Wharton SPORTS ANALYTICS
Student Research Journal

2022, Vol 4

wsb.wharton.upenn.edu/student-research-journal

Quantifying NBA Play Style Drift Using Finite Mixture Multinomial Model

Andrew Castle, *W'21*
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

Grant Cho, W'21
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

David Fan, W'21
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

Advisor: Abraham Wyner, PhD
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

Abstract

NBA games have been changing. In particular, such play style shift, through 'positionless' basketball, has often been associated with more three- pointer attempts and fewer inefficient mid-range shots. However, while such aggregate trend is trivial to analyze, we pose three outstanding questions that the aggregated trend fails to encapsulate. First, is that there is a growing volume of three-point specialists coming into the league. Second, is that there is a change in play style across all players. Third, is that such trend is heterogeneous across top and bottom performing teams. Through player clusters created from a Finite Mixture Multinomial model, we found that the overall shift towards three-pointers seen in the NBA is attributable both to a growing volume of rookies who are focused on predominantly shooting threes and a more general trend where all-around players are transitioning to become three-point specialist, albeit decreasing over time. In addition, while both the top and bottom performing teams subscribe to similar trends of acquiring more three-point shooters, we do see weak evidence suggesting that top teams appear to be capturing this trend at a faster rate.
model

key words: basketball, player style, cluster, finite mixture multinomial, shot

Introduction

In recent NBA history, there has been a widely-observed trend towards more efficient play styles (e.g. characterized by more three-pointers and less inefficient mid-range) through what has been called “positionless basketball”. While basketball has had five “positions” historically, the boundaries between those positions have shifted and the actual roles played by players are not well-described by these simplistic position labels. As such, there is a real need for a system that labels players based on their actual skills and tendencies. Once we have this set of labels, we can use them to generate useful insights about the evolution of the league over time and the optimal way to construct a roster or lineup. While the primary purpose of this paper is to use finite mixture multinomial models to create clusters of player roles in a disciplined way, we will also use these labels to explore the evolution of the league over the past two decades.

Methods

While there are any number of factors we might want to include in our clusters, we decided to primarily focus on a dataset of shot types by player. Using the RVest package, we were able to write a scraper that crawled Basketball Reference pages to extract counts of shot makes by location. For each player, we were able to pull in the count of shot makes at the rim, from short range, from mid-range, from deep mid-range, and from beyond the three-point arc. We ran the scraper on all players who appeared on an NBA roster from 1999 to 2020. After scraping the data, we had to do a substantial amount of work to clean up the output and turn it into a usable dataframe. Our ultimate data consisted of labeled player-seasons with shot make counts for each location/bucket; using this output, we were able to proceed to modeling the data and creating our final clusters.

Modeling Approach & Results

To be able to answer the aforementioned questions, we must find a way to principally model similarity and differences across players and their respective playing style. In this section we detail the methods we have applied.

Finite Multinomial

A finite mixture multinomial clustering system was used to cluster players based on their shot selection. To first provide some context, we believe that the shots taken by the players exist in a choice space, where every shot taken is a choice between the five categories above. This naturally leads to a choice based model, the multinomial. To briefly describe the model, we start with a regular multinomial model. Under a multinomial model, we assumed the following for a single player:

- An individual has some fixed underlying propensity that govern their shot selection: [p_1, p_2, ... p_N]
- The actualization of these propensity to actual shots made will follow a multinomial distribution where the parameters are the same as the underlying propensity listed earlier

To extend this in a multi-player context, one can naturally see two simple extensions. First, where we fit one multinomial model per player or second, where we fit one multinomial model for all the players. The former poses a world where every player is completely heterogeneous and independent where no information sharing can be obtained, where the latter poses a world where every player is exactly the same, sharing the same propensities that can be jointly estimated.

Both approaches can be problematic. The first approach poses a big problem for players with limited sample size as the estimated propensities for these players will be based on very limited data points. The second approach is problematic due to the homogeneity assumption - clearly players like Steph Curry and Shaquille O'Neal wouldn't share the same propensities for the kind of shots they attempted.

