Defensive Player Classification in the National Basketball Association

Neil Seward

University of Ontario Institute of Technology, Oshawa Ontario, Canada
neil.seward@uoit.ca

Abstract. The National Basketball Association(NBA) has expanded their data gathering and have heavily invested in new technologies to gather advanced performance metrics on players. This expanded data set allows analysts to use unique performance metrics in models to estimate and classify player performance. Instead of grouping players together based on physical attributes and positions played, analysts can group together players that play similar to each other based on these tracked metrics. Existing methods for player classification have typically used offensive metrics for clustering [1]. There have been attempts to classify players using past defensive metrics, but the lack of quality metrics has not produced promising results. The classifications presented in the paper use newly introduced defensive metrics to find different defensive positions for each player. Without knowing the number of categories that players can be cast into, Gaussian Mixture Models (GMM) can be applied to find the optimal number of clusters. In the model presented, five different defensive player types can be identified.

Keywords: defense, efficiency

1 Introduction

Traditionally, players in the NBA would be evaluated against other players playing the same position. The evaluation would only include traditional box score metrics in the evaluation, and would largely be concerned with offensive output. This evaluation methodology assumes that players in the five defined positions play the same as other players in the same position. This is not always the case. Players in the NBA are acquired for different reasons. One of the most notable differences in play style is the difference between a low-post power forward and a stretch forward. Both player styles are for the power forward position, but have vastly different offensive strategies. A stretch forward like Channing Frye (Cleveland Cavaliers) plays the outside key frequently due to his accurate three point shot. Frye plays offense differently than a low-post forward like Greg Monroe (Milwaukee Bucks). Monroe finds power in being close to the net to be able to get offensive rebounds and use his many post maneuvers to score baskets. To see the difference in offensive style, observe the shot charts for both of these players in Figure 3 and Figure 4 in the Appendix. In this analysis, we assume the same is true for defensive players. Although players exist within a set of positions, not all players of the same position play the same defensively.

1.1 Offensive Player Clustering

Lutz [2] has proposed evaluating players using a combination of moderately advanced offensive production rates and traditional box score defensive metrics. The evaluation methodology then uses Gaussian Mixture Models for cluster number estimation. With the Gaussian model, ten clusters are used to classify the players. With the players identified, team season records can be used to determine which distribution of the ten different play types is the most effective at winning. By comparing team records and player classification totals for each team, the most effective player types can be identified.

1.2 Attempts at Defensive Player Clustering

Previous attempts at clustering defensive play types have used the opponent's physical characteristics as a weight to determine how a player defends[2]. By using average guarding player height as the average weight of the player guarded, Willard estimates player defense types. By using opponent height and weight, the clustering selects player types based on the size of the player that the defensive player is guarding and how effective the player is at guarding that size of a player. This attempt presents bias against centers and power forwards. Both power forwards and centers usually stay near the hoop during while on defense. Their hoop presence allows the players to create help defense whenever another player on the team loses their man and needs help. This is an effective defensive strategy, but does not capture the point of the defensive player classification. The distribution of player sizes being defended against as a center or a forward is much different that the distribution of players as a small forward or a guard. Centers and power forwards give help defense around the rim, while other players typically do not offer help defense. Guards driving to the net are going to be contested by a center or a forward. As guards are typically small players, this bias will throw the distribution of players guarded as a center off. Giving help defense should not attribute a rim protector with defending a small guard.

2 Data

Over the last four years, the NBA has widely expanded data collection efforts. The available data has invited an explosion in research on the sport. The NBA employs several SportVU cameras that track the X,Y positions of all players on the court, as well as the X,Y, and Z positions of the ball at 25 frames per second. The tracking capabilities of the cameras allow for the collection of more advanced metrics that were never tracked in box scores. The expanded metrics can contribute to better insight in classification efforts. For defensive player classification, these new metrics have been identified to showcase defensive effort in different areas.

The NBA has recently expanded the set of publicly available defensive metrics tracked. These metrics go much farther than blocks and steals. Some of the metrics introduced include contested shots and deflections, where both of these efforts weaken efficiency in offenses. These new metrics were released as part of the playoffs last year, so there is not a full season's worth of data to do the analysis with. The clustering was done on all recorded data of the 2016-17 regular season. This type of analysis should be done with a full season's worth of data in order to provide full results.

