SESUG Paper 234-2018

# Data Driven Approach in the NBA Pace and Space Era

Thomas Ferrara, Kobie Marketing, Inc., Decision Sciences, Manager of Decision Sciences

## ABSTRACT

Whether you're an NBA executive or Fantasy Basketball owner or a casual fan, you can't help but begin the conversation of who is a top tier player? Currently who are the best players in the NBA? How do you compare a nuts and glue defensive player to a high volume scorer? The answer to all these questions lies within segmenting basketball performance data.

## OVERVIEW

A k-means cluster is a commonly used guided machine learning approach to grouping data. I will apply this method to human performance. This case study will focus on NBA basketball individual performance data. The goal at the end of this case study will be to apply a k-means cluster to identify similar players to use in team construction.

## INTRODUCTION

I'm currently a manager of decision sciences for a loyalty marketing firm located in St. Petersburg, Florida, doesn't really scream basketball fan does it? My childhood was spent in Brooklyn, New York. I'm a diehard New York Knicks fan. My formative years were spent watching my favorite team get handled by arguably the greatest basketball player of all time, Michael Jordan.

Several moments throughout my life and to this day it crosses my mind, only if we had that player on our team. Over time I have come to terms with we would never have Michael Jordan or player of his caliber, but wouldn't it be interesting if a NBA team could find complimentary parts or look-a-like players?

This is why I'm writing a paper about finding these look-a-likes, these diamonds in the rough, or as the current term is "Unicorns". Let's begin this journey together in search for a cluster of basketball unicorns.

## WATCHING THE GAME TAPE

What do high level performers have in common? In most cases you'll find they study their sport, study their own game performance, study their opponents and study the performance of other athletes they strive to be like. The data analyst equivalent to watching game tape would be to gather as many independent and dependent variables as possible to perform an analysis.

For the NBA data used in this k-means cluster analysis, I took the approach of what contributes to success in winning a game. Outscoring your opponent was a no-brainer starting point, but I'll need to dig deeper. How many ways can and what methods can you outscore an opponent? The avid basketball fan would agree how a player scores a basket (i.e. field goal vs behind the three point line) will determine how they fit into an offensive scheme and defines their game plan.

Beyond scoring there are other equally as important contributors to basketball performance. This where I began to think of how much hustle and defensives metrics could I gather (i.e. rebounds, assists, steals, blocks, etc.). Could I normalize all of these metrics to come to get a baseline on player efficiency and more importantly effectively identify an individual player's role in a team's overall performance?

To normalize my metrics I made the decision to produce my raw data on a per minute level, this way I wouldn't show biases to high usage players or low usage players. To identify how a player fits into an offensive scheme and their scoring tendencies I calculated an individual level what percent of points scored comes from all methods of scoring (i.e. free throw percentage, three pointers made, two point field goals).

Once I went through all of my data analyst game tape, I was ready to hold practice and cluster.

## HOLDING PRACTICE

Practice makes perfect, but everything in moderation (i.e. the New York Knicks of the 1990's overworked themselves during practice, they would lose steam in long games). Similar to I wouldn't want to over-fit a model on sample data, I won't get too complicated with my approach to standardizing my variables. Utilizing ***proc standard***, I'll standardize my clustering variables to have a mean of 0 and a standard deviation of 1.

After standardizing the variables I'll run the data analyst version of a zone defense (***proc fastclus*** and use a macro to create max clusters from 1 through 9). I don't anticipate to use a 9 cluster solution once running the game plan and evaluating my game time results. Ideally I want to keep my cluster size to small manageable number while still showing a striking difference between the groups.

To evaluate how many cluster I'll analyze to come to a final solution, I'll extract the r-square values from each cluster solution and then merge them to plot an elbow curve. Using ***proc gplot*** to create my elbow curve, I'll want to observe where the line begins to curve (creating an elbow).

Finally, before we're kicked off the court for another team's practice, I'll use ***proc anova*** to validate my clusters. As a validate metric I'll use the variable "ttll_pts_per_m" this should help identify the difference between a team's "go-to" option and a player whom is more of a complimentary piece at best.

