Clustering NFL Quarterbacks using QBR Distributions

Douglas S. Anderson

Department of Mathematical Sciences
Montana State University

May 6, 2016

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Douglas S. Anderson

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date

Mark Greenwood
Writing Project Advisor

Date

Mark Greenwood
Writing Projects Coordinator

Abstract: Clustering NFL Quarterbacks using QBR Distributions

Quarterbacks are becoming the face of most NFL teams. QBR, a metric designed by ESPN, has been used since its creation in 2011 to compare the efficiency (and impact) a quarterback had for his team in a given game. A method of functional data analysis, inspired by the work by Noah Davis and Michael Lopez, has been used to cluster quarterbacks using the QBR distributions of each quarterback. Methods were first used to attempt to replicate the work of Davis and Lopez, which involved the k-means clustering algorithm. Other methods were proposed to better compare the estimated QBR densities using L2 and Kullback-Leibler distances. The results suggest five clusters of QBR distributions were present that were described as “Elites”, “Second-tiers”, “High Expectations”, “Some-potential”, and “Still Around”.

Acknowledgements

There is a great deal of thanks that should be directed towards my family for their support, encouragement, and love. I would also like to thank my professors through the years for helping me develop an analytic mind. An additional thanks must be given to Dr. John Borkowski, who helped make it possible for me to be in the masters program in statistics at Montana State University when I was lacking direction. Lastly, but not least, I would like to thank my advisor, Dr. Mark Greenwood, who has been an endless resource of R packages and assistance for this research.

1 Introduction

Jameis Winston and Marcus Mariota, quarterbacks from Florida State University and the University of Oregon, respectively, were picked first and second overall in the 2015 National Football League (NFL) draft. Prior to the 2015 draft there have been 85 players taken first overall since 1936 (see Figure 1). Of the 85 players drafted, 31 have been quarterbacks; the next most being running-backs, of which there were 12.

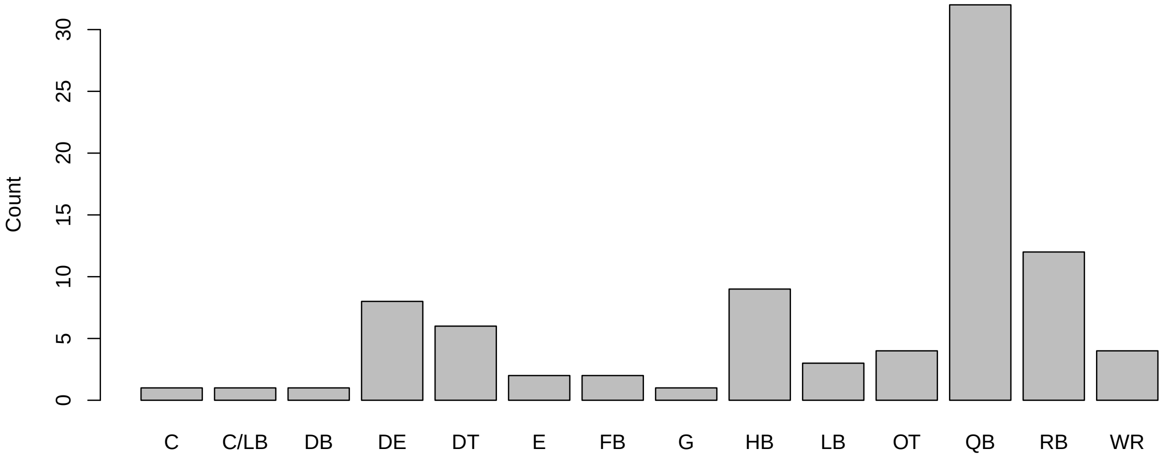

Figure 1: *The positions of the first pick taken in the NFL draft from each year, dating back to 1936. Data were obtained from [1]. The positions are center (C), center/linebacker (C/LB), defensive back (DB), defensive end (DE), defensive tackle (DT), switchable end (E), fullback (FB), guard (G), halfback (HB), linebacker (LB), offensive tackle (OT), quarterback (QB), runningback (RB), and wide receiver (WR). Some of the positions have changed over time, thus the reason there are DEs and Es as well as halfbacks and runningbacks.*

The quarterback has become, arguably, the most sought after position by NFL teams. This is further evidenced by Figure 2, which shows beanplots (Kampstra, 2008) of the top ten salaries of players at each position from the 2014 season. Notice how much higher, on average, the top ten quarterbacks were paid in comparison to runningbacks who have been taken first overall the second most among positions. Although there were a couple of right tackles that were being paid very lucratively, as well as a well-paid outside linebacker in a 3-4 defensive system, on average, the top ten quarterbacks were being paid close to 10 million dollars more per year.

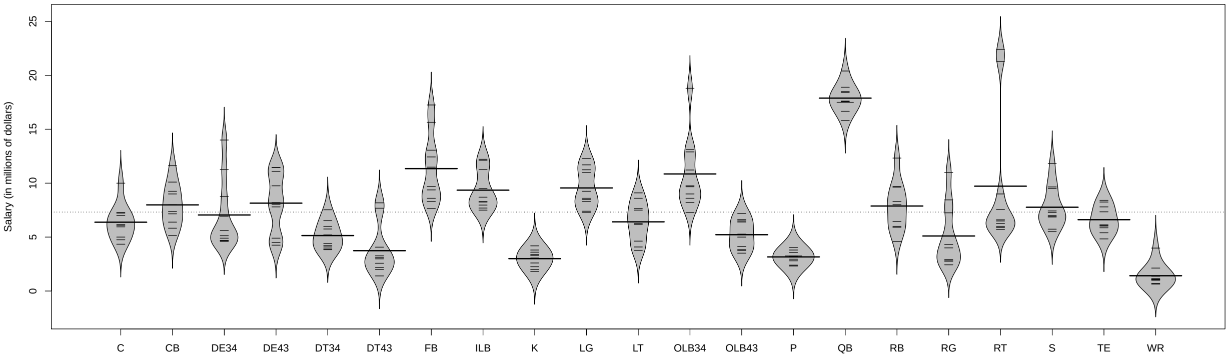

Figure 2: *Beanplots of the top ten salaries of players at each position in 2014. Data were obtained from [2].*

Due to the high value placed upon quarterbacks, a great amount of attention is given to the position (both for fans and team managers). As a natural outcome of this focus, methods have been created to assess and compare quarterback performance. In 2011, ESPN (the Entertainment and Sports Programming Network) designed a new metric called the Total Quarterback Rating, or QBR for short. ESPN has not released the exact formula for calculating QBR due to it being proprietary information; however, what has been released about the formula is that it takes every play of a team's offense into account when calculating the QBR for any quarterback. Dean Oliver, at ESPN, wrote a guide to the rating (Oliver, 2011) which illustrates some of the background behind the metric. The reader is encouraged to read Oliver's explanation if interested in learning more about the measure, which uses terms such as “win probability”, “expected points”, “division of credit”, and “clutch index”. What is important about QBR for understanding this paper is that it is a measure that has been re-scaled to be between 0 and 100, that a higher rating is better, and that a QBR of 100 is a perfect performance for a quarterback.

1.1 Validity of QBR?

There are a lot of critics of ESPN's QBR metric. A great deal of criticism of the metric is due to the hidden nature of the calculations. Although Oliver does provide an explanation of the concepts the group at ESPN contemplated for creation of the metric, it does not provide enough information on how QBR is *actually* computed. Another criticism is due to the subjective nature of many of the concepts. How are the weightings for each play determined? How was a throw good or bad in this situation but in another situation it was the opposite? For a time, some critics were bewildered that Charlie Batch, the backup quarterback for Ben Roethlisberger in 2010, had the best QBR game ever with a score of 99.9 (Smith, 2015). The results of which appeared to have vanished or changed from the ESPN database as it was not found upon searching. Despite the many concerns with regards to this metric, this report will not explore the validity of this metric.

1.2 Purpose of this Analysis

Davis and Lopez (Davis & Lopez, 2015) wrote an article entitled "The 10 Types Of NFL Quarterback" for the website FiveThirtyEight in January of 2015 that serves as the inspiration for this analysis. Davis and Lopez took NFL quarterbacks and grouped them into 10 categories based on their respective QBR density curves. We investigate this idea further. The steps of Davis and Lopez's analysis are replicated and then the clusters and methods used to find the clusters are explored further.

2 Data

2.1 Source

ESPN makes the game-level QBR data available to the public [3]. The QBR of a quarterback is measured for each game and ESPN stores it by the week (one game per week) of the season. Only the quarterback and his rating were used for the purpose of this analysis. In an attempt to be consistent with Davis and Lopez, only results from quarterbacks on regular season games between 2006 and 2014 were included. It was unclear if postseason results should be included and thus were left out. The article was written in January 2015 so it is possible that the 2014 season may not have been fully available for Davis and Lopez.

2.2 Scraping

QBR data are available publicly from ESPN. However, it is available from each week in each season and not separately for each player. For convenience and simplicity, the R package `rvest` (Wickham, 2015) was used to scrape the data from their publicly available websites. For more information on scraping data, see the sample code availaible at R-bloggers (Rudis, 2014).

2.3 Processing

The initial data set contained 5,107 QBR results from 160 quarterbacks. Some of the quarterbacks included were backups and only had one or two games played. To be consistent with Davis and Lopez, the data were further reduced to include only the 46 quarterbacks they discussed. This, for the most part, meant including only quarterbacks who had at least 10 starts over the previous two years or 50 career starts since 2006. See Davis & Lopez (2015) for a list of quarterbacks included (it is unclear which D. Carr was included, though it was our belief it was David and not Derek). The final dataset, which may or may not be the same data Davis and Lopez worked with, included 3,202 QBR observations, total, on 46 quarterbacks.

3 Densities

Non-parametric kernel-density estimators are one way to approximate the distribution of a random variable. It helps to think in terms of histograms to understand them as Härdle and Simar (2012) write “Histograms are density estimates.” If the variable of interest is measured in some fashion (quantitative) then it can be appropriate to use a histogram to display the data. In contrast to a barplot or bargraph (see Figure 1) a histogram is not naturally grouped by some characteristic. In Figure 1, the bars were given to each position that was drafted. To make a histogram for a quantitative variable, the data must be binned in some fashion. For example, the QBRs for Peyton Manning, who recently retired from the NFL, are displayed in Figure 3 with a density curve that is discussed below.

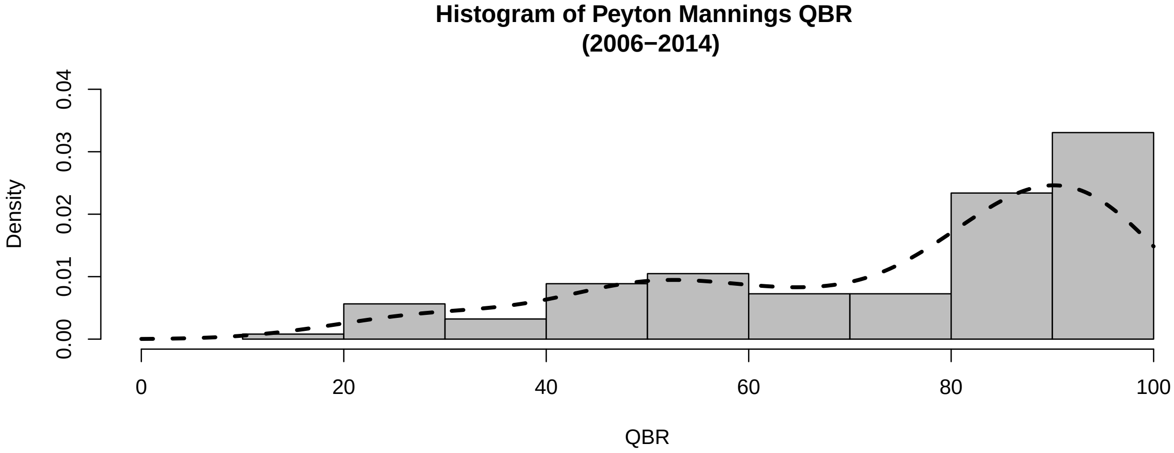

Figure 3: *Histogram (gray bars) of QBR for Peyton Manning from 2006 to 2014 with a density curve (dashed line) superimposed.*

The function `hist()`, in base R (R core team, 2016), created 11 break points defining 10 mutually exclusive and exhaustive bins of QBR values. The observations were then counted in each bin with the results shown in Table 1. Assuming bars are all of equal width, the choice of the bandwidth h (or width of bins) acts as a smoothing parameter. If h is too large, it leads to too few blocks and over-smoothing, while h too small can create false peaks and little summarization, as evidenced by Figure 4.