The metrics used in this paper are:

- average minutes played min
- average defensive speed $speed_{def}$
- average shots blocked blk
- average shots contested 2 point range $contest_2$
- average shots contested 3 point range $contest_3$
- average deflections def
- average loose balls recovered lbr
- average steals stl
- average shots contested 2 point range $contest_2$
- average field goals attempted in restricted area fga_{ra}
- average field goals attempted in non restricted area fga_{nra}
- average field goals attempted in mid range fga_{mr}
- average field goals attempted in left corner three fga_{lc}
- average field goals attempted in right corner three fga_{rc}
- average field goals attempted in above key three fga_{ak}
- height of player $height$

2.1 Time Frame of Metrics

Not all players average the same amount of play time during a game and not all players play the same amount of games. All players that played less than 5 games and averaged less than 10 minutes a game were ignored in the classification. In order for all players to be evaluated within the same time period, each of the tracked averages need to be converted to a 40 minute frame. Let each set of metrics for a player be defined as

$$f(x_{player}, min_{player} | x \in X^n) = \sum_{i=1}^{n} \frac{40x_i}{min_{player}} \tag{1}$$

where x_{player} is all metrics tracked for a player that belong to all tracked metrics X of size n and min_{player} is the average number of minutes played for the player.

2.2 Converting to Efficiency

Although the metrics listed would be useful to classify defensive play types, there is no emphasis on efficiency. Knowing how many times a player contests shots is useful, but it does not reflect how efficient the player was at contesting the shots that they faced. Not all metrics used in the classification can be converted to reflect efficiency. The metrics that can be converted to efficiency are listed below.

Let the average blocks made against all field goal attempts, blk_{fga}, defended be defined as

$$blk_{fga} = \frac{blk}{fga_{ra} + fga_{nra} + fga_{mr} + fga_{lc} + fga_{rc} + fga_{ak}} \tag{2}$$

Let the average contest made against all two point field goal attempts, $contest_{2fga}$, defended be defined as

$$contest_{2fga} = \frac{contest_2}{fga_{ra} + fga_{nra} + fga_{mr}} \tag{3}$$

Let the average contest made against all three point field goal attempts, $contest_{3fga}$, defended be defined as

$$contest_{3fga} = \frac{contest_3}{fga_{lc} + fga_{rc} + fga_{ak}} \tag{4}$$

2.3 Normalization

The Gaussian Mixture Model is used to estimate the number of clusters with normalized data [3]. The metrics used in the model need to be modified to get the z-score. Let each normalized metric x be defined as

$$f(x|x \in X_{player}) = \sum_{i=1}^{n} \frac{(x_i - \mu_i)}{\sigma_i} \tag{5}$$

where x are all metrics tracked that belong to player-specific metrics X_{player} of length n, μ_i is the mean of metric x_i, and σ_i is the standard deviation of metric x_i.

3 Gaussian Mixture Models

The Gaussian Mixture Model(GMM) is a probability density function that is modeled as a weighted sum of a number of individual Gaussian probability distributions [3], [4], [7], [8]. The model is used to estimate the number of clusters present in a multi-variant distribution. One of the benefits of using GMM when clustering is that the assignments are same every time the model is applied.

The results do not have an element of randomness to them. The parameters and weights used for the model are created from iterations of the Expectation Maximization algorithm [4]. The GMM can be defined as

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \sum_i) \tag{6}$$

where x is the vector representation of a basketball player, w_i is the representation of the mixture weights in the total weight space of M weights, and $g(x|\mu_i, \sum_i)$ represents the Gaussian distribution i.

3.1 Expectation Maximization

Given a sequence of length T of basketball players $X = x_1, ..., x_T$, the GMM can be applied to assign clusters in a single iteration as

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \tag{7}$$

There can be no maximization of the assignments in just one iteration of the training set. In order to maximize the probability of each assignment, the optimal parameters for the model must be obtained through iterations of the Expectation Maximization (EM) algorithm [2]. Both the means μ, and mixture weights w across the mixtures $i = 1, ..., M$ can be optimized using these parameter optimization functions.

$$\bar{w}_i = \frac{1}{T} \sum_{t=1}^{T} Pr(i|x_t, \lambda) \tag{8}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} Pr(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} Pr(i|x_t, \lambda)} \tag{9}$$

These parameter optimization methods will keep updating the parameters until a threshold is met during the parameter changes. Once the optimal parameters are determined, the new model $\bar{\lambda}$ will satisfy $p(X|\bar{\lambda}) \geq p(X|\lambda)$ and create optimal assignments of basketball players among the mixtures of different types.