Mathematical Formulation of Multinomial, for an individual player:

$$x \sim Multinominal\ (p_1,\ p_2\ , \dots\ p_n)$$

$$x \sim \frac{n!}{x_1!\ x_2!\ \dots\ x_k!}\ (p_1^{x_1}, p_2^{x_2}, \dots\ p_k^{x_k})$$

$$Joint\ Log\ Likelihood \sim \sum_{j=1}^{N} \log\left(\frac{n!}{x_{1j}!\ x_{2j}! \dots\ x_{kj}!}\ (p_1^{x_1}, p_2^{x_2}, \dots\ p_k^{x_k})\right)$$

where,

$$k = different\ catories\ of\ shots$$
$$n = total\ shots\ taken\ by\ the\ specific\ player$$
$$x_{ij} = shots\ taken\ at\ i\ category\ and\ by\ the\ j\ player$$
$$N = number\ of\ player$$

That said we propose a model that sits in between these two extremes. Where instead of having just one multinomial model, we propose that the entire shot space is composed of a finite cluster of different multinomial models. In this specification we presume the following:

- Every cluster has some fixed underlying propensity that govern the shot selection of players within the cluster: [p_1, p_2, ... p_N]
- Every player belongs to a specific cluster, where all players within the cluster share the same propensity
- The actualization of these propensity to actual shots made will follow a multinomial distribution where the parameters are the same as the underlying propensity listed earlier

Mathematical Formulation of Finite Mixture Multinomial,

For a player in cluster c:

$$x \sim Multinominal\ (p_{1c},\ p_{2c}\ , \dots\ p_{nc})$$

$$x \sim \frac{n!}{x_1!\ x_2!\ \dots\ x_k!}\ (p_{1c}^{x_1}, p_{2c}^{x_2}, \dots\ p_{kc}^{x_k})$$

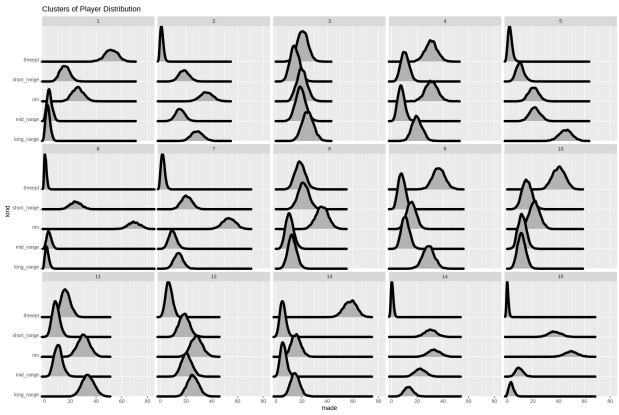
for all players,

$$Joint\ Log\ Likelihood \sim \sum_{i=1}^{N} \log\ (\sum_{c=1}^{C} q_c\ \frac{n!}{x_{1j}!\ x_{2j}!\ \dots\ x_{kj}!}\ (p_1^{x_1}, p_2^{x_2}, \dots\ p_k^{x_k}))$$

Such finite mixture approach allows us to regularize players with limited sample to his nearest cluster. It also allows us to generate cluster inferences easily and principally by applying bayes rule on each player. Such application will generate a probability of each player being in a respective cluster, where the final cluster selected for the player will be the one that maximizes the aforementioned probability.

$$P(Player\ A\ in\ Cluster\ 1) = \frac{(Likelihood(Cluster\ 1\ shot\ of\ Player\ A) * Prior(Cluster\ 1))}{\sum_{n=1}^{k}(Likelihood\ (Cluster\ n \mid shots\ of\ Player\ A) * (Prior\ (Cluster\ n))}$$

The model is to be estimated using Maximum Likelihood, where we seek to maximize the joint log likelihood of observing the shots of all players. The final number of clusters was decided through an iterative likelihood ratio test approach, where Bonferroni correction was applied to a p-value threshold of 0.05. That is, we compared the log-likelihood of (n + 1) clusters to n clusters using a p-value threshold of 0.05/n to account for repeated testing.



