset that contains the past five years of NBA game-by-game player statistics. With our data set collated and preprocessed, we are utilizing the R package Shiny to allow for the immediate development of an interactive web application. The analysis contained within our application is divided into three distinct pages each with its own focus. The first tab highlights team performance, while tabs two and three are player centric. We have developed some initial interactive team plots that illustrate performance across defensive statistics like Field Goals Defended at Rim Attempted (DFGA) and Field Goals Defended at Rim Percent (DFG



Teams: MIL

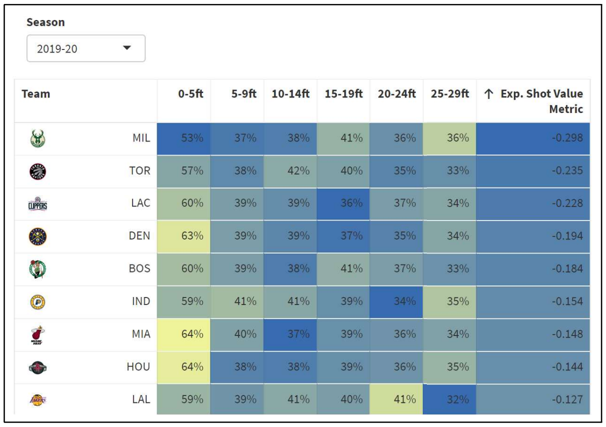
Distance: 0-5ft

Season: 2018-19

Figure 1: Defensive FG Percentage Plot Outputs

To enhance readability we are testing various color schemes, alpha compositing, and various chart types and axes. These refinement activities remain on going as we continue to develop new visuals and enhance established ones.

Additional interactive graphics we have already developed include the following table which depicts opposing team's expected shot value based on each team's ability to defend field goals relative to the league average.



Season

2019-20

Team	0-5ft	5-9ft	10-14ft	15-19ft	20-24ft	25-29ft	↑ Exp. Shot Value Metric
MIL	53%	37%	38%	41%	36%	36%	-0.298
TOR	57%	38%	42%	40%	35%	33%	-0.235
LAC	60%	39%	39%	36%	37%	34%	-0.228
DEN	63%	39%	39%	37%	35%	34%	-0.194
BOS	60%	39%	38%	41%	37%	33%	-0.184
IND	59%	41%	41%	39%	34%	35%	-0.154
MIA	64%	40%	37%	39%	36%	34%	-0.148
HOU	64%	38%	38%	39%	36%	35%	-0.144
LAL	59%	39%	41%	40%	41%	32%	-0.127

Figure 2: Reactable Output

In parallel, we are developing our player comparison model which will be underpinned by our clustering algorithm. This clustering model will help demarcate defensive styles as well as defensive similarities amongst players. Prior to conducting the clustering algorithms, we are performing dimensionality reduction, primarily through iterative executions of principal component analysis (PCA) to dampen collinearity and extract as much of the independent internal structure of the data as possible. Though PCA will cause use to lose some interpretability directly traceable to the original features, we believe this to be a critical step due to the high correlation between certain statistics. Additionally, it's a necessary step to facilitate the visualization of the data in a lower dimensional space in order to illustrate clustering or proximity between certain players.

We are currently contemplating several different visualization techniques to best illustrate our cluster analysis. Our leading approach is to cast a 2 or 3-dimensional projection of our component factors (with their clusters) using t-SNE or UMAP to aid users in exploring the player profiles. Although the stochastic nature of these models does not make it a good fit for creating the clusters themselves, we believe it may provide the most compelling visual to explain our high-dimensional clusters.

3.3 Experiments and Model Evaluation

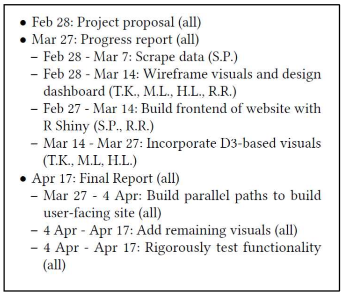
We are in the process of evaluating multiple clustering algorithms to serve as the basis of our analysis. Amongst the approaches we are assessing are industry standards, such as k-means and k-nearest neighbors, as well as more computationally intensive approaches like DBSCAN. We are carefully taking into consideration the trade-offs of each approach and how our data lends itself to the application of each methodology.

Since the algorithm will be run initially to develop the clusters and then held static, the initial performance and processing time is of little concern. Additionally since both DBSCAN and k-means are intended for unsupervised data sets, we'll need to assess the validity of the clustering based on our own evaluation of players and what we deem sensible groupings. Since this clustering will be performed on the PCA components and not the raw feature set, it will make the deconstruction of the clustering harder to both trace back and ground in the original features.

In addition to evaluating the sensibility of the clustering, we will need to consider the distinguishability of each cluster. Since we want to render these clusters as a visual in low-dimensional space (2-3 dimensions) it's important that the clusters have clear separation and distinctness. Otherwise they may appear arbitrary or a product of the parameters passed to the algorithm. Our research on t-SNE shows that it has shown a lot of promise in this particular area, but its quadratic scaling complexity and it's extremely resource intensive processing require further testing against our particular data set.