3.2 Bayesian Information Criterion

The number of M mixtures used in the GMM can be infinite. In order to select the optimal number of clusters, or mixtures, Bayesian Information Criterion(BIC)[5] can be used. BIC is used to evaluate the changes between the different models created with different cluster numbers. The score evaluates the maximum assignment likelihood for all models specified. Using this score across a subset of models, an optimal number of clusters can be determined for the GMM. The BIC score across 1 to 10 cluster GMM's are shown in Fig. 1. From the BIC scores across the different clusters, 5 clusters were chosen as the number of possible defense types. The maximum likelihood is achieved at 2 clusters, but will not offer the type of descriptions in the cluster assignments that having 5 clusters does.

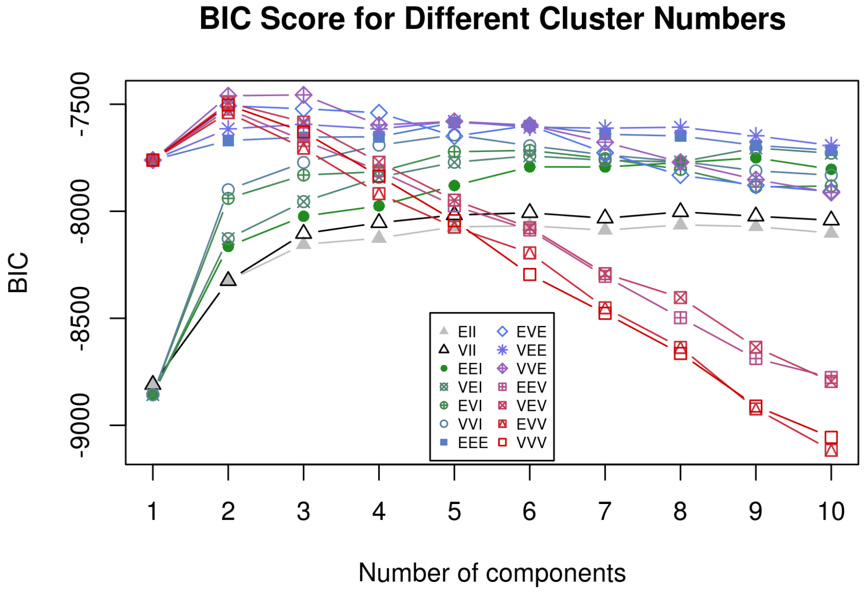


Fig. 1. BIC Score for 1-10 Clusters in GMM

4 Results

The Gaussian Mixture Model creates cluster assignments for five types of defense present during the regular season. In Table 1, summaries of each cluster is present. Each cluster has means associated with each metric tracked. The scatter plot containing the cluster assignments can be seen in Fig. 2.

From the results in Table 1 we can give proper names to each of the styles. The first player type is excellent in contesting two point shots, getting blocks, getting rebounds, and is very tall. The player profile is slow and does not contest three point shots. The first play type would be identified as a Rim Protector. A player that fits this profile is Andrew Bogut.

The second player type excels at defending three point shots and two point shots, as well as getting defensive rebounds. The player is typically taller than