Bin	Bin Count	Density
[10,20]	1	0.01
(20,30]	7	0.06
(30,40]	4	0.03
(40,50]	11	0.09
(50,60]	13	0.10
(60,70]	9	0.07
(70,80]	9	0.07
(80,90]	29	0.23
(90,100]	41	0.33

Table 1: *The number of games that P. Manning had a QBR in a bin chosen by the* `hist()` *function and the resulting density of the bin obtained by dividing by the total number of games recorded for P. Manning between 2006 and 2014. By default the bins are right-closed, meaning a 30 would put the observation in the 20-30 grouping.*

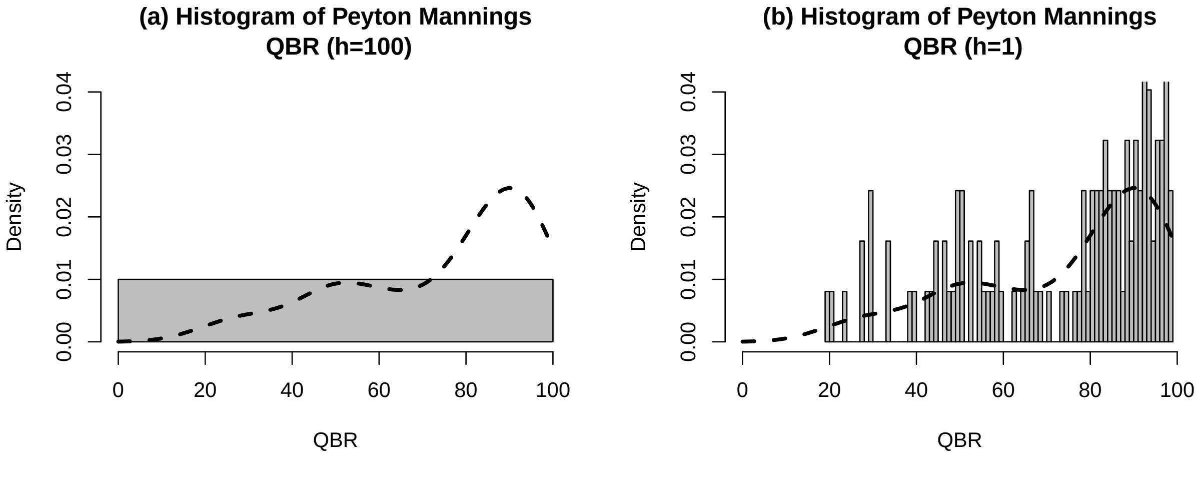

Figure 4: *Histograms of QBR for Peyton Manning from 2006 to 2014 with density curves superimposed using a bandwidth of (a) 100 and (b) 1.*

Figure 4 (a) with a bandwidth of 100, gives the illusion that P. Manning was equally likely to have a game that results in a QBR of 30 or lower as a QBR of 70 or higher (which is clearly not the case). Figure 4 (b) with a bandwidth of 1, on the other hand, makes it appear that there is a non-zero probability of P. Manning having a performance that results in a QBR between 62 and 63 but zero probability of P. Manning having a game with a QBR between 60 and 61. There is little to indicate that his QBR should act accordingly! In comparison, Figure 3 shows a histogram with a bandwidth of 10 that appears to be more reflective of what the distribution of P. Mannings QBR could look like, showing that this default bandwidth gave a more reasonable choice than either of the choices in Figure 4.

Härdle and Simar (2012) present some difficulties with histogram estimation: (1) the choice of binwidth, (2) the choice of bin origin, (3) the loss of information from replacing observations with interval-centered points, and (4) the underlying density is usually assumed to be smooth yet that is not the case with histograms. The distribution using a histogram can be computed via

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} I\left(|x - x_i| \leq \frac{h}{2}\right), \tag{1}$$

The `I()` in Equation 1 is called an indicator function where it is 1 if the observation is within half the bandwidth of the center point of the interval and 0 otherwise. Those ones and zeroes are then summed up for all the bins and then divided by the number of bins times the binwidth size to determine the density function for any x.

Instead of using boxes (each observation in a range) as a way to build the distribution, a smooth kernel function can be used. This approach may aid in avoiding one of the difficulties presented, but is still subject to the choice of bandwidth. We can modify Equation 1 by defining $K(u) = I(|u| \leq \frac{1}{2})$ (Härdle & Simar, 2012), which is the general form of a kernel estimator, and use it to provide a density estimator of

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2}$$

It is important to note that the density function is scaled so that the total area under the curve is 1. This is consistent with the definition of a probability distribution function. All of the densities for the QBRs of each quarterback are also defined on the range of 0 to 100. There are still decisions to be made for the densities as the analysis moves forward: the choice of kernel, the bandwidth, and the evaluation grid of the density estimate for clustering.

3.1 Choice of Kernel

There are seven choices of kernel available in R's `density` function: "gaussian", "epanechnikov", "rectangular", "triangular", "biweight", "cosine", and "optcosine". Figure 5 displays the estimated density for the 14 QBRs of Blake Bortles from 2006 to 2014 using each of these kernels with a bandwidth of $h = 5$. The rectangular kernel is the least smooth of the choices but all of the kernels appear to find similar estimates of $\widehat{f}(x)$. The default choice in R is to use the Gaussian kernel, due to its smooth behavior and common usage in many fields. The kernel types are summarized in Table 2. As evidenced by Figure 5, the choice of kernel

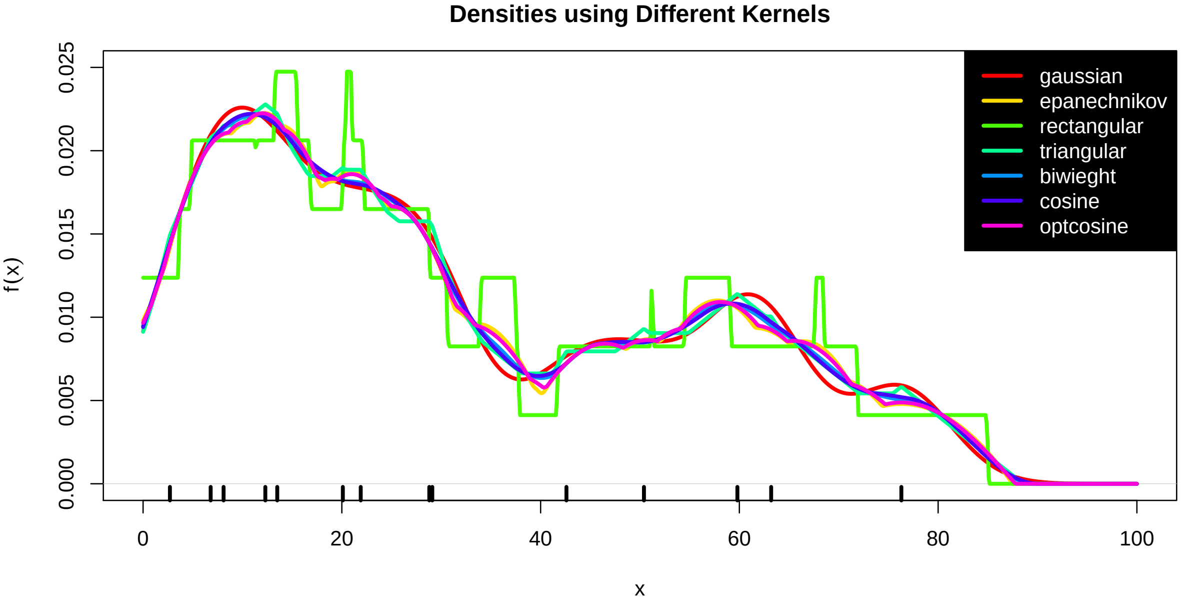

Figure 5: *The estimated density of Blake Bortles QBR from 2006-2014 using each kernel. A rug plot (bottom of plot) was used to show the 14 observed QBRs.*

would not appear to matter as each formed a similar estimate, except for rectangular. Due to its common usage and by reviewing the provided estimate densities in the article, we assume that Davis and Lopez used Gaussian kernels. They certainly did not use a rectangular kernel.

$K(\bullet)$	Kernel
$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u}{2})$	Gaussian
$K(u) = \frac{3}{4}(1 - u^2)I(\lvert u\rvert \leq 1)$	Epanechnikov
$K(u) = \frac{1}{2}I(\lvert u\rvert \leq 1)$	Rectangular (or Uniform)
$K(u) = (1 - \lvert u\rvert)I(\lvert u\rvert \leq 1)$	Triangular
$K(u) = \frac{15}{16}(1 - u^2)^2 I(\lvert u\rvert \leq 1)$	Biweight (or Quartic)

Table 2: *Formulas for the more common kernels as described by Härdle and Simar.*

3.2 Choice of Bandwidth

There are many rules of thumb about the choice of bandwidth. One idea for the Gaussian kernel is to use $h_G = 1.06\widehat{\sigma}n^{-1/5}$, where $\widehat{\sigma}$ denotes the sample standard deviation of the responses and n the number of observations, and to use $h_Q = 2.62 \cdot h_G$ for the quartic kernel (Härdle & Simar, 2012). Venables and Ripley (2002) suggest a slight adjustment to the bandwidth with $h = 1.06 \cdot \min(\widehat{\sigma}, R/1.34) \cdot n^{-1/5}$ where R is the range of the sample values, which they found wasn't always desirable. Härdle has other suggestions and explanations of bandwidth choices (Härdle, 1990). Delicado (Delicado, 2007) argues that if the bandwidth sizes or evaluation points of the density curves differ between the density estimates when considering multiple estimated densities then the comparison of the estimated probability distributions would have some sort of dependence on either selection.

The bandwidth plays a large role in the shape of the estimate. Too small a bandwidth and the estimate is too choppy, but too large and the estimate loses information, such as modality, as evidenced in Figure 6. Using Bortles' QBRs, and code similar to Everitt and Hothorn (2011), a Gaussian kernel was used to estimate the density with bandwidths of size 1, 10, and 30. Using the bandwidth of 1 (Figure 6 (a)) results in a very separated overall density estimate (bold line) which would be illogical. The bandwidth choice of 10

(Figure 6 (b)) seems to find a fairly smooth estimate but the bi-modality is still captured. In comparison, the bandwidth of 30 (Figure 6 (c)) creates a very smooth estimate and does not capture the bi-modal feature that the bandwidth of 10 captured.

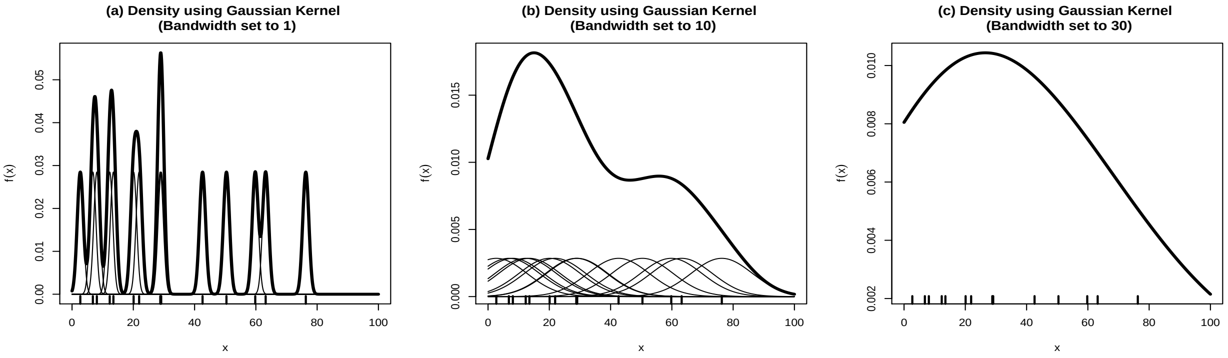

Figure 6: *Density estimates (bold lines) for Blake Bortles QBR using bandwidths of (a) 1, (b) 10, and (c) 30. Below the overall density estimates are the Gaussian-kernel, building blocks that are summed to create the overall density estimates. The building blocks are extremely flat in (c) due to the large bandwidth choice making them difficult to see.*

3.3 Choice of Grid Resolution for Density Evaluation

Given a smooth density curve, it is often necessary to evaluate the curve on a discrete grid of points. The default grid of evaluation points of a density estimate in R is 512 evenly spaced points from the minimum to maximum values that need to be defined; if these are not specified, then the smallest value used is the minimum observation minus three times the bandwidth and and largest value used is the maximum observation plus three times the bandwidth. The choice of size makes little impact on the shape of the estimate if it is over 20, as evidenced in Figure 7. In Figure 7 (a), density estimates for Blake Bortles QBR are displayed using a Gaussian kernel and bandwidth of 10 with the number of evaluation points ranging between 100 and 1000 (by 100). All of the density estimates overlap each other, thus to see the estimated densities they were shifted down (from the estimate formed using 1000 evaluation points) proportional to their grid resolution size. In Figure 7 (b), the estimated densities using grids with 3 to 18 points are displayed along with the result for

1000 evaluation points. As we get closer to 18 evaluation points the densities are closer to the estimated density using a size of 1000. The choice of size does not appear as impactful on the shape but it would be good to reiterate Delicado's (2007) thoughts on the matter and choose equal evaluation points for each density estimate if multiple densities are going to be compared.