## RUNNING GAME PLAN AND GAME TIME RESULTS

A k-means cluster analysis was conducted to identify underlying subgroups of National Basketball Association athletes based on their similarity of responses on 11 variables that represent characteristics that could have an impact on 2016-17 regular season performance and play type. Clustering variables included quantitative variables measuring:

perc_pts_ft (percentage of points scored from free throws)

perc_pts_2pts (percentage of points scored from 2 pt field goals)

perc_pts_3pts (percentage of points scored from 3 pt field goals)

'3pts_made_per_m'N (3 point field goals made per minute)

reb_per_min (rebounds per minute)

asst_per_min (assists per minute)

stl_per_min (steals per minute)

blk_per_min (blocks per minute)

fg_att_per_m (field goals attempted per minute)

ft_att_per_min (free throws attempted per minute)

fg_made_per_m (field goals made per minute)

ft_made_per_m (free throws made per minute)

to_per_min (turnovers per minute)

All clustering variables were standardized to have a mean of 0 and a standard deviation of 1.

Data was randomly split into a training set that included 70% of the observations (N=341) and a test set that included 30% of the observations (N=145). A series of k-means cluster analyses were conducted on the training data specifying k=1-9 clusters, using Euclidean distance. The variance in the clustering variables that was accounted for by the clusters (r-square) was plotted for each of the nine cluster solutions in an elbow curve (see figure 1 below) to provide guidance for choosing the number of clusters to interpret.