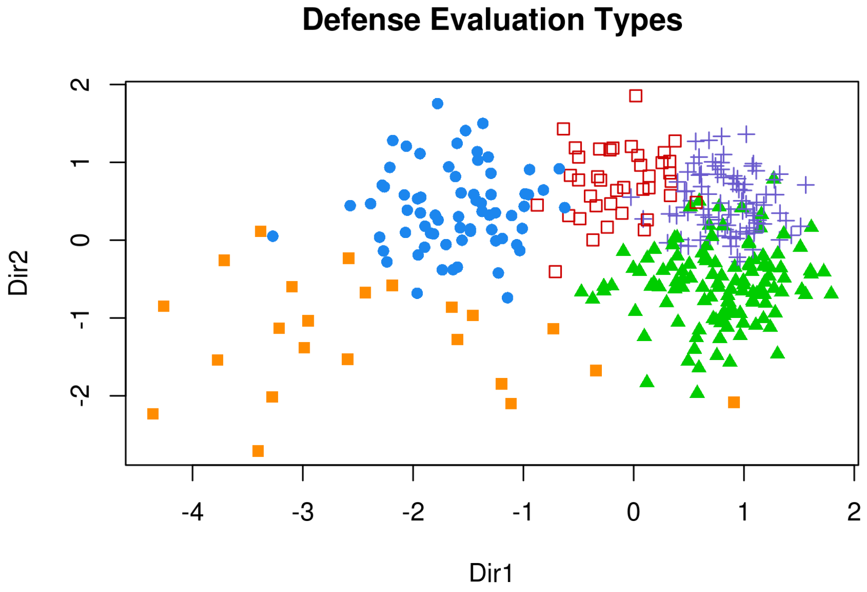


Fig. 2. Defensive Cluster Assignments Using GMM

type	*contest*$_{2fga}$	*contest*$_{3fga}$	*blk*$_{fga}$	*def*	*stl*	*lbr*	*drb*	*speed*$_{def}$	*height*
1	1.21	-0.81	0.79	-0.48	-0.49	-0.29	0.88	-0.23	1.05
2	0.24	0.53	-0.27	-0.39	-0.35	-0.14	0.52	-0.24	0.66
3	-0.60	0.45	-0.37	0.55	0.48	0.41	-0.46	0.23	-0.82
4	-0.69	-0.09	-0.61	-0.44	-0.28	-0.39	-0.59	0.07	-0.19
5	1.66	-0.64	2.23	0.64	0.47	0.26	1.05	-0.35	0.81

Table 1. Defensive Cluster Information

average and can be cast as a Wing Defender. A player that would fit this profile is Richard Jefferson.

The third player type can be categorized as a player that excels at getting deflections, getting loose balls, getting steals, is quick on defense, and is short. The third play type can be identified as an Elite Defensive Ball Handler. An example of a player that fits this profile would be Kemba Walker.

The fourth play type classification categorizes players as ones that do not particular stand out in any one area. The fourth play type can be labeled as a Below Average Defender. A player that fits this category would be D'Angelo Russell.

The fifth play type describes players unique players that would typically fit the profile of a Rim Protector, but have an expanded skill set to get deflections, steals, and loose balls. With the Rim Protecting skills, these players even do better than the typical Rim Protectors. These players are untouchable in getting blocks, getting defensive rebounds, and contesting inside shots. These players can be categorized as Versatile Rim Protectors. One player that fits this profile is Joel Embiid.

5 Future Works

Offensive shot charts are extremely effective at describing offensive direction for all players. Using these charts, which can be seen in the Offensive Shot Chart section of the Appendix, weak and strong shooting areas can be defined. Franks et. al [6] have creating two metrics from player tracking data that are very useful in evaluating players. Using Markov modeling, the model they have created can tracked each offensive player and each of the defensive assignments for the defending team. With information on which player is guarding which player during defensive possessions, defensive players can finally have shots attributed to them without presenting bias to large players that are on help defense.

The metrics proposed under the model are disruption score and volumetric score. Disruption score measures the tenancies of the defending player to force the offensive player to shoot at a worse rate in locations than their expected rate at that location. An inflated example would be a defensive player like Kawhi Leonard guarding Lebron James on the wing, and forcing him to shoot sub 30% at the right corner three spot, when Lebron would be shooting an expected 60% from the location. Kawhi would be attributed with having a high disruption score in that region. Volumetric score is focused comparing on number of attempts made for an offensive player when guarded by a specific player against their expected attempt frequency during the same time. Using the same example as before, if Kawhi can force Lebron to shoot 25% less from the corner three, then Kawhi would be attributed with a low volumetric score. The aim of these metrics is to identify players that have a high disruption score and a low volumetric score. These metrics are useful when creating defensive shot charts. See the Appendix for examples of Defensive Shot Charts.

6 Conclusion

Offensive player classification can assist coaches and general managers in the NBA with evaluating players on other teams. It allows them to gain better insight into players that are not well known stars and get high box scores to backup their stardom. The different evaluation classes can assist management in finding niche players that fit into the organization's offensive direction. The same can be done with defensive player classification. With expanded tracking metrics, players can have a wide array of skills recorded and evaluated to describe their effort on defense.

When clustering a set of data with no knowledge of the initial number of clusters, a Gausian Mixture Model(GMM) can be applied to maximize the assignment probabilities across multiple Gaussian distributions. In order to select the optimal number of clusters, Bayesian Information Criterion (BIC) scores can be calculated to observe the change in assignments as mixtures are added or taken away from the GMM.

After applying the GMM to the defensive metrics, the model finds five distinct defensive categories for players. The categories are: Rim Protector, Wing Defender, Below Average Defender, and Versatile Rim Protector. These player categories are created to give accurate descriptions of players on defense outside of the typical box score evaluations. By using clustering in evaluations of defense, new player types can be found. Using this methodology, general managers and coaches can have the ability to gain insight into how each player performs on defense and to which category each player would fit into on defense.