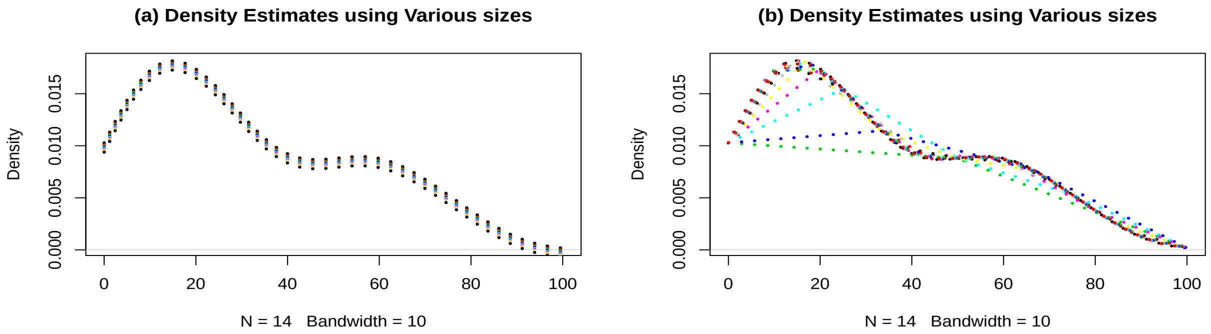

Figure 7: *The estimated densities of Blake Bortles QBR using a Gaussian kernel and bandwidth of 10 with (a) sizes of 100 to 1000 by 100 (shifted down by N/1,000,000 to be visible) and (b) sizes of 3 to 18 with one of size 1000 (right).*

3.4 Choice on whether or not to *"lift"*

In creating the estimated densities using the `density()` function in R (R Core Team, 2016) and then cropping them to be between 0 and 100, all of the estimated QBR densities lose some information. In other words, the densities created do not integrate to 1 over QBR values of 0 to 100, and more troublesome is that they do not all integrate to the same constant over that range. Take, for example, the estimated QBR densities for Blake Bortles and Peyton Manning using a Gaussian kernel, bandwidth of 10, and 100 evaluation points. Using a left Riemann sum, the area under Bortles' QBR density between 0 and 100 is 0.927 while the area under Manning's QBR density on that range is 0.879, as illustrated in Figure 8. To make the comparison more fair between all of the estimated QBR densities, they will be lifted by their area to ensure all QBR densities will integrate to 1 between QBR's of 0 and 100. This means finding the constant c for each quarterback such that

$$\frac{1}{c}\int_0^{100} \widehat{f}(x)dx = 1.$$

Figure 9 illustrates the effect this could have on the analysis where the differences are enhanced where density is largest. If only one curve is of interest then this may be unnecessary, but in comparing many densities the heights do matter.

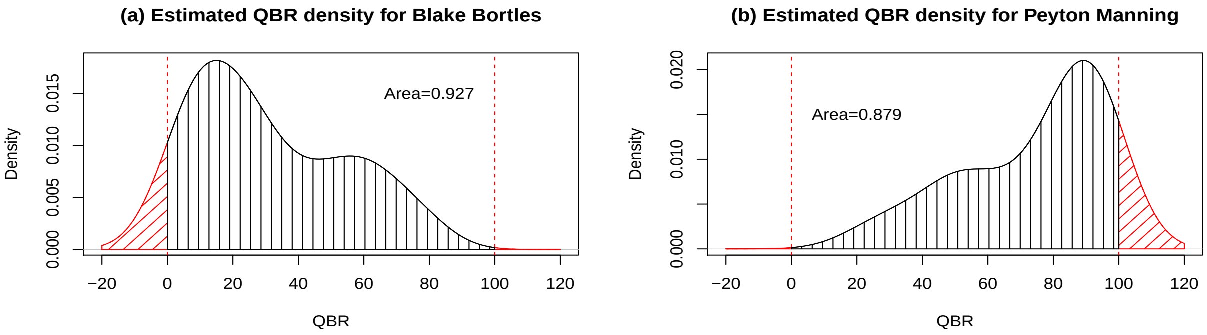

Figure 8: *(a) Blake Bortles' estimated QBR density showing that the area between 0 and 100 (shaded in black) integrates to 0.927. (a) Peyton Manning's estimated QBR density showing that the area between 0 and 100 (shaded in black) integrates to 0.879.*

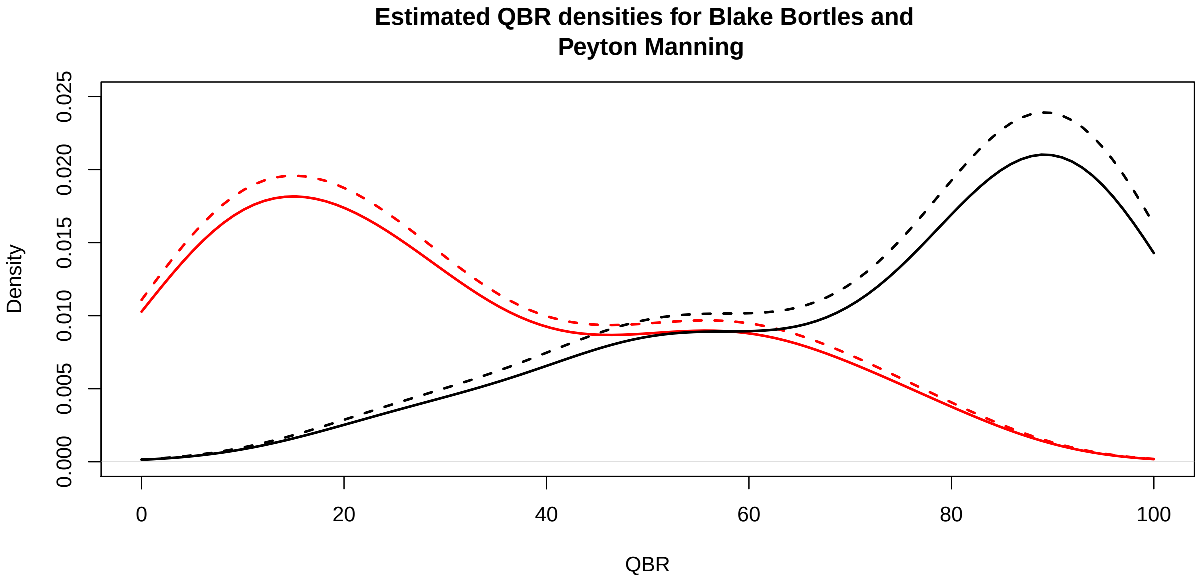

Figure 9: *The esimated QBR densities for both Blake Bortles (red lines) and Peyton Manning (black lines). The clipped QBR densities are solid lines and the lifted estimates that integrate to 1 between 0 and 100 are dashed lines.*

3.5 QBR densities

Davis and Lopez did not report how they created the density curves they used for the analysis, and as a result, they are difficult to recreate. It is also unknown what resolution they used to evaluate their density curves. For the purposes of this analysis, the density curves were evaluated at 100 evenly spaced points from 0 to 100. As for the bandwidth used by Davis and Lopez in their density estimates, it is also unknown if the bandwidths were subjectively selected or decided by the software defaults, or if the bandwidths were equal across each density curve. If they were to use the default choice in R, then the bandwidth changed across quarterbacks since the number of games played varied drastically. For the purposes of replicating the work of Davis and Lopez, a bandwidth of 15 appeared to get a close match to many, but not all, of the density curves in their report. The unknown choice of kernel used by Davis and Lopez further complicates recreation of their results. The densities they included appeared rather smooth so we selected the Gaussian kernel, which is

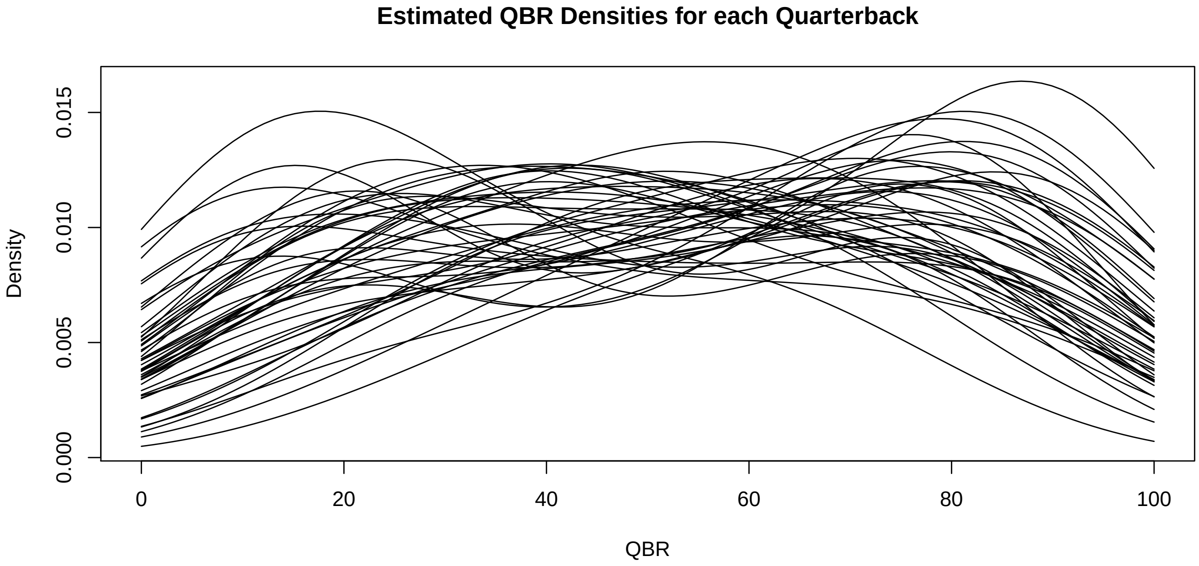

Figure 10: *The estimated QBR densities of all 46 quarterbacks using a Gaussian kernel, 100 evaluation points between 0 and 100, and a bandwidth of 15 for each. The QBR densities were not lifted.*

often the default choice for density estimates, for this analysis. The densities, which are not lifted due to our uncertainty that Davis and Lopez considered this, are shown in Figure 10. There do appear to be some groupings of curves with many having modes occurring at QBRs of either 20, 50, or 80, but it is difficult to determine exactly how many clusters may exist.

4 Clustering

As an illustration of conventional clustering methods, an example involving all of the players looking to be drafted by NFL teams from a single year will be considered. These players are all looking to play a certain position, and the characteristics required to play the position are not the same for each position; but perhaps the builds of two players are not that different despite playing different positions. For example, a wide receiver and a cornerback (defends the wide receiver) may have similar characteristics such as height, weight, and speed. Of interest is whether there are only certain types of builds of football players. A cluster analysis will be used to find groups of players that are considered similar on a multitude of physical attributes and to explore whether these characteristics relate to player positions.

4.1 Choice of Dissimilarity

The way in which the observations are clustered is determined by the dissimilarity matrix. The "dissimilarity" between two observations is a way of discerning how they differ. The smaller the dissimilarity, the more like each other the observations are. Take, for example, the data collected on players who participated in the 2015 Combine [5]. There were 174 players that had their position, height (in inches), weight, forty yard dash, vertical jump, broad jump, and reps on the bench press recorded. The one quarterback and two fullbacks in the dataset were removed to trim the sample to 171 players and the three strong safeties were re-labeled to free safeties to create a set of player position categories.

Now, if the focus is on the quantitative variables (ignoring position), then there are six dimensions of a player. It is possible, but not always useful, to plot each observation in three dimensions. A good starting point to understanding the parameter space is to create scatterplots for each pairwise comparison, for example see Figure 11 where heights and weights are displayed by position. With six dimensions there are 15 pairwise comparisons that can be visualized using a scatterplot matrix as seen in Figure 12. There are many strong correlations, either negative or positive, between variables as calculated in the lower diagonal. Many linear relationships can also be seen between the variables. Some distinct groups are present in these plots but it is hard to see them clearly.

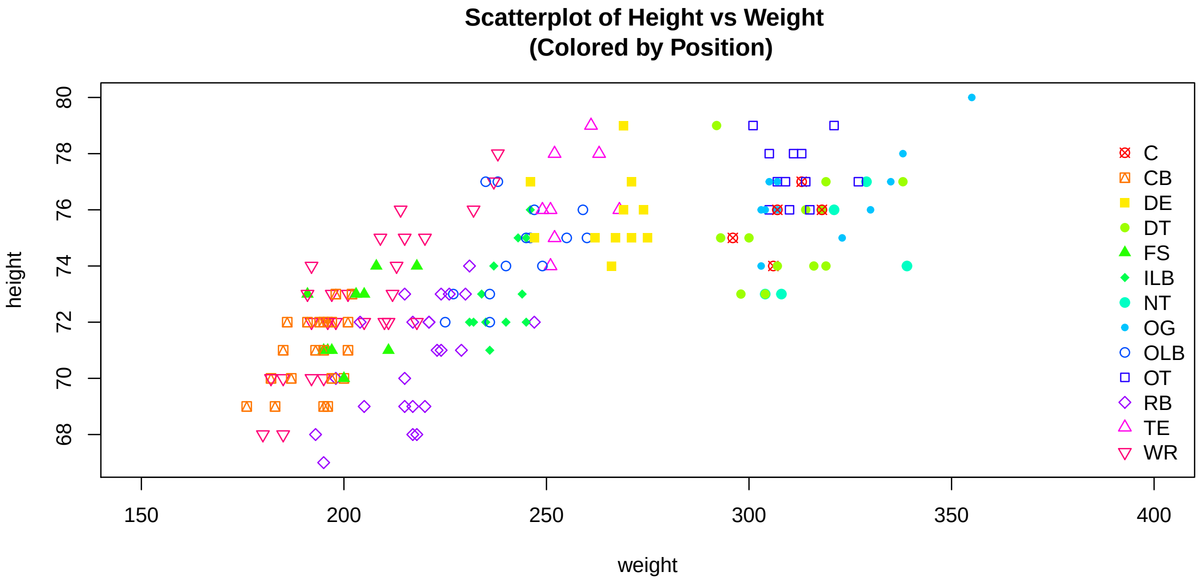

Figure 11: *The heights and weights of each player plotted against each other using his position to determine color and plotting character.*

As these data (and the QBR discussed in Section 5) are quantitative variables or being treated as quantitative, the discussion here will focus on measures of dissimilarity of quantitative variables that are metrics (or distance measures), for alternatives refer to Härdle and Simar (2012) or Hastie, Tibshirani, and Friedman (2001). Variables are often standardized (by subtracting the mean value over all the observations and dividing by the standard deviation of observed values of that variable) prior to forming the distance matrix.