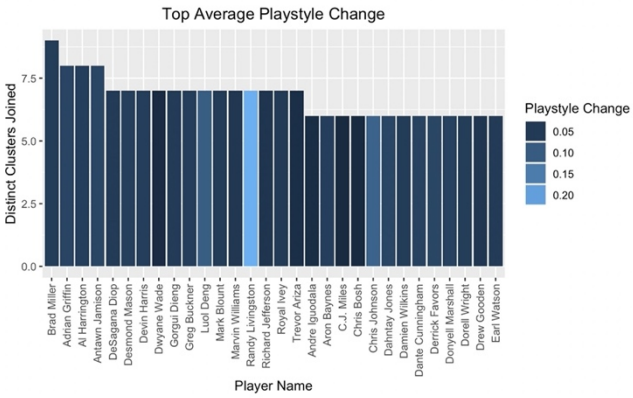


Figure 1a, *Shot-making distributions based off of clusters,* *Fig 1b, Players with the most amount of cluster changes*

Results

The model forms 15 clusters after the iterative Log- Likelihood test. Each cluster has a distinct shooting frequency distributions, spread out through three- pointers, long-range twos, shots near the rim, and mid- range shots. The dependent variable highlight in the graph is the number of shots made from that particular range should a player from the specified cluster randomly attempt 100 shots. Suppose we take a random center in the league who likes to shoot a lot from under the rim and a bunch of fade-away jumpers. That particular player might be classified as someone in cluster 7 since he is most dominant at the rim but also enjoys shooting at a distance further from the basket while also not taking a lot of three-point attempts.

Individually, we can note who's had to change their play style the most throughout their career. We started by creating and then normalizing a Kullback-Leibler Divergence matrix before finding the maximum values of the averaged outputs results, which would provide a readily interpretable estimate as to how much a player's play style over the course of that person's career. Specifically, for each value in the KLD matrix, the normalization would be as follows:

$$Z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Suppose we compare cluster 1 and cluster 2, the KLD matrix would provide a value of 1.35. The maximum dissimilarity in the matrix was 1.94, while the minimum was 0, so the normalized value of cluster 1-2 would be 1.35 = 0.7. This means that, relative to other values within the matrix, clusters 1 and 2 are very dissimilar.

Applying this concept across all players in the NBA with minimum 100 shots taken per season, we can see that Randy Livingston not only jumped multiple clusters throughout his time as a player, but that the average transition he had to make was extremely large. This means that Randy had to alter his play style significantly over the course of his career, which makes sense given the fact that he was with 14 different teams while he was a player.

Hierarchical Clustering

We can take this a step further with hierarchical clustering by discerning how popular some of the shots taken are for each cluster and then categorize them into super clusters to further generalize our inference. Figure 1 displays our "taxonomy" of NBA playstyles. Note how clusters 6, 7, and 15 have a lot of shots taken under the rim; players in clusters 1, 9, 10, and 13 tend to take a lot of three-point shots; players in clusters 2, 12, and 14 tend to take short-range jumpers; players in clusters 3, 4, 8, and 11 are generally all-around players; and finally, players in cluster 5 tend to take a lot of long-range two-pointers. We will return to this point at length in later sections.

Cluster	*Cluster Name*	*Representative Player*	*Famous Example*	*Super Cluster*
1	3pt + Rim Specialist	Matt Barnes	James Harden	Long-Range
2	Short+Mid	Carlos Boozer	Karl Malone	Short Range
3	Balanced 3pt + Mid	Dirk Nowitzki	Kawhi Leonard	All-Around
4	Balanced 3pt + Mid	Leandro Barbosa	Lebron James	All-Around
5	Long Twos	Kurt Thomas	Michael Jordan (Wizards)	Long 2pt
6	Can't Shoot	Tyson Chandler	Ben Simmons	Non-Shooters
7	Rim + Mid	Zaza Pachulia	Blake Griffin (2010-11)	Non-Shooters
8	Rim w/Some 3's	Lamar Odom	Russell Westbrook	All-Around
9	3+Long2	Jamal Crawford	Reggie Miller	Long-Range
10	3pt Primary	Manu Ginobili	Bradley Beal	Long-Range
11	Rim + Long 2s	Carmelo Anthony	Carmelo Anthony	All-Around
12	Versatile Rim	Andre Miller	Derrick Rose (2010-11)	Short Range
13	3pt Specialist	Kyle Korver	Kyle Korver	Long-Range
14	No Threes	Elton Brand	Tim Duncan	Short Range
15	Short Range	Kendrick Perkins	Shaquille O'Neal	Non-Shooters