3.4 Plan of Activities Revisited

All members have contributed a similar amount of effort to this point, but we have shifted responsibilities to match our team's strengths. Below we have depicted our prior proposed plan as well as our updated plan of activities and accomplishments. It's worth noting that this plan of activities remains subject to change and that for the sake of brevity it does not capture all of the work that has been performed to date nor all of the remaining work to be completed:



- Feb 28: Project proposal (all)
- Mar 27: Progress report (all)
  - Feb 28 - Mar 7: Scrape data (S.P.)
  - Feb 28 - Mar 14: Wireframe visuals and design dashboard (T.K., M.L., H.L., R.R.)
  - Feb 27 - Mar 14: Build frontend of website with R Shiny (S.P., R.R.)
  - Mar 14 - Mar 27: Incorporate D3-based visuals (T.K., M.L, H.L.)
- Apr 17: Final Report (all)
  - Mar 27 - 4 Apr: Build parallel paths to build user-facing site (all)
  - 4 Apr - Apr 17: Add remaining visuals (all)
  - 4 Apr - Apr 17: Rigorously test functionality (all)

Figure 3: Initial Proposed Plan of Activities

Updated Plan of Activities

- Project Proposal (all)
- Data scraping (SP, RR))
- Build frontend/backend repositories (HL)
- Wireframe visuals and design dashboard (SP, TK)
- Visualization of initial analysis (SP, TK)
- Progress/Final Report writing lead (ML)
- Dimensionality/Cluster Analysis with PCA/TSNE (RR, HL)
- Visualizing Dimensionality Analysis (SP, TK)
- Rigorously Test Functionality (all)

.

4 CONCLUSIONS

Though we have a large thrust of development work and unit testing left to complete, our team is satisfied with the progress we've made to date and feel confident in our ability to deliver our project on-time and as designed. Our highest remaining priorities are down-selecting and tuning our clustering algorithm and developing additional complementary visualizations for each page of our web application. Given the time we have budgeted for each activity, we believe that we remain on schedule and are well-positioned for final deliverables. We will continue to collaborate efficiently and respond to the challenges that remain and the adaptations they demand.

REFERENCES

[1] Cervone Bornn Franks, D'Amour. Meta-analytics: tools for understanding the statistical properties of sports metrics. *Journal of Quantitative Analysis in Sports*, 3.3, 2007.

[2] Pelton Rosenbaum Kubatko, Oliver. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 12.4, 2017.

[3] Kirk Goldsberry and Eric Weiss. The dwight effect: A new ensemble of interior defense analytics for the nba. *Sports Aptitude, LLC*, pages 1–11, 2013.

[4] Slobodan Simovic. Analysis of influence of basketball game-related statistics on final result based on differences at the 2017 fiba asia cup. *Asian Journal of Sports Medicine*, 10.1, 2019.

[5] Jeff Su Sheldon Kwok Tal Levy Adam Wexler Noel Hollingsworth Rajiv Maheswaran, Yu-Han Chang. The three dimensions of rebounding. *MIT Sloan Sports Analytics Conference*, 2014.

[6] James T. Bartholomew and David A. Collier. The benefits of forcing offensive basketball players to their weak side. *Journal of Multidisciplinary Research*, 4:19–27, 2012.

[7] Miguel Ángel Gómez Jesús Salado Alejandro Álvarez, Enrique Ortega. Study of the defensive performance indicators in peak performance basketball. *Sports Psychology Magazine*, 18, 2009.

[8] Mark Wilson. Moneyball 2.0: How missile tracking cameras are remaking the nba. https://www.fastcompany.com/1670059/moneyball-20-how-missile-tracking-cameras-are-remaking-the-nba.

[9] S. Weil. The importance of being open: What optical tracking data can say about nba field goal shooting. *MIT Sloan Sports Analytics Conference*, 2011.

[10] Alexander Franks. Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9.1:94–121, 2015.

[11] Bornn Goldsberry D'Amour, Cervone. Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. *MIT Sloan Sports Analytics Conference*, 2014.

[12] Andrew Miller, Luke Bornn, Ryan Adams, and Kirk Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. *31st International Conference on Machine Learning*, 2014.

[13] Joachim Gudmundsson and Michael Horton. Spatio-temporal analysis of team sports – a survey. *ACM Computing Surveys*, 50, 2016.

[14] Fivethirtyeight: A better way to evaluate nba defense. https://fivethirtyeight.com/features/a-better-way-to-evaluate-nba-defense/.

[15] Espn: The next big thing: real plus-minus. https://www.espn.com/nba/story/_/id/10740818/introducing-real-plus-minus.

[16] Jeff Su Sheldon Kwok Tal Levy Adam Wexler Kevin Squire Yu-Han Chang, Rajiv Maheswaran. Quantifying shot quality in the nba. *MIT Sloan Sports Analytics Conference, year = 2014*.

[17] Kirk Goldsberry. Courtvision: New visual and spatial analytics for the nba. 2012.

[18] Counterpoints : Advanced defensive metrics for nba. *MIT Sloan Sports Analytics Conference*, 2015.

[19] Etienne Becht. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 31, 2019.

[20] Laurens van der Maaten and Geoffrey Hinto. Visualizing data using t-sne. *Journal of machine learning research*, pages 2579–2605, 2008.

[21] Kirk Goldsberry. Sprawlball. sprawlball: A visual tour of the new era of the nba. houghton mifflin harcourt. 2019.