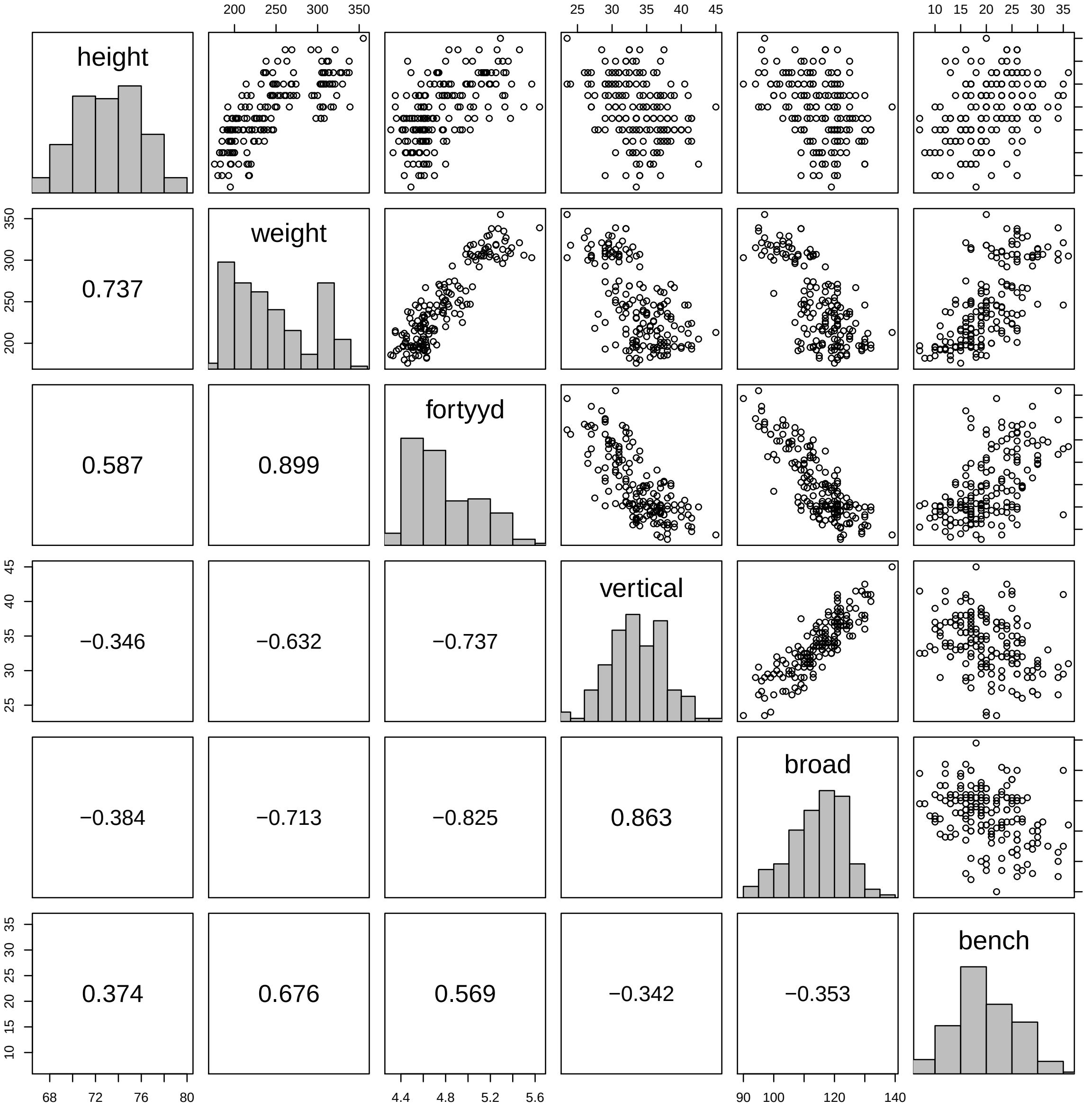

Figure 12: *Upper diagonal: Pairwise scatterplots for each possible pair of variables. Diagonal: histograms of the variables. Lower Diagonal: correlations between variables.*

In multivariate settings, this is beneficial as the ranges and different units of the variables could impact the results, putting larger emphasis on the variables with higher ranges. Take for example the weight (in pounds) of NFL players vs the hand size (in inches) of the players. The range of weight is between roughly 175 and 350 lbs while the hand size is between roughly 7 and 12 inches; a distance between two players would mostly be taking the weights of the players into account if the relative variability is not made comparable across the two variables.

The most common distance measures are L_r-norms where $r \geq 1$. In an L_r-norm the distance d_{ij} between two observations i and j can be computed across the P variables (Härdle, 2003) by

$$d_{ij} = ||x_i - x_j||_r = \left(\sum_{k=1}^{P} |x_{ik} - x_{jk}|^r \right)^{1/r}. \tag{3}$$

The benefit of L_r-norms is that being distance metrics they are non-negative ($d(x, y) \geq 0$), they obey the identity of indiscernibles ($d(x, y) = 0$ if and only if $x = y$), they obey symmetry ($d(x, y) = d(y, x)$), and they obey the triangle inequality ($d(x, z) \leq d(x, y) + d(y, z)$) (Wikipedia, *Metric*). When $r = 1$, the norm is called the 'City Block' or 'Manhattan' distance as using the absolute values is similar to travelling on a city block, i.e., determine how many blocks up and how many blocks over are being traveled, accumulating absolute differences on each variable. The Euclidean distance is calculated when $r = 2$. In two dimensions, the Euclidean distance is calculating the hypotenuse of a triangle or walking straight through the building to the chosen destination instead of walking the streets. It is common, and the default of the `dist()` function in R (R Core Team, 2016), to apply the Euclidean distance.

After a decision of which measure is to be used, a $n \times n$ dissimilarity matrix can be made, where n is the number of observations in the data set. The upper (or lower) diagonal portion of the matrix will be all that is necessary to focus on. An example of a distance matrix is given in Table 3 between the first 5 observations of the Combine data using only the six standardized quantitative dimensions and the L_2-norm. The symmetry of Euclidean distance is ovious in this result. Ajayi and Alexander were most similar and Alford and Abdullah were most different, among the five players.

	A. Abdullah	J. Ajayi	K. Alexander	M. Alford	J. Allen
A. Abdullah	0.00	3.10	3.75	4.84	4.21
J. Ajayi	3.10	0.00	1.37	3.99	1.79
K. Alexander	3.75	1.37	0.00	4.30	2.24
M. Alford	4.84	3.99	4.30	0.00	3.28
J. Allen	4.21	1.79	2.24	3.28	0.00

Table 3: *Euclidean distance matrix between five players in the Combine dataset using standardized variables.*

4.2 Choice of Algorithm

There are many types of algorithms for clustering: most popular are hierarchical clustering, k-means clustering, and model based clustering (Everitt & Hothorn, 2011).

4.2.1 Hierarchical Clustering

In hierarchical clustering, the observations are taken through several partitioning steps either by agglomerative (bottom-up) or divisive (top-down) methods (Hastie, Tibshirani, & Friedman, 2001). When the algorithm is finished, a dendrogram can be displayed to help the researcher determine an optimal cut point from which the clusters will result. A dendrogram is a tree-like structure that shows the steps between the first (or last) partition (all observations clustered together) to the final (or initial) partition (all observations belong to their own cluster). Looking from the top to the bottom of the dendrogram, it is possible to see where the splits occurred between clusters. The "height" that the split occurs on the dendrogram is proportional to the value of the inter-group dissimilarity, which is monotone increasing for merged clusters, between the clusters at the split (Hastie, Tibshirani, & Friedman, 2001).

In order for the algorithm to perform its duties, an inter-group dissimilarity needs to be defined. One approach is to apply the nearest-neighbor technique or single linkage agglomerative clustering. This method computes all pairwise dissimilarities between the observations in cluster 1 and all the observations in cluster 2 and then records the smallest of the dissimilarities. This approach can lead to extended, trailing clusters from the algorithm combining single observations one at a time (James et al., 2013). Applying this method on the Combine data results in the dendrogram shown in Figure 13. Single linkage also is susceptible to classifying differing groups into one cluster if they are not well separated (Härdle & Simar, 2003). The single linkage approach created many singleton clusters from this data set where few truly unique player types were expected so this cluster solution does not seem ideal here.

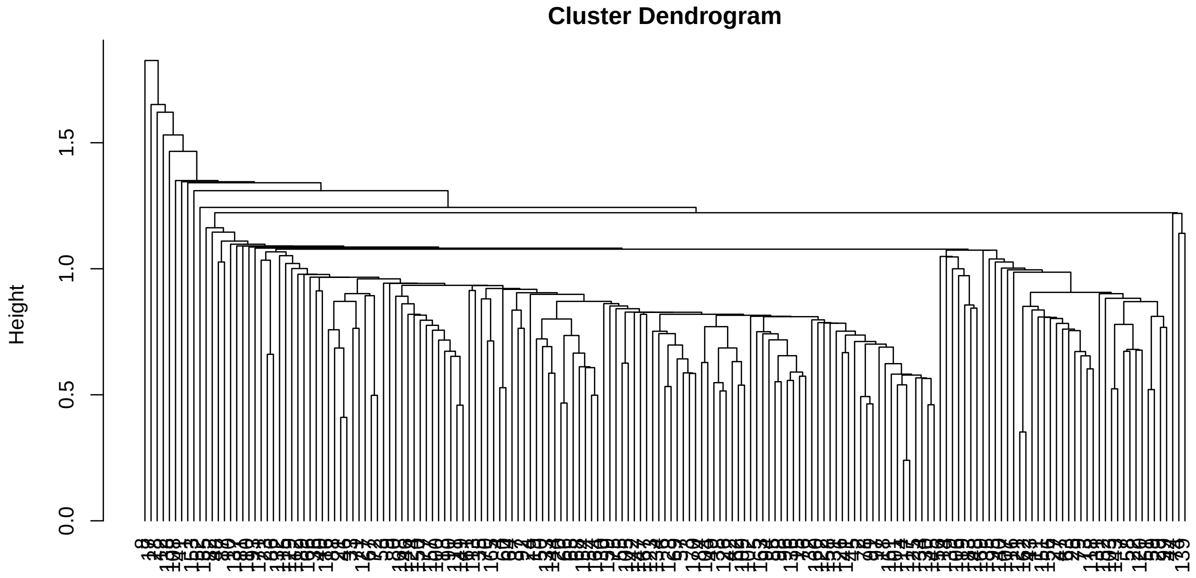

Figure 13: *A dendrogram of the Combine data using single linkage.*

Another hierarchical approach is the complete linkage agglomerative technique (or the furthest-neighbor technique (Hastie, Tibshirani, & Friedman, 2001)). This method computes all pairwise dissimilarities between the observations in cluster 1 and all the observations in cluster 2 to represent the distance between the clusters, but then records the largest of the dissimilarities (James et al., 2013). Like single linkage before, a benefit to this approach is that it is invariant under monotone transformations of the inter-individual distances (Everitt & Hothorn, 2011). It also tends to produce larger clusters than single linkage and is less prone to chaining. Applying this technique to the Combine data results in the dendrogram shown in Figure 14. If a height of 5 was chosen to cut the dendrogram, then five clusters would result, as displayed in the rectangles. From there, one could inspect the data to see what observations were clustered together.

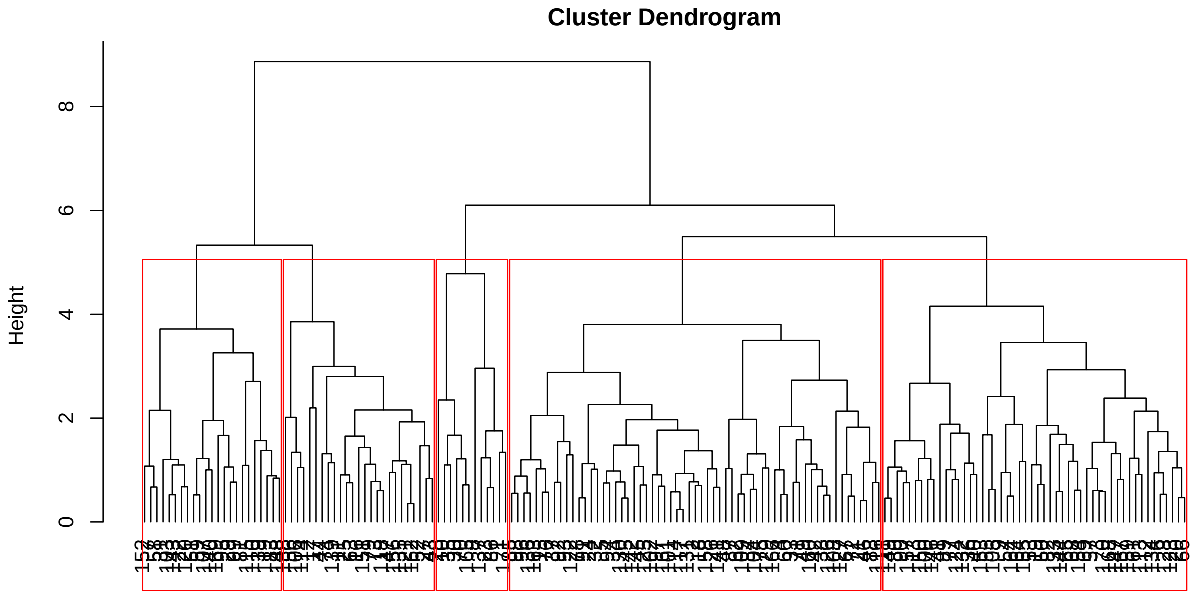

Figure 14: *A dendrogram of the Combine data using complete linkage.*

A third common linkage method is average linkage. Using the pairwise dissimilarities as calculated before for either the single- or complete-linkage algorithms, it is also possible to use the average dissimilarity between the clusters, and merge clusters that are closest on average. This approach was taken to form the dendrogram in Figure 15 for the Combine data. A cut at a height of 3 resulted in three clusters that are perhaps not optimal. Cutting any lower in the dendrogram would result in some clusters with very few players.