Figure 2: Hierarchical Clustering: clusters with example players

Assessing the Results

With any unsupervised learning or clustering approach, there is of course no “ground truth" with which we can compare our results. As such, we need to use our own intuition and knowledge of basketball to assess whether or not our clusters are generating meaningful and useful groups of playstyles. To do this, can manually look at the players the computer has assigned to each cluster and assess whether these players are actually similar to one another. We can also create meaningful labels for each cluster - for instance, rather than calling it “Cluster 13", we might refer to players in this cluster as being “3pt Specialists." Figure 2 shows our manually assigned names for each cluster, the “most representative" player for each cluster, and a famous player who epitomizes the nature of this playstyle. For instance, Kyle Korver is the prototypical 3pt Specialist, while James Harden has defined the archetype of the player who almost exclusively shoots threes and drives to the rim.

For the sake of clarity, we assigned players who have spent most of seasons in a particular cluster next to their respective cluster. To demonstrate, Kyle Korver has spent 12 seasons in cluster 13, which is the most amount of seasons of all players for that particular cluster. As such, his playing style is the most representative of that cluster, so anyone within that cluster can be assumed to play similarly to Korver. We used our discretion to choose the “famous player" example for each cluster.

We generally find that these results make sense; the cluster labels mostly well-represent the players within them and identify important differences between clusters. Some of the clusters are a bit too similar (is a Balanced 3pt + Mid really all that different than a 3pt + Rim player?) although even in these cases, the clusters are picking up meaningful differences between these players. In this particular instance, Lebron James - the prototype of the Balanced 3pt + Mid - has a materially different playstyle from James Harden, the classic 3pt + Rim player. Someone like Russell Westbrook (a similarly named Rim + Some 3pt player) is different still; he relies far more on dunks and layups, with fewer threes than Lebron or Harden and more midrange than Harden.

By applying this concept of dissimilarity to form super clusters through hierarchical clustering, we're able to broadly generalize different playing styles into something that's much easier to interpret. The five super clusters we formed categorized players into rim players, three-point shooters, short-range shooters, all- around players, and inefficient players. Players who are placed within the inefficient cluster are known to take long-range twos and not much of anything else.

For those cases where the final 15 clusters are some- what similar, we can rely on our "super-clusters" to abstract away some of these marginal differences be- tween players. For instance, Lebron and Westbrook might be a bit different from one another, but they are both wildly different than Shaq or Ben Simmons, neither of whom shoots from anywhere outside the paint. Our five super-clusters can also be used in cases where we do not have large enough sample sizes in our more refined buckets; we can in essence borrow strength from closely related buckets in order to make more informative analyses.

Putting this all together, we can create a full "taxonomy" of NBA player types, shown in Figure 3.