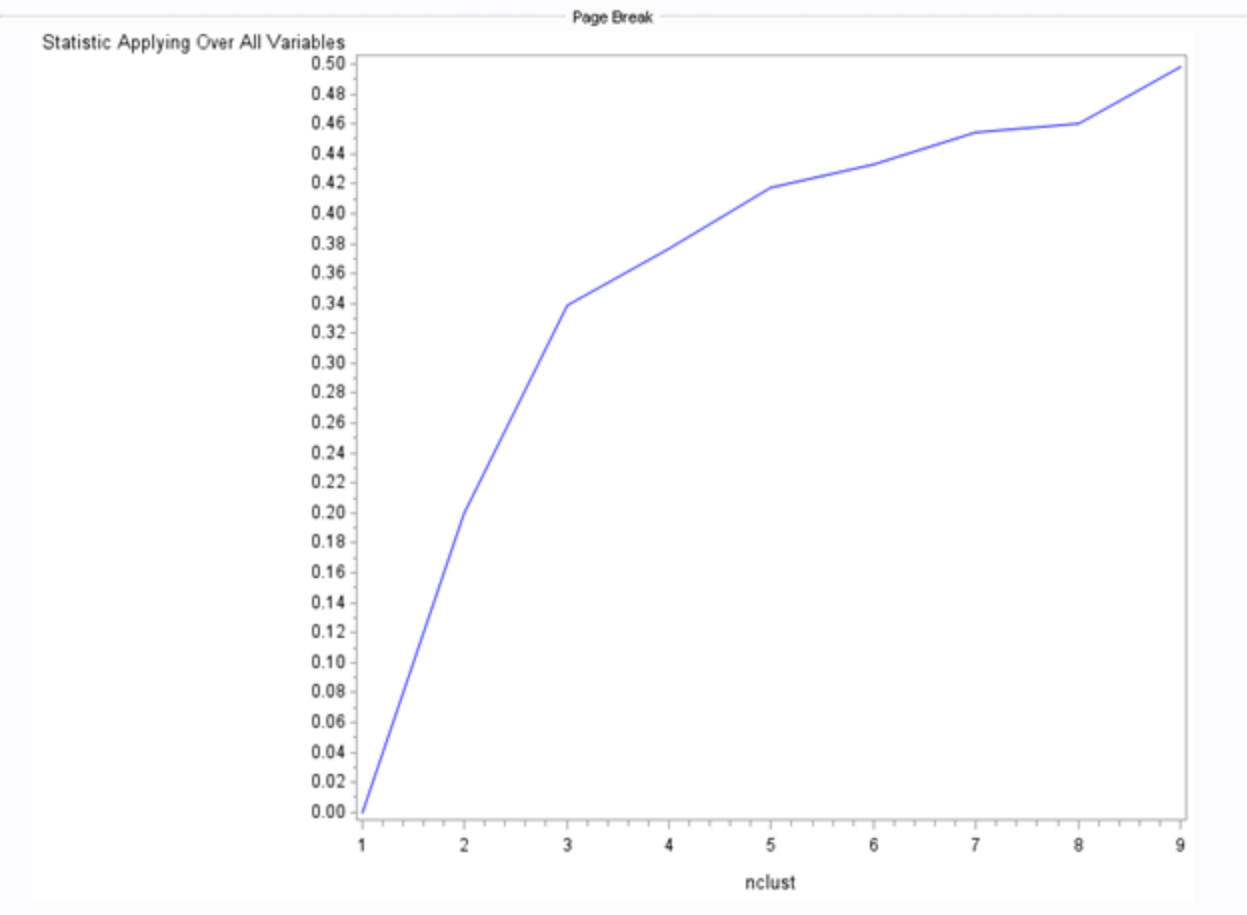


**Figure 1. Elbow curve of r-square values for the nine cluster solutions**

The elbow curve was inconclusive, suggesting that the 3, 5 and 7-cluster solutions might be interpreted. The results below are for an interpretation of the 3-cluster solution.

Canonical discriminant analyses was used to reduce the 11 clustering variable down a few variables that accounted for most of the variance in the clustering variables. A scatterplot of the first two canonical variables by cluster (Figure 2 shown below) indicated that the observations in cluster 3 is the most densely packed with relatively low within cluster variance, and did not overlap very much with the other clusters. Cluster 1's observations had greater spread suggesting higher within cluster variance. Observations in cluster 2 have relatively low cluster variance but there are a few observations with overlap.