The last hierarchical clustering agglomeration method that will be discussed here is called Ward's method. Ward's method relies on joining groups that results in the smallest increase in heterogeneity (measured via the error sum of squares or ESS). The hope is that this increase is not too drastic. At each step of the algorithm, every possible pairing of clusters is considered and then the merger that is chosen is the one that provided the smallest increase in the ESS. Figure 16 displays the dendrogram that was the result of Ward's algorithm for the Combine data. A cut at a height of 7 resulted in six clusters as shown by the rectangles.

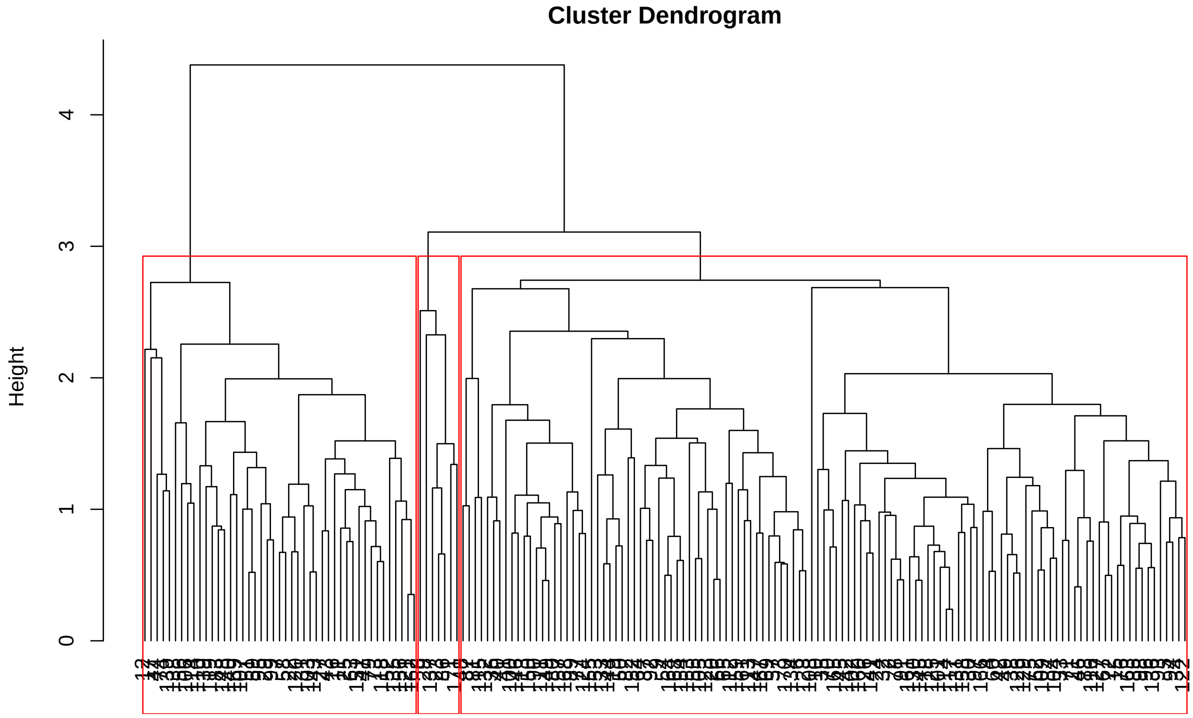

Figure 15: *A dendrogram of the Combine data using group average linkage.*

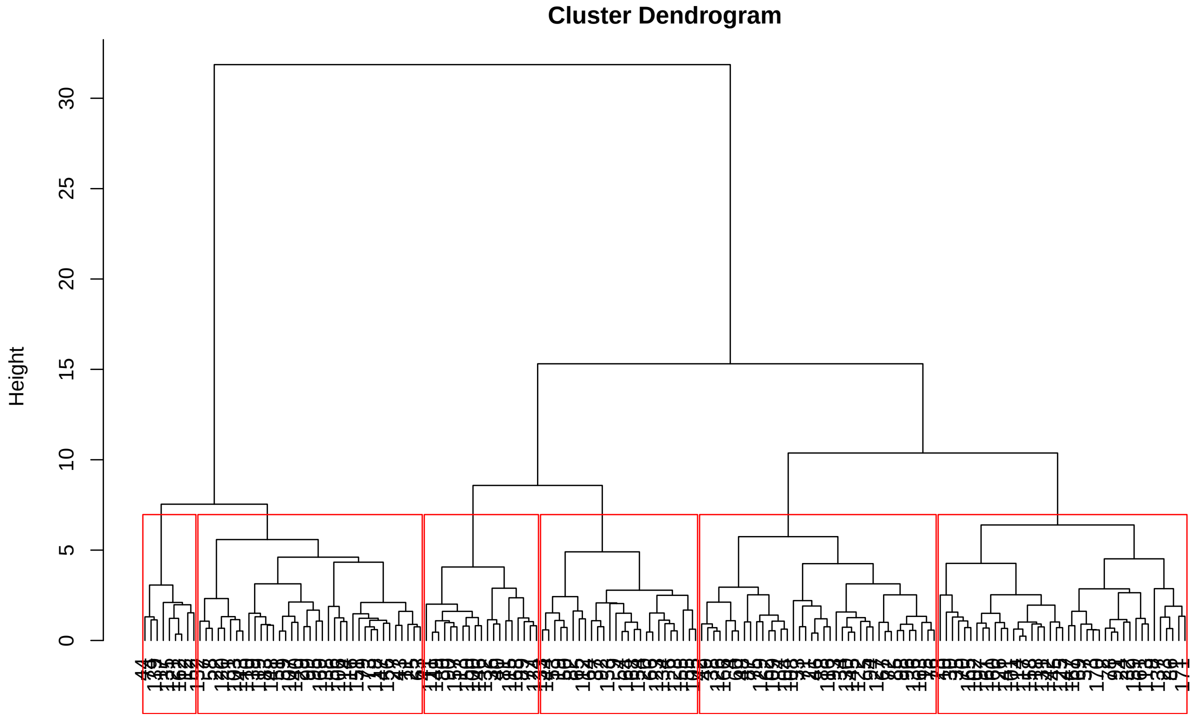

Figure 16: *A dendrogram of the Combine data using Ward's algorithm.*

The cluster solution from Ward's method vs. the number of times a player at a position that were placed in each cluster are displayed in Table 4. For example, cluster 6 is composed of 2 DE's (defensive ends), 6 ILB's (inside linebackers), 5 OLB's (outside linebackers), 2 RB's (runningbacks), 3 TE's (tight ends), and a WR (wide receiver). Cluster 3 is composed similarly to 3. Clusters 1 and 2 contain the same positions to each other, as do clusters 4 and 5, this would make it appear that there may be different types of players that play the same positions.

	1	2	3	4	5	6
C	0	0	0	4	1	0
CB	11	10	0	0	0	0
DE	0	0	9	1	0	2
DT	0	0	0	12	0	0
FS	5	5	0	0	0	0
ILB	3	1	2	0	0	6
NT	0	0	0	5	0	0
OG	0	0	0	8	3	0
OLB	4	0	5	0	0	5
OT	0	0	0	7	5	0
RB	8	13	0	0	0	2
TE	0	0	5	0	0	3
WR	10	10	5	0	0	1

Table 4: *Table showing the positions of players in each cluster using Ward's method.*

The cluster solution from average linkage had only 3 clusters and the results are shown in Table 5. The results were not very satisfactory with cluster 1 containing only 7 of the 171 players. Upon inspection, the seven players were all in the upper 25% in vertical jump and broad jump (Table 6 and Table 7). Many of the players in this cluster had a fast forty yard dash time (but none were the fastest) suggesting that they are also quick relative to the rest of the players. The cluster solution from Ward's method appears more interesting despite clusters 1 and 2, 3 and 6, and 4 and 5 sharing similar compositions of player types.

In comparing clusters 1 (Table 8) and 2 (Table 9) from the solution obtained using Ward's method, cluster 2 contains players that on average are shorter and smaller in size with slower forty yard times and shorter jumping distances on average. Likewise, cluster 6 (Table 11)

	1	2	3
C	0	0	5
CB	1	20	0
DE	0	12	0
DT	0	0	12
FS	0	10	0
ILB	0	12	0
NT	0	0	5
OG	0	0	11
OLB	1	13	0
OT	0	0	12
RB	3	20	0
TE	0	8	0
WR	2	24	0

Table 5: *Table showing the positions of players in each cluster using average linkage.*

	height	weight	fortyyd	vertical	broad	bench
1	Min. :67.00	Min. :176.0	Min. :4.310	Min. :23.5	Min. : 90.0	Min. : 7.00
2	1st Qu.:72.00	1st Qu.:203.5	1st Qu.:4.570	1st Qu.:31.0	1st Qu.:109.5	1st Qu.:16.00
3	Median :74.00	Median :236.0	Median :4.690	Median :33.5	Median :116.0	Median :20.00
4	Mean :73.65	Mean :245.6	Mean :4.786	Mean :33.8	Mean :115.0	Mean :20.33
5	3rd Qu.:76.00	3rd Qu.:294.5	3rd Qu.:5.010	3rd Qu.:37.0	3rd Qu.:121.0	3rd Qu.:25.00
6	Max. :80.00	Max. :355.0	Max. :5.640	Max. :45.0	Max. :139.0	Max. :36.00

Table 6: *Summary statistics on each dimension of all 171 players in the Combine data.*

	position	height	weight	fortyyd	vertical	broad	bench
1	RB	69.00	205	4.60	42.50	130.00	24.00
8	OLB	75.00	246	4.53	41.00	130.00	35.00
23	WR	73.00	212	4.43	41.00	131.00	23.00
28	WR	74.00	213	4.35	45.00	139.00	18.00
86	RB	73.00	224	4.50	41.50	127.00	25.00
137	CB	72.00	201	4.44	37.50	130.00	26.00
171	RB	71.00	223	4.60	41.00	121.00	25.00

Table 7: *Data on the 7 players in cluster 1 from the cluster solution using average linkage.*

has slightly smaller players in height and in size than cluster 3 (Table 10) does; cluster 3 also appears to be more athletic than cluster 6 with cluster 6 only being more skilled at bench press on average. Cluster 5 (Table 13) has taller and bigger players on average than cluster 4 (Table 12) who are less athletic, performing worse in each of the other four dimensions. Clusters 1 and 2 tend to have smaller players in weight than the other four while clusters 4 and 5 tend to have larger players in weight than the other 4. The names for the types of players could then be “small-athletic” (cluster 1), “small-moderate” (cluster 2),

	height	weight	fortyyd	vertical	broad	bench
1	Min. :69.00	Min. :183.0	Min. :4.310	Min. :35.00	Min. :117.0	Min. : 7.0
2	1st Qu.:71.00	1st Qu.:196.0	1st Qu.:4.440	1st Qu.:37.00	1st Qu.:121.0	1st Qu.:16.0
3	Median :72.00	Median :205.0	Median :4.530	Median :38.00	Median :124.0	Median :18.0
4	Mean :72.07	Mean :209.3	Mean :4.523	Mean :38.35	Mean :124.7	Mean :18.9
5	3rd Qu.:73.00	3rd Qu.:223.0	3rd Qu.:4.600	3rd Qu.:40.00	3rd Qu.:129.0	3rd Qu.:22.0
6	Max. :75.00	Max. :246.0	Max. :4.880	Max. :45.00	Max. :139.0	Max. :35.0

Table 8: *Summary measures of the 6 dimensions on the players in cluster 1 determined by Ward's method.*

	height	weight	fortyyd	vertical	broad	bench
1	Min. :67.00	Min. :176.0	Min. :4.430	Min. :29.00	Min. :108.0	Min. : 7.00
2	1st Qu.:69.00	1st Qu.:192.5	1st Qu.:4.545	1st Qu.:33.00	1st Qu.:114.0	1st Qu.:11.50
3	Median :70.00	Median :198.0	Median :4.600	Median :33.50	Median :116.0	Median :15.00
4	Mean :70.67	Mean :201.7	Mean :4.590	Mean :33.99	Mean :116.5	Mean :15.13
5	3rd Qu.:72.00	3rd Qu.:213.0	3rd Qu.:4.630	3rd Qu.:35.50	3rd Qu.:120.0	3rd Qu.:18.50
6	Max. :74.00	Max. :231.0	Max. :4.800	Max. :37.00	Max. :125.0	Max. :26.00