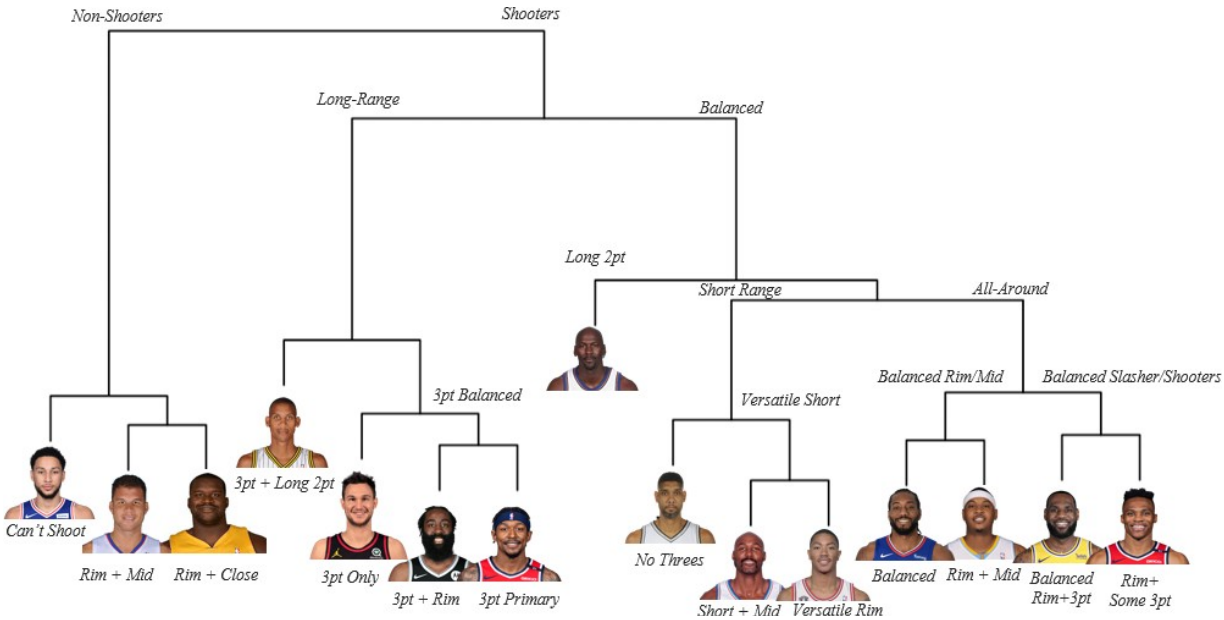


Figure 3. Shot-making distributions based off of clusers1

Inferences & Discussion

There's no question that the league has changed considerably over the past 20 years in favor long-distance shooting and versatility. The clustering we've done can directly provide insight as to how much the league has transformed.

General League Evolution, 1999- 2020

What we found was that three-point shooters have be- come far more popular since the 1999 season, with players with the James Harden "only three's and layups" play style becoming more prevalent in today's league. On the other hand, the number of inefficient shooters has seen a precipitous drop most likely due to players opting for more efficient shots, such as three-pointers or rim shots. In fact, as classified by the model there have been no inefficient players in the league since at least 2016.

Furthermore, as a result of an increase in shot-taking efficiency, the number of players in the “Long 2pt / inefficient" cluster has declined drastically. Observing the distribution chart of shot-taking per cluster, we can indeed see that players who’s most popular shot- making selection is the long-range two are becoming increasingly unpopular.

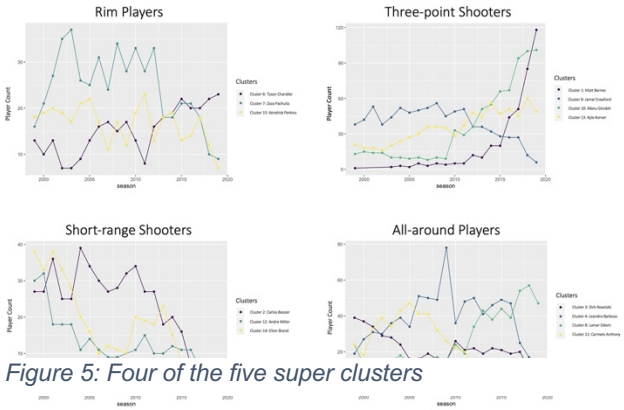


Figure 5: Four of the five super clusters

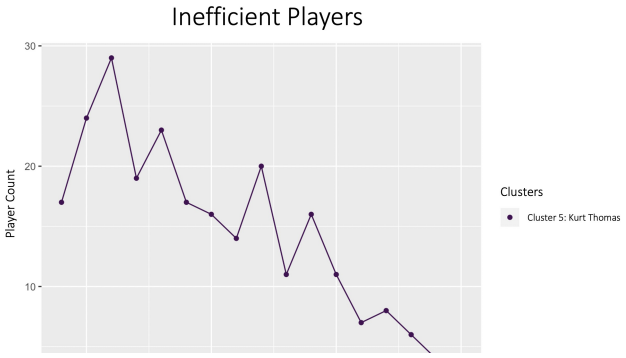


Figure 6: Players who take lots of long-range and mid-range shots but not much else

Disaggregating League Evolution

While the overall trend in the league is obvious, a key question remains: what is actually driving these changes? Are players shifting their playstyle over time? Are players of a certain style being pushed into retirement early? While we explored many of these topics at length, we will cover only a brief summary of each.

Year-to-Year Transitions Between Playstyles

Given our cluster labels for each player-season, we can very easily compute the empirical frequency with which a player in a given cluster will end up in another cluster in the following year. Using these probabilities, we can make some interesting statements about which transitions are most likely Very few short-range and rim players make the transition to long-range shooters, which makes sense given they may only shoot at short-range because they may not believe they can confidently make a lot of threes. Instead, these kinds of players make the switch, if any, to either all-around players or rim players.