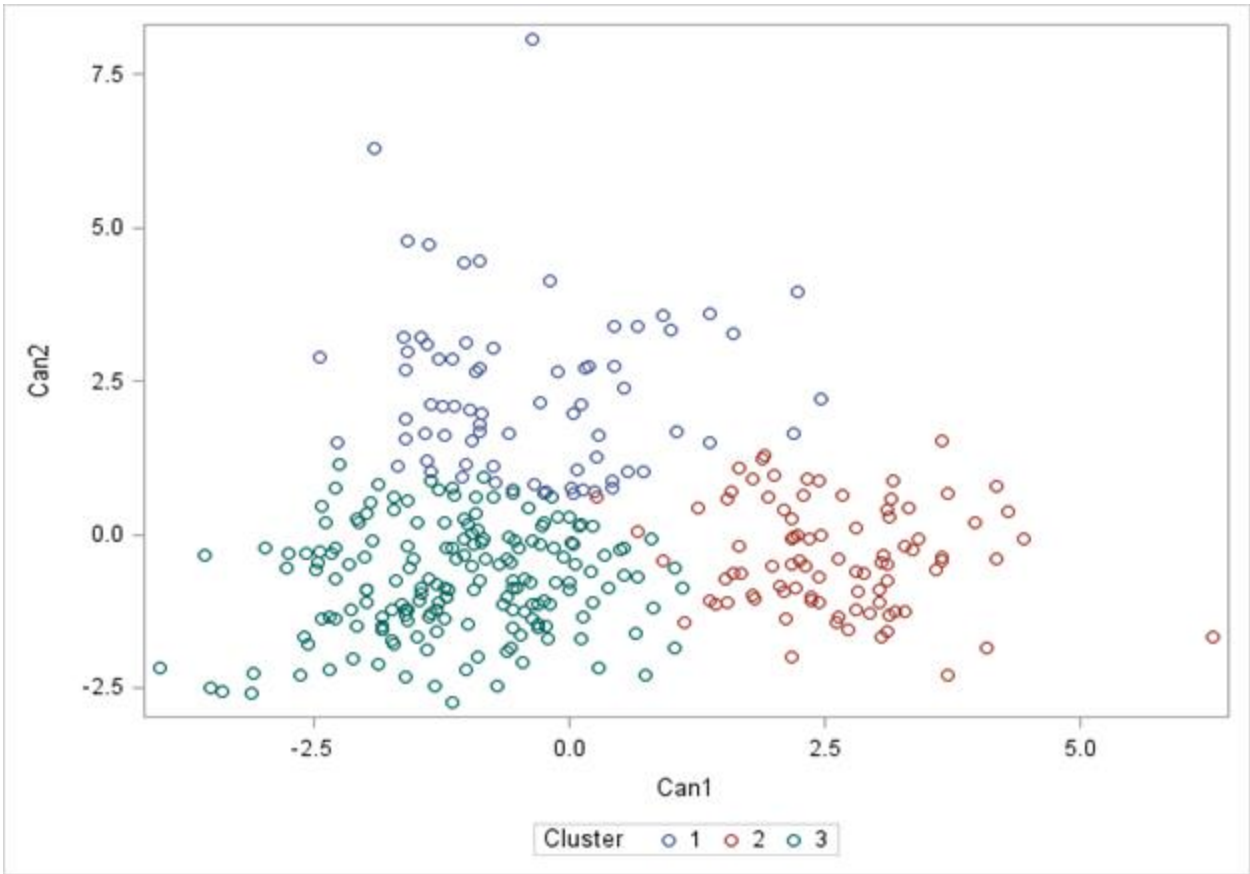


**Figure 2. Plot of the first two canonical variables for the clustering variables by cluster.**

The results of this plot suggest that the best cluster solution is 3 clusters, but due to a lack of data points there will be overlap of clusters.

The means on the clustering variables showed that, athletes in each cluster have uniquely different playing styles.

Cluster 1:

These athletes have high values for percentage of points from free throws, moderate on percentage points from 3 point field goals and low on percentage of points from 2 point field goals. These athletes attempt more field goals per minute, free throws per minute, make more 3 point field goals per minute and have the highest value for assists per minute; these athletes are focal points of a team's offensive strategy.

Athletes in this cluster:

Kevin Durant

Anthony Davis

Stephen Curry

Cluster 2:

The athletes have extremely high values for percentage of points from 2 point field goals, moderate on percentage points from free throws, and extremely low values for percentage of points from 3 point field goals. These athletes rarely make perimeter shots and have low values for assists.

Athletes in this cluster:

Rudy Gobert

Hassan Whiteside

Myles Turner

Cluster 3:

The athletes have high values for percentage of points from 3 point field goals, and low values for point 2 point field goals and free throws. These athletes stay on the perimeter (high values for 3 point field goals made) but are a secondary option at best, observed by a low field goal attempts per minute.

Athletes in this cluster:

Otto Porter

Klay Thompson

Al Horford

In order to externally validate the clusters, an Analysis of Variance (ANOVA) was conducting to test for significant differences between the clusters on total points scored per minute (ttl_pts_per_m). A tukey test was used for post hoc comparisons between the clusters. The results indicated significant differences between the clusters on ttl_pts_per_m ($F(2, 340)=86.67$, $p<.0001$). The tukey post hoc comparisons showed significant differences between clusters on ttl_pts_per_m, with the exception that clusters 2 and 3 were not significantly different from each other. Athletes in cluster 1 had the highest ttl_pts_per_m (mean=.541, sd=0.141), and cluster 3 had the lowest ttl_pts_per_m (mean=.341, sd=0.096).

## CONCLUSION

Using a k-means cluster is a data driven approach to grouping basketball player performance. This method can be used in constructing a team when a salary budget is constricted. The elephant in the room is this essentially is human behavior, therefore the validation step using proc anova is critical. The approach I've applied to the NBA data is a guide machine learning approach.

## REFERENCES

Data used in this case study was sourced from:

Basketball Reference

https://www.basketball-reference.com/

## RECOMMENDED READING

Detail on proc fastclus for a k-means clustering:

https://support.sas.com/documentation/cdl/en/statugfastclus/63675/PDF/default/statugfastclus.pdf

Detail on proc standard:

http://support.sas.com/documentation/cdl/en/proc/65145/HTML/default/viewer.htm#p08lchizi3ii8yn10lzjjegwwazv.htm

Detail on proc gplot:

http://support.sas.com/documentation/cdl/en/graphref/63022/HTML/default/viewer.htm#gplot-syn.htm

Detail on proc anova:

https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#anova_toc.htm

## CONTACT INFORMATION

Comments, questions, and additions are welcomed.

Contact the author at:

Thomas Ferrara

Kobie Marketing, Inc.
Decision Sciences
Manager of Decision Sciences
100 2nd Ave S.
St. Petersburg, FL 33701
Phone: (727)-822-5353
Email: Thomas.ferrara@kobie.com