Table 9: *Summary measures of the 6 dimensions on the players in cluster 2 determined by Ward's method.*

	height	weight	fortyyd	vertical	broad	bench
1	Min. :74.00	Min. :214.0	Min. :4.460	Min. :32.50	Min. :109.0	Min. :12.00
2	1st Qu.:75.00	1st Qu.:238.5	1st Qu.:4.612	1st Qu.:34.50	1st Qu.:118.0	1st Qu.:16.25
3	Median :76.00	Median :248.0	Median :4.725	Median :36.25	Median :120.0	Median :20.00
4	Mean :75.88	Mean :249.0	Mean :4.699	Mean :35.98	Mean :119.5	Mean :20.00
5	3rd Qu.:77.00	3rd Qu.:261.8	3rd Qu.:4.790	3rd Qu.:37.38	3rd Qu.:121.0	3rd Qu.:24.00
6	Max. :79.00	Max. :271.0	Max. :4.930	Max. :40.50	Max. :127.0	Max. :28.00

Table 10: *Summary measures of the 6 dimensions on the players in cluster 3 determined by Ward's method.*

	height	weight	fortyyd	vertical	broad	bench
1	Min. :71.00	Min. :218.0	Min. :4.610	Min. :27.50	Min. :100.0	Min. :16.00
2	1st Qu.:72.00	1st Qu.:236.5	1st Qu.:4.780	1st Qu.:30.00	1st Qu.:110.0	1st Qu.:19.00
3	Median :74.00	Median :247.0	Median :4.880	Median :31.00	Median :111.0	Median :20.00
4	Mean :74.11	Mean :247.6	Mean :4.866	Mean :30.82	Mean :110.4	Mean :20.68
5	3rd Qu.:75.50	3rd Qu.:259.5	3rd Qu.:4.950	3rd Qu.:32.00	3rd Qu.:112.0	3rd Qu.:23.00
6	Max. :78.00	Max. :275.0	Max. :5.040	Max. :35.00	Max. :115.0	Max. :30.00

Table 11: *Summary measures of the 6 dimensions on the players in cluster 6 determined by Ward's method.*

	height	weight	fortyyd	vertical	broad	bench
1	Min. :73.00	Min. :269.0	Min. :4.860	Min. :26.50	Min. : 95.0	Min. :20.00
2	1st Qu.:74.00	1st Qu.:304.0	1st Qu.:5.060	1st Qu.:29.00	1st Qu.:103.0	1st Qu.:24.00
3	Median :76.00	Median :307.0	Median :5.150	Median :30.50	Median :106.0	Median :26.00
4	Mean :75.84	Mean :310.2	Mean :5.162	Mean :30.22	Mean :106.5	Mean :27.32
5	3rd Qu.:77.00	3rd Qu.:318.0	3rd Qu.:5.250	3rd Qu.:31.50	3rd Qu.:111.0	3rd Qu.:30.00
6	Max. :79.00	Max. :339.0	Max. :5.640	Max. :34.00	Max. :117.0	Max. :36.00

Table 12: *Summary measures of the 6 dimensions on the players in cluster 4 determined by Ward's method.*

	height	weight	fortyyd	vertical	broad	bench
1	Min. :76.00	Min. :303	Min. :5.250	Min. :23.50	Min. : 90.00	Min. :16.00
2	1st Qu.:76.00	1st Qu.:313	1st Qu.:5.310	1st Qu.:24.00	1st Qu.: 95.00	1st Qu.:17.00
3	Median :77.00	Median :318	Median :5.340	Median :26.50	Median : 97.00	Median :20.00
4	Mean :77.44	Mean :322	Mean :5.366	Mean :26.89	Mean : 96.22	Mean :20.89
5	3rd Qu.:78.00	3rd Qu.:327	3rd Qu.:5.390	3rd Qu.:29.00	3rd Qu.: 97.00	3rd Qu.:23.00
6	Max. :80.00	Max. :355	Max. :5.570	Max. :32.00	Max. :101.00	Max. :27.00

Table 13: *Summary measures of the 6 dimensions on the players in cluster 5 determined by Ward's method.*

"medium-athletic" (cluster 3), "medium-moderate" (cluster 6), "large-athletic" (cluster 4), and "large-moderate" (cluster 5).

A drawback to the hierarchical clustering methods is that once an observation has been allocated to a cluster it cannot be reallocated to another cluster. This can result in observations that are more similar to each other being placed in separate clusters and being grouped with less similar observations. Another issue is the need to determine an appropriate cut on the dendrogram which forms the clusters; this is typically done where the height changes are small for splitting clusters. Other clustering methods are available that alleviate at least the first issue.

4.2.2 k-means Clustering

The k-means algorithm is an optimization, non-hierarchical clustering algorithm. Prior to starting the algorithm, the number of clusters (k) must be defined. The algorithm then, perhaps randomly, assigns the observations to exactly one of the k clusters. It then loops through computing the centroid (located at the mean of each of the P dimensions in the cluster P_k) for each of the $l = 1, \ldots, k$ clusters and moving observations to the cluster whose centroid is nearest until the within-group sums of squares (WGSS) (Everitt & Hothorn, 2011),

$$WGSS = \sum_{j=1}^{P} \sum_{l=1}^{k} \sum_{i \in P_l} \left(x_{ij} - \overline{x}_j^{(l)} \right)^2 , \tag{4}$$

has been minimized. Equation 4 shows the formula for the WGSS where $\overline{x}_j^{(l)}$ is the mean of observations in the l^{th} partition on variable j. In theory, the algorithm would find every possible partition of the n observations into the the k clusters. However, the number of partitions becomes very large, very fast depending on k and n and computational speeds necessitate a partial search. The algorithm must randomly choose the initial partitions and then move observations between clusters if the change leads to the greatest improvement in the WGSS.

Everitt and Hothorn (2011) mention that possible drawbacks to the k-means algorithm are that scaling the variables may result in different clusters than the raw data (not unique to k-means), that the algorithm imposes a spherical structure on the clusters even if the clusters are truly of some other shape, and that repeated runs of the algorithm can produce different solutions. Hastie, Tibshirani, and Friedman (2001) also warn that the clusters formed from different k's need not be nested and that the clusters made with $k = 4$ may not be sub-clusters of the ones formed when $k = 3$. A further drawback, with a reasonable solution, is that the k-means algorithm requires a set number of clusters k to be established prior to starting. A reasonable approach is to create a plot of the change in the WGSS for each number of clusters. A choice of k would then be determined by the last “significant” change in the WGSS.

Like before, the scaled Combine data are used to demonstrate the algorithm. From Figure 17 it seems reasonable to choose four (vertical dashed line) clusters based on the relatively slow decline in total WGSS beyond four clusters. After determining the number of clusters, the percent of the variation that is explained by the clusters provides a summary of the quality of the cluster solution. In this case, the percent of variation explained by the four clusters was 68.2%. Perhaps that was not ideal and if we wanted clusters that explain closer to 80%, then at least 10 clusters would be necessary, as shown in Figure 18.

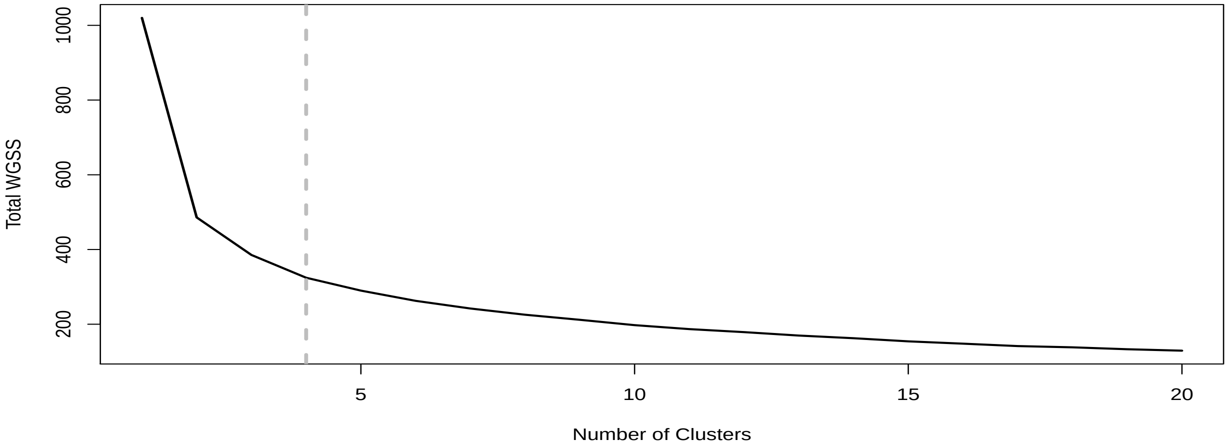

Figure 17: *The total within group sums of squares plotted against the number of clusters (1 to 20) for the scaled Combine data.*

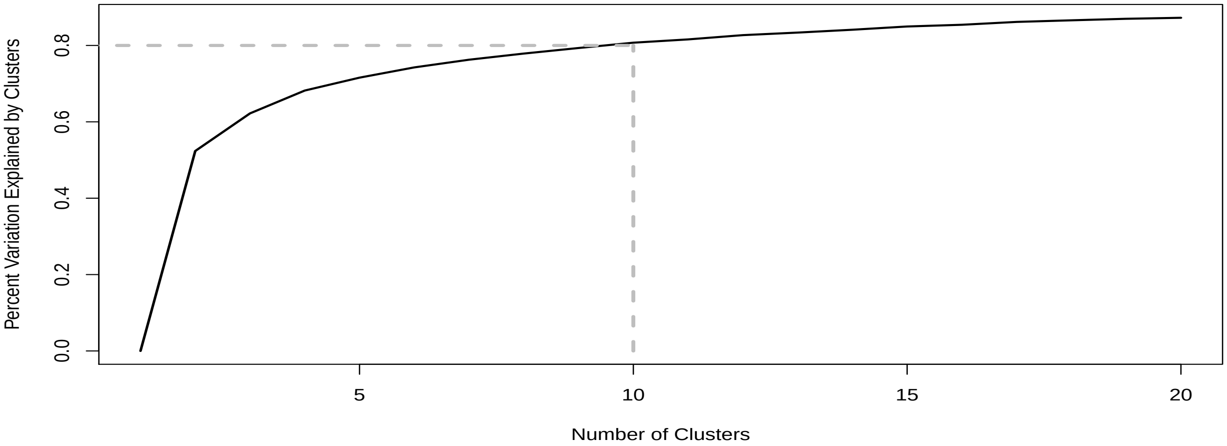

Figure 18: *The percent of variation that is explained by the model for the number of clusters (2 to 20).*

The results from one particular k-means run on the Combine data resulted in the cluster solution given in Table 14. The results are not as easily interpretable as the solution found from Ward's method. From knowledge of football, it is hard to believe a wide receiver, an offensive tackle, and a defensive tackle would belong to the same cluster as was the case in cluster 3. The cluster solution obtained using Ward's method appears to have performed better than the k-means algorithm or at least produced more easily explained results.

	1	2	3	4
C	5	0	0	0
CB	0	10	0	11
DE	0	1	11	0
DT	10	0	2	0
FS	0	6	0	4
ILB	0	4	7	1
NT	5	0	0	0
OG	11	0	0	0
OLB	0	4	9	1
OT	11	0	1	0
RB	0	9	2	12
TE	0	1	7	0
WR	0	12	3	11

Table 14: *Table showing the positions of players in each cluster using the k-means algorithm with 4 clusters.*

4.2.3 Choice of Clustering Method

The k-means algorithm can be a useful approach to clustering. However, it did not appear to perform as well as the hierarchical approach using Ward's method in this case. In other situations, the k-means algorithm could perform better. For the Combine data, it would appear that six types of players exist that are entering the NFL as determined from Ward's method.

There is a similar algorithm for more robust optimization using the medoids instead of the centroids to group the responses called k-medoids or k-medians algorithm (see Hastie, Tibshirani, and Friedman (2001), for example). Another method is model-based clustering, also known as latent class cluster analysis due to the groups being the latent or unobserved variable. For more details on model-based clustering refer to Everitt and Hothorn (2011).

5 Methods used for QBR Clustering

5.1 Replicating Davis and Lopez

For the portion of the analysis where we attempt to replicate the work of Davis and Lopez, some choices were made for clustering QBR densities of each quarterback. Evaluated estimated densities at each grid point from QBRs of 0 to 100 are treated as the "variables" and scaling these "variables" did not make sense because all responses are on the same density scale. The k-means algorithm was employed with 10 clusters to parallel the work of Davis and Lopez. The algorithm defaults to partitioning overall sums of squares using a Euclidean distance approach.

To attempt to match the results of Davis and Lopez, the QBR densities were estimated using Gaussian kernels which were not adjusted to integrate to 1 between QBRs of 0 and 100 as discussed in Section 3.4. Because no information was provided on bandwidth and the number of evaluation points on the densities, we explored all combinations of bandwidths between 1 and 20 and evaluation point grids with 5-15, 20, 30,..., 100, and 200, 300, ..., 1000. The k-means algorithm was run with ten centers for each combination of bandwidth and evaluation grid. The different solutions are compared using the adjusted Rand index (ARI) using the `adjustedRandIndex()` function from the mclust R package was used (Fraley & Raftery, 2002; Fraley et al., 2012). The adjusted Rand index provides an agreement measure that is corrected for chance that provides 0 if the agreement is no more than would be expected by chance and 1 if the agreement is perfect.