Long-range shooters are the least likely to switch clusters. This may be because their play style is more reliant on shooting rather than getting up close and personal at the rim. Players like Kyle Korver and Joe Harris are, for example, almost exclusively used for their extremely efficient three-point shooting abilities. As three-pointers become increasingly valuable, there’s little reason for a sharpshooter to make the switch.

Inefficient players who do end up switching clusters mostly transition into either short-range or all-around players; very few become long-range shooters. Michael Jordan is twice a noteworthy member of the inefficient cluster because of his strong preference for mid to long-range twos; however, since he was so accurate with those shots, he was never truly inefficient with his shot-making and remained valuable as a result.

If we hold constant the player cohort (only looking at those who played in every season between 2014 and 2020), we see an essentially identical trend. Clearly, players are actually transitioning towards these higher- value playstyles, rather than this just being a change in the overall player

mix.

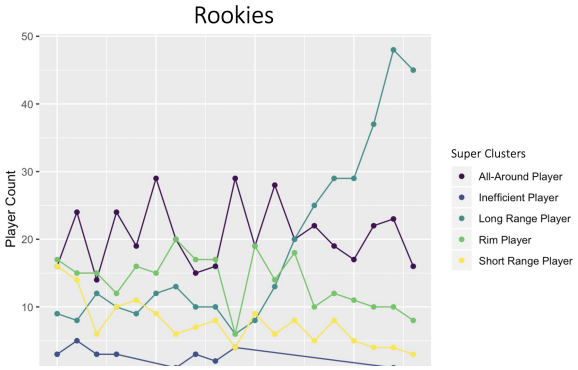


Figure 7: Rookie players play styles over the past twenty years have also mostly shifted in line with the rest of the league

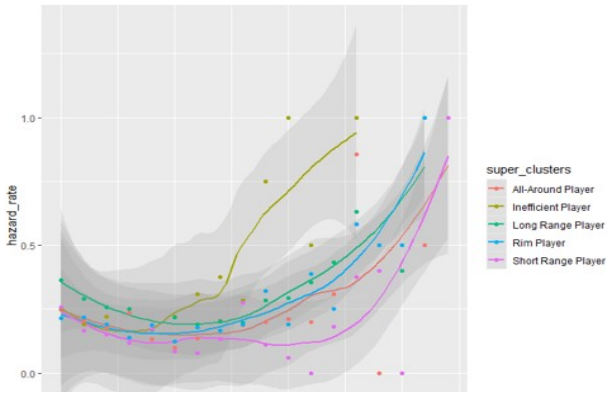


Figure 8: Hazard rate for retirement, by career year and player super cluster

Rookie Composition

In addition to the actual changes in the playstyles of existing players, greater number of rookies are entering the league as three-point shooters (Figure 7), which we might infer to be a result of collegiate and pre- collegiate programs also beginning to realize the value behind long-range shooting. The number of all-around players entering the league have steadied while inefficient players have generally become non-existent since at least 2018. We believe this is a big driver of the overall trend in league results.

Retirement Risk

It is also worth asking if there are some players who have been forced to leave the league because they could not adapt to a new playstyle. Indeed, we find some weak evidence that players in the “Long 2-pt" bucket face increased career risk, even after controlling for the number of years they have already spent in the league. Figure 8 shows this graphically, with the retirement rate on the y-axis. We constructed a logistic regression that shows a similar result; details are shown in Figure 9.

```
Call:
lm(formula = retired ~ nseasons_sq + year + super_clusters, data = hazard_dat %>%
    mutate(nseasons_sq = nseasons^2))

Residuals:
    Min      1Q  Median      3Q     Max
-0.4888 -0.2613 -0.1677 -0.0318  1.0055

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -4.585e+01  2.824e+00 -16.235   <2e-16 ***
nseasons_sq                    -3.394e-04  1.614e-04  -2.103   0.0355 *
year                            2.292e-02  1.406e-03  16.303   <2e-16 ***
super_clustersInefficient Player  9.189e-02  3.587e-02   2.562   0.0104 *
super_clustersLong Range Player   2.764e-02  1.547e-02   1.787   0.0740 .
super_clustersRim Player          1.939e-03  1.663e-02   0.117   0.9072
super_clustersShort Range Player -1.283e-02  1.888e-02  -0.679   0.4969
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4034 on 4687 degrees of freedom
Multiple R-squared:  0.06634,   Adjusted R-squared:  0.06514
F-statistic:  55.5 on 6 and 4687 DF,  p-value: < 2.2e-16
```