References

1. Lutz D.: A Cluster Analysis of NBA Players. MIT Sloan Sports Analytics Conference. (2012)
2. Willard J., NBA Positions by Clustering, Nylon Calculus, 2016. [Online]. Available: http://nyloncalculus.com/2015/09/29/nba-positions-by-clustering/. [Accessed: 12-Dec- 2016].
3. Rasmussen C.E: The Infinite Gaussian Mixture Model. NIPS. 12, 554–560 (1999)
4. Reynolds, D.: Encyclopedia of Biometrics, 1st ed. New York: Springer, pp. 659-663 (2009).
5. Chen S.: Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, Proc DARPA Broadcast News Transcription and Understanding Workshop. 8, pp. 127-132 (1998)
6. Franks A., Miller A., Bornn L., Goldsberry K.: Counterpoints: Advanced Defensive Metrics for NBA Basketball, MIT Sloan Sports Analytics Conference, (2015)
7. Fraley C., Raftery A. E., Murphy B. T., Scrucca L.: mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597, Department of Statistics, University of Washington, (2012)
8. Fraley C., Raftery A. E.: Model-based Clustering, Discriminant Analysis and Density Estimation Journal of the American Statistical Association 97, pp. 611-631, (2002)
9. http://nbasavant.com/apps/compare.php

A Appendix

A.1 Offensive Player Shot Charts

On offense, players in the same positions do not always play the same. In Figures 5 and 6, a prime example of different styles are presented. Greg Monroe is shown to play inside the key, while Channing Frye stays outside the three point line to stretch the floor and get three point shots.

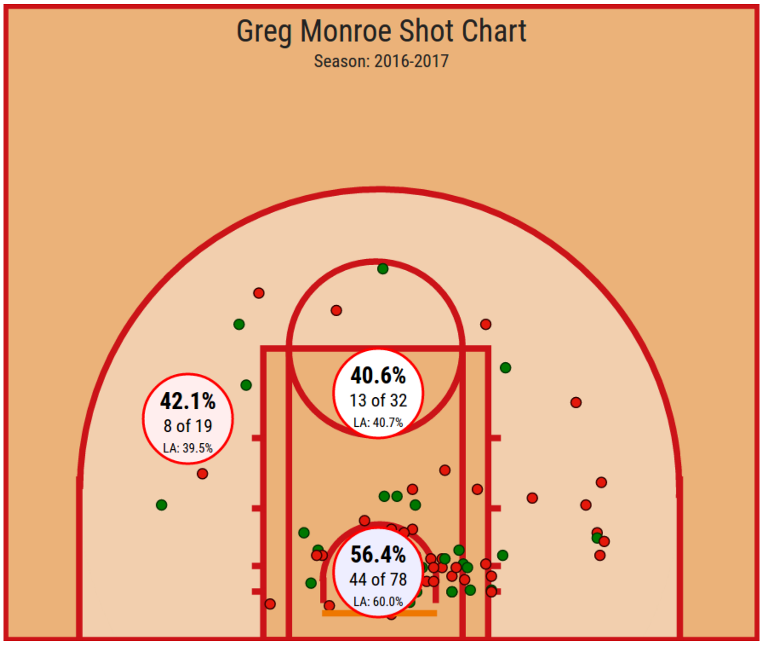


Fig. 3. Greg Monroe Offensive Shot Chart[9]

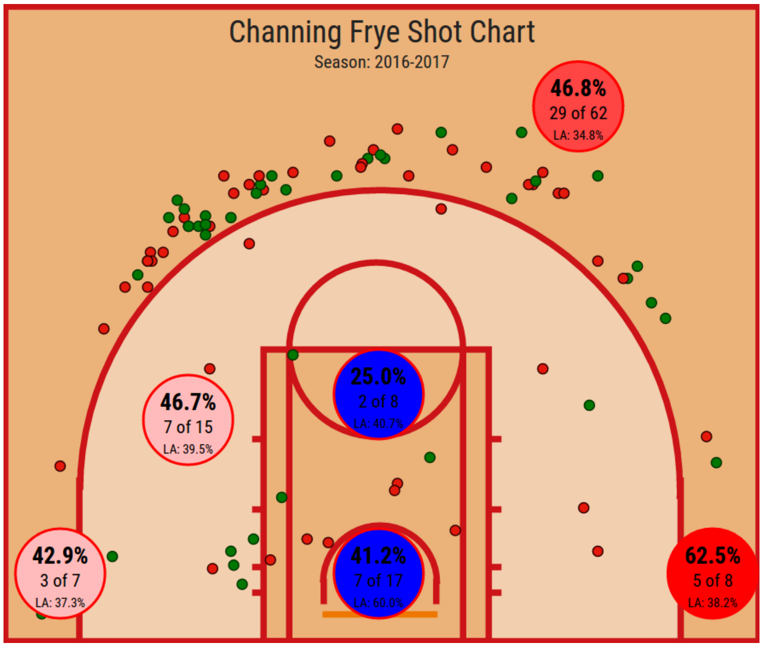
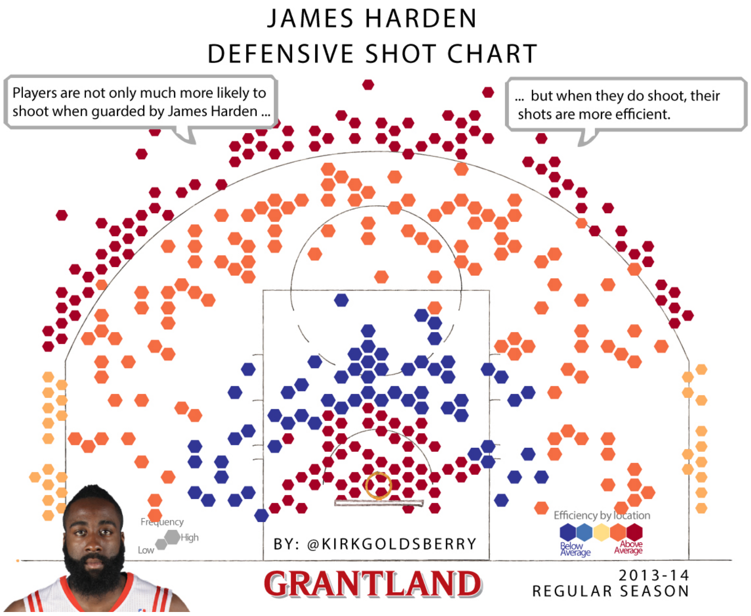


Fig. 4. Channing Frye Offensive Shot Chart[9]

A.2 Defensive Player Shot Charts

As seen in Figure 5 and Figure 6, volumetric score can be used to increase the size of ticks, while disruption score can be used to change the colour of ticks. As ticks become warmer in colour, the defender has a higher disruption score in that area.



JAMES HARDEN
DEFENSIVE SHOT CHART

Players are not only much more likely to shoot when guarded by James Harden ...

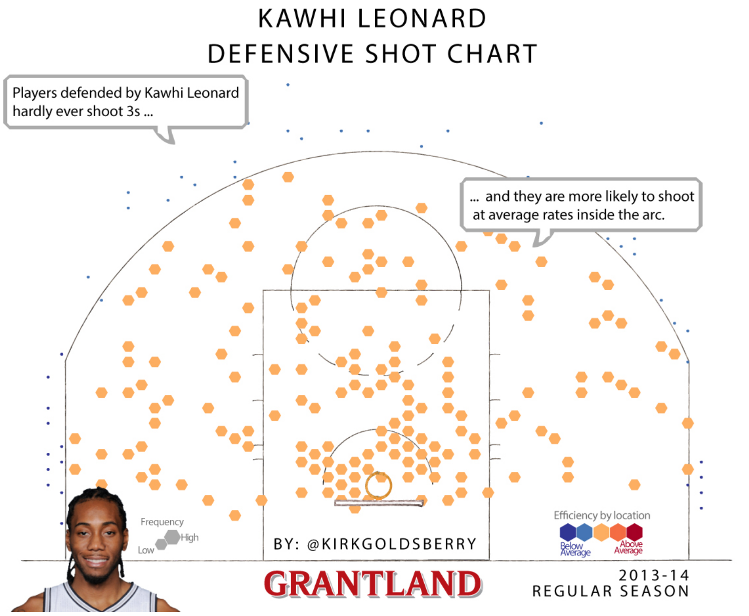
... but when they do shoot, their shots are more efficient.

BY: @KIRKGOLDSBERRY

GRANTLAND

2013-14
REGULAR SEASON

Fig. 5. James Harden Defensive Chart[6]



KAWHI LEONARD
DEFENSIVE SHOT CHART

Players defended by Kawhi Leonard hardly ever shoot 3s ...

... and they are more likely to shoot at average rates inside the arc.

BY: @KIRKGOLDSBERRY

GRANTLAND

2013-14
REGULAR SEASON

Fig. 6. Kawhi Leanord Defensive Shot Chart[6]