Figure 19 shows the results for the ARIs of the k-means cluster solutions with different combinations of bandwidth and evaluation grids. The largest ARI score that occurred was 0.55 for a bandwidth of 13 with 100 evaluation points. However, it is not certain that these numbers would be reproduced if the k-means algorithm were run again as k-means is influenced by the random starting points. It was apparent that the number of evaluation points from each QBR distribution had less effect on matching their cluster solution than the bandwidth choice used in the density estimator. From the plot, it appeared that a

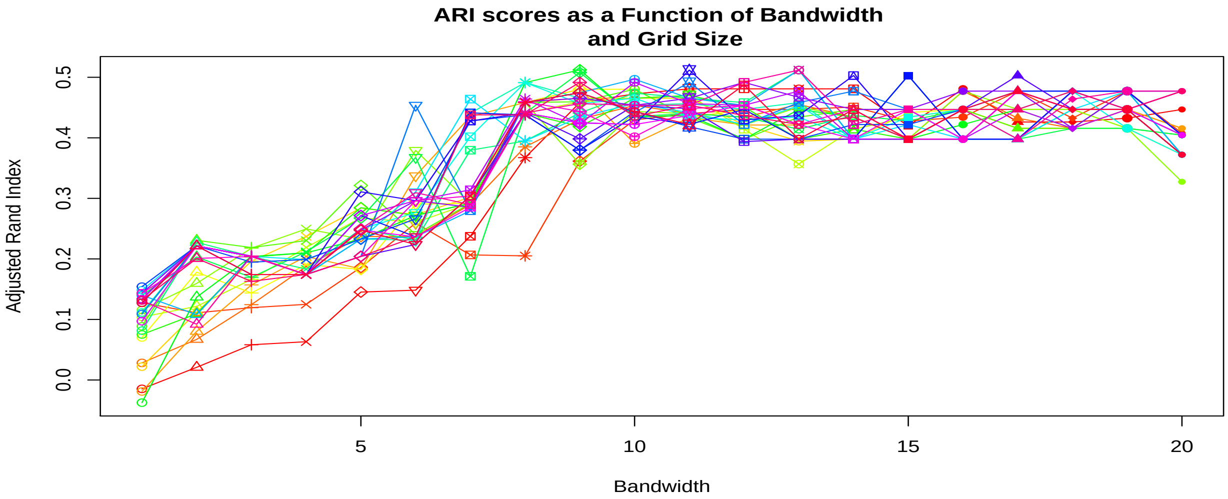

Figure 19: *The Adjusted Rand Index between cluster solutions found using the kmeans algorithm with 10 groups on QBR distributions and the cluster solution in the article plotted against the bandwidth of the QBR distributions and colored by the number of evaluation points.*

bandwidth greater than 8 for each QBR distribution found cluster solutions that were relatively similar to the article cluster solution.

Based on these results, a ten cluster k-means cluster analysis of the QBR distributions (not lifted) with bandwidths of 10 and grid sizes of 100 was performed. A plot showing all the QBR distributions (dashed lines) colored by the cluster solution is provided in Figure 20. The pointwise mean QBR distribution of each cluster is plotted with solid lines. The cluster solution was unsatisfying with three groups of size 8, two groups of size 6, one of size 4, two of 2, and two with only 1 quarterback (Peyton Manning and Blake Bortles in singleton clusters). Davis and Lopez did not provide any information on why they chose ten clusters. We could consider other numbers of clusters with k-means but chose to explore the benefits of hierarchical clustering for clustering density curves.

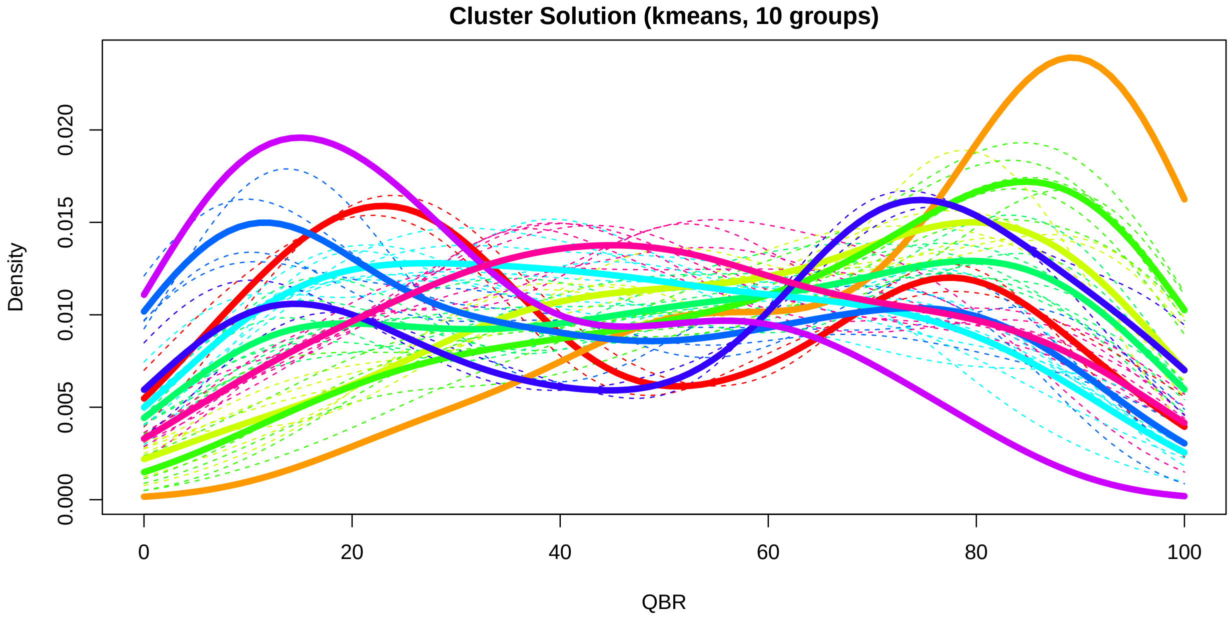

Figure 20: *All QBR distributions (dashed lines) colored by the cluster solution of the kmeans algorithm with 10 groups. The pointwise mean QBR distributions of each cluster are plotted with solid, thick lines.*

5.2 Hierarchical clustering of QBR densities

As mentioned previously (Section 4), when performing hierarchical clustering it is necessary to form a dissimilarity matrix, here it is to compare each QBR distribution. Densities are continuous functions of, here, QBR that are defined from 0 to 100 and integrate (if "lifted") to 1 over that interval. The distance metrics used should reflect the specific nature of density curves. In this case, two metrics were considered to compare distributions $f(t)$ and $g(t)$: the L2-norm

$$d(f(t), g(t)) = \sqrt{\int_0^{100} (f(t) - g(t))^2 dt}$$

and the Kullback-Leibler divergence

$$d(f(t), g(t)) = E_{f(t)}(\log(f(t)/g(t))) = \int_0^{100} \log(f(t)/g(t)) \cdot f(t) dt.$$

The Kullback-Leibler divergence is not symmetric as $d(f(t), g(t)) \neq d(g(t), f(t))$, but we can use a symmetric version (Febrero-Bande & Oviedo de la Fuente, 2012) of $0.5 \cdot [d(f(t), g(t)) + d(g(t), f(t))]$. In order to create the distance matrix to perform the hierarchical clustering the R package `fda` (Ramsay et al., 2014) was used to create a continuous representation of the densities so that the functions `metric.lp()` (for L2-norm) and `metric.kl()` (for Kullback-Leibler divergence) from the R package `fda.usc` (Febrero-Bande & Oviedo de la Fuente, 2012) could be used to form the distance matrices.

The L2-norm (Euclidean metric) was chosen for its proven validity in most clustering situations involving quantitative data and the Kullback-Leibler metric was chosen for its proven validity in comparing distributions that has ties to Akaike's Information Criterion (see Cavanaugh (1999) for details). The L2-norm was of primary interest to calculate the distance between distributions, but the Kullback-Leibler divergence could include additional information of interest. Figure 21 shows that they contain similar information but compare QBR densities in different ways since they are not following the 1-1 line.

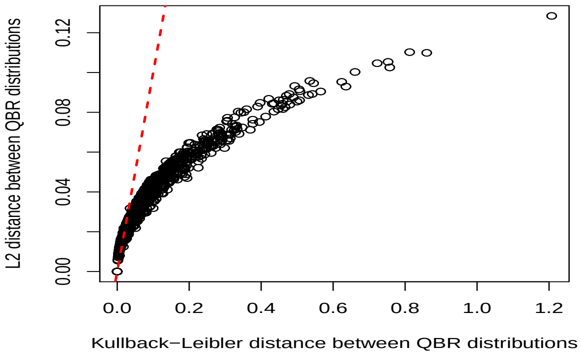

Figure 21: *A comparison of the distance metrics used to compare QBR distributions. The red (dashed) line is the 1-1 line.*

6 Results

Performing hierarchical clustering using Ward's agglomeration method on the dissimilarity matrix formed using the L2-distance between the lifted QBR distributions yields the dendrogram in Figure 22 (a). A height of 0.06 seemed to be an appropriate place to cut the dendrogram to form the clusters. Figure 22 (b) shows the QBR distributions of all quarterbacks (dashed lines) that are colored by the cluster solution formed from this cut with each clusters pointwise mean QBR distribution represented with solid lines. The cluster names we have provided do have some subjectivity to them and are inspired by Davis and Lopez as well as some knowledge of the quarterbacks.

The "Elites" group feature quarterbacks such as Peyton Manning or Tom Brady who are known to be top-of-the-charts quarterbacks. In looking at the pointwise mean QBR distribution for the "Elites", we can see that quarterbacks in this category typically perform well with the single mode being at a QBR above 80 which has a density close to 0.02.

The "Second-Tiers" group contain quarterbacks that typically perform well, like Andrew Luck or Cam Newton, but not as well as the "Elites" (at least not as of yet, for some). They

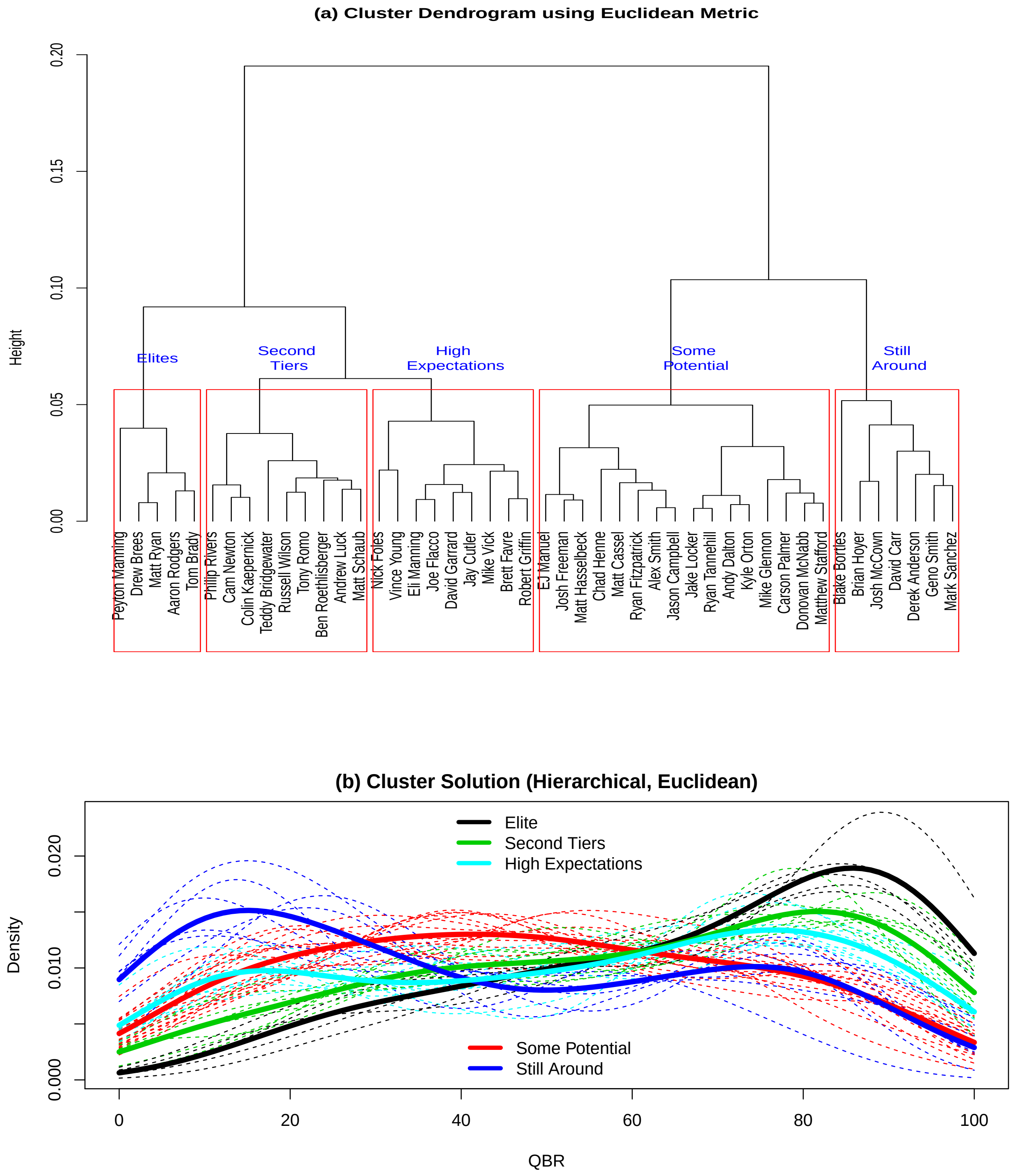

Figure 22: *(a) The dendrogram formed from Ward's agglomeration method applied to the L2-distances between the QBR distributions. A cut at a height of 0.06 results in 5 clusters. (b) The QBR distributions for all quarterbacks (dashed) colored by the cluster solution formed from the cut at height of 0.06. The solid lines are the pointwise mean QBR distributions for each cluster.*

too have a mean QBR distribution with a single mode occurring at a QBR above 80, however, the density is not as high as for the "Elites". This would suggest that they are capable of having performances that rank among the top quarterbacks in the league, but have a few more bad games than the "Elites".

The "High Expectations" group is the first group with a mean QBR distribution that is bi-modal. The naming of this group is due to it containing quarterbacks such as Jay Cutler or Joe Flacco, who, despite being paid very well, often do not perform as expected. The bi-modal shape of the mean QBR distribution helps illustrate this as one mode occurs at a QBR of around 80 and the other around 20, with the former mode having higher density. Quarterbacks in this group often have really good games with several poor performances in the mix.

The group we have labeled as "Some Potential" was named for it containing quarterbacks such as Kyle Orton or Alex Smith who have shown that they can perform well if given the right offense to match their talent. This group, on average, has a fairly flat distribution suggesting equal probability of the quarterback performing well or poorly. The mode for this group's mean QBR distribution is around 40, which is not great but also is not the worst.

The final group from this cluster solution we have named "Still Around". It was named for the fact that, on average, these quarterbacks do not play very well but do have the occasional good game that have inspired teams to keep them around, often for lack of anyone else. This includes quarterbacks such as Brian Hoyer or Mark Sanchez, who have never quite had the blame placed on them for their respective teams poor records. The mean QBR distribution for this group is another bi-modal distribution with modes occurring at about 15 and 75 on the QBR scale, with the mode at 15 having a higher density. The shape of the mean QBR distribution exposes the tendency of this group to perform poorly, but with the occasional good game.

The cut of 0.06 we believe has resulted in clusters that subjectively appear appropriate with objective evidence to support the similarities between quarterbacks in each cluster. A cut made higher would have grouped the “Second-Tiers” cluster and the “High Expectations” cluster, which we feel contain quarterbacks that do perform differently as evidenced by the mean QBR distributions for each cluster. On the other hand, a cut made lower would have resulted in splitting Blake Bortles into his own cluster away from the “Still Around” cluster, which we didn’t feel was necessary as the shape of his distribution agreed well with that of the mean QBR distribution for that cluster. The lower cut could also have split the “Some Potential” cluster in two causing quarterbacks we believe to be similar in nature (Orton and Smith, or Matt Hasselback and Matthew Stafford) to be separated, remembering that in hierarchical clustering a quarterback is not able to change clusters despite being similar to a quarterback in another cluster.

Figure 23 (a) shows the dendrogram after performing hierarchical clustering via Ward’s agglomeration method on the dissimilarity matrix formed using the symmetric Kullback-Leibler divergence between the lifted QBR distributions. A height of 0.15 was deemed appropriate to cut the dendrogram, which formed six clusters. Upon careful inspection of the labels, it was apparent that the sixth cluster formed was Bortles forming his own cluster. The remaining clusters were the same as before (with the “Still Around” group missing Bortles). This equivalence is shown in the contingency table (Table 15) between the two cluster

		K-L Clusters					
	-	1	2	3	4	5	6
L2 Clusters	1	5	0	0	0	0	0
	2	0	16	0	0	0	0
	3	0	0	9	0	0	0
	4	0	0	0	1	0	6
	5	0	0	0	0	9	0

Table 15: *A contingency table comparing the cluster solution from the hierarchical clustering using the L2-norm (rows) as a dissimilarity between QBR distributions vs. using the Kullback-Leibler divergence (columns) as a dissimilarity between QBR distributions.*

solutions. And if a six-cluster solution had been selected with L2 distances, the clusters would perfectly agree.

Figure 23 (b) shows the QBR distributions of all quarterbacks (dashed lines) that are colored by the cluster solution formed from this cut with each cluster's medoid QBR distribution represented with solid lines. The medoid is the distribution that is most similar to all other distributions in its cluster; compared to the other quarterbacks in the cluster, it has the minimum average distance to all other QBR distributions in said cluster. Figure 23 (b) includes the name of the quarterback whose QBR distribution is the medoid for that group. Each of the shapes are similar to what was shown in Figure 22 with the "Still Around" group having their lower mode flattened due to the removal of Bortles (and it being a medoid QBR distribution and not a pointwise mean QBR distribution). While, the representative distribution for the "Still Around" cluster hints at performance slightly better than previously suggested as the poor games appeared to be evened out a little (compared to the pointwise mean from before), the distribution still looks fairly similar to when Bortles belonged to the cluster.

In going back to the dendrogram (Figure 23 (a)), in order to allow for Bortles to be included in the "Still Around" cluster a cut would have to be made higher. This would have resulted in the clusters "Second-Tiers" and "High Expectations" forming one cluster. The shapes of the representative QBR distributions for each appeared different (one unimodal, the other bi-modal) and probably should not be joined. A cut lower in the dendrogram would have resulted in the splitting of the "Some Potential" group which we didn't feel was subjectively appropriate.

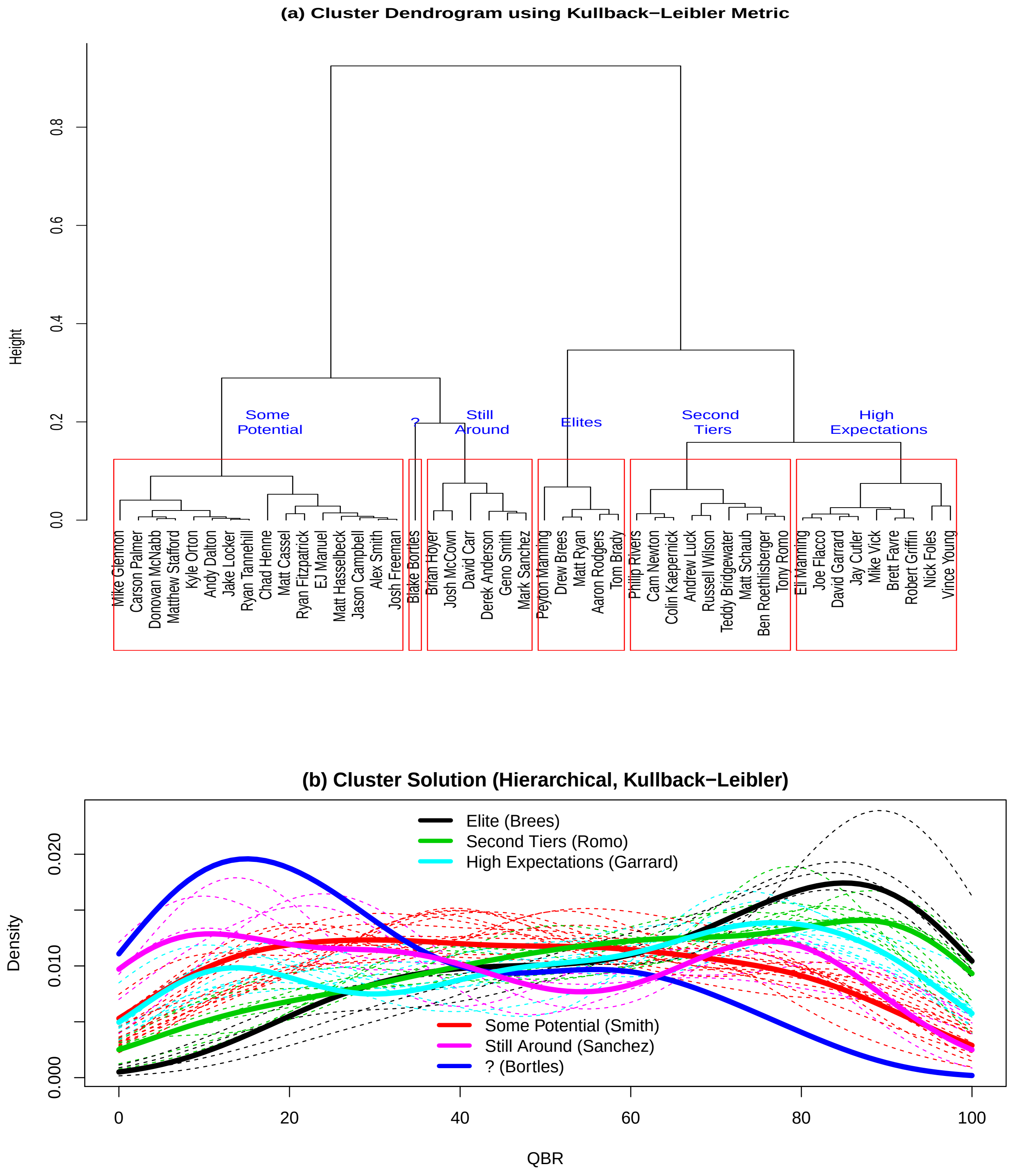

Figure 23: *(a) The dendrogram formed from Ward's agglomeration method applied to the symmetric Kullback-Leibler divergences between the QBR distributions. A cut at a height of 0.15 results in 6 clusters. (b) The QBR distributions for all quarterbacks (dashed) colored by the cluster solution formed from the cut at height of 0.15. The solid lines are the medoid QBR distributions for each cluster.*

7 Conclusion

This approach to clustering distributions proved to be a useful technique at comparing not only the means of the distributions but also the shapes of the distributions. In referencing Figure 22 (b), if we were to compare the means of the QBR distributions that ended up in the “Some Potential” and “High Expectations” clusters, then we might not have found that quarterbacks in these two groups differed. However, in taking into account the shapes of the distributions, the differences between the two groups became apparent with one being bi-modal and the other being almost uniform across the range of QBR.

As was demonstrated throughout, to apply this method requires many decisions to be made that could potentially affect the results: deciding on a density estimator, choosing a metric, choosing a clustering algorithm, and determining the number of groups. We were unable to reproduce the results established by the analysis of Davis and Lopez, which could have been the result of any of these choices or due to having different data. The clusters that Davis and Lopez determined were nice, but ultimately we made choices that we felt were more appropriate in forming our QBR distributions as well as in arriving at our cluster solution. In deciding between the two cluster solutions resulting from the hierarchical methods applied above, our preference was the five cluster solution from Ward’s method because we do not feel the representative curve changed enough in the “Still Around” cluster to warrant Bortles being in his own cluster.

8 References

[1] `http://www.nfl.com/draft/history/alltimeno1`

[2] `http://overthecap.com/position/`

[3] `http://espn.go.com/nfl/qbr/`

[4] www.nfl.com/combine/workouts

[5] http://nflsavant.com/about.php

[6] Carroll, Bob, Pete Palmer, and John Thorn. *The Hidden Game of Football.* New York, NY: Warner, 1988.

[7] Cavanaugh, Joseph E. “A Large-sample Model Selection Criterion Based on Kullback’s Symmetric Divergence.” *Statistics & Probability Letters* 42.4 (1999): 333-43. http://myweb.uiowa.edu/cavaaugh/kic.pdf.

[8] Davis, Noah, and Michael Lopez. “The 10 Types Of NFL Quarterback.” *FiveThirtyEight.* FiveThirtyEight, 16 Jan. 2015. 14 Mar. 2016. http://fivethirtyeight.com/features/the-10-types-of-nfl-quarterbacks/

[9] https://michaellopez.shinyapps.io/my_app/

[10] Delicado, Pedro. “Functional K-sample Problem When Data Are Density Functions.” *Computational Statistics* 22.3 (2007): 391-410. http://www-eio.upc.es/~delicado/my-public-files/FANOVA.density.pdf

[11] Everitt, Brian, and Torsten Hothorn. *An Introduction to Applied Multivariate Analysis with R.* New York: Springer, 2011.

[12] Febrero-Bande, Manuel and Manuel Oviedo de la Fuente (2012). *Statistical Computing in Functional Data Analysis: The R Package fda.usc.* Journal of Statistical Software, 51(4), 1-28. http://www.jstatsoft.org/v51/i04/.

[13] Fraley, Chris and Adrian E. Raftery (2002) *Model-based Clustering, Discriminant Analysis and Density Estimation* Journal of the American Statistical Association 97:611-631

[14] Fraley, Chris et al. (2012) *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

[15] Härdle, Wolfgang. *Applied Nonparametric Regression.* Cambridge: Cambridge UP, 1990.

[16] Härdle, Wolfgang, and Léopold Simar. *Applied Multivariate Statistical Analysis, third edition.* Berlin: Springer, 2012.

[17] Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-color Illustrations.* New York: Springer, 2