Figure 9: Logistic regression for retirement based on career length and player type

Team Success vs. Role Composition

While a thorough assessment of the effect of role com- position on team success could fill several papers in its own right, it is worth briefly exploring how the best and worst teams in the league have been composed over time. Figures 13 and 14 show these comparisons for our five super-clusters. In general, we find few substantive differences here. The absolute worst teams have more “Long 2pt” players than the best teams and for a longer time. On the flip-side, the

best teams have shifted towards three-point shooting sooner than the worst teams.

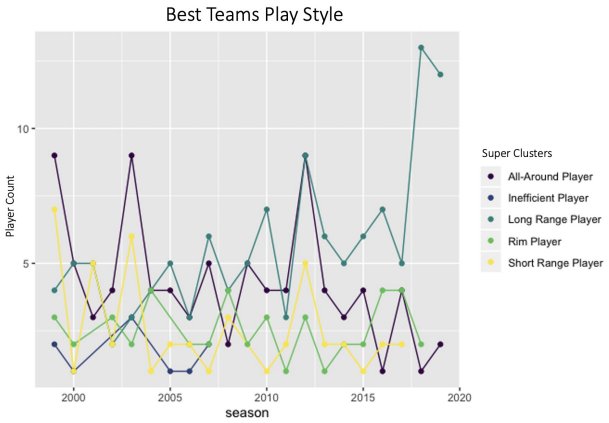


Figure 2The best teams haven't had inefficient players since the mid-2000's and have had more all-around and three-point shooters

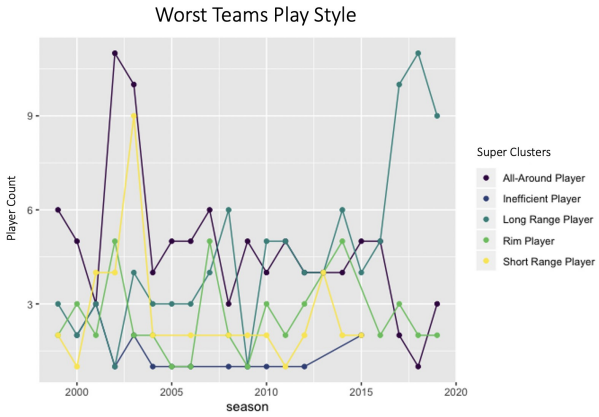


Figure 1The worst teams have had more inefficient players for longer and fewer three-point players

Conclusion

Using the finite mixture clusters we found that the obvious shift to a three-pointer driven league can be attributed to both the transition of all-around players into long-range shooters as well as the increase in number of rookies coming into the league as players who are already inclined to shoot from three-point land. We find little evidence that the changes in roster composition and playstyle have been driven by increased retirement or career risk to players who stick with an out-of-favor playstyle, although we do see some evidence of this dynamic among the population of inefficient players who opt for shooting long 2-point shots.

Rim players are the least likely to change their play style likely due to their lack of versatility in terms of shooting. With the increase of three-point shooters, we've also observed the decline in inefficient shooters. In fact, since 2016, the number of inefficient players in the league, including rookies, have dropped to zero. Moreover, the best teams have also identified the value of three-point shooters earlier and made the transition sooner than the worst teams.

In terms of next step, we envision that more sophisticated form of clustering such as a finite mixture Dirichlet Multinomial model may help improve the model fit even further. In addition, this kind of analysis could further include dimensions such as locations where passes are made/received to further gauge the overall contribution that a player makes. We do believe, over- all this modelling approach offers a principally justified way to derive player segmentation analysis.

Additionally, there are obvious opportunities to use this clustering methodology for lineup optimization or player valuation applications. Adding the player type or supercluster in a plus-minus model would be an interesting way to find the best combination of player roles. For instance, one might find that a certain star plays well with 3pt Specialists, regardless of the quality of the other player's overall quality. This might allow teams to create better lineups or identify players who would be more valuable in a certain offensive scheme or in combination with another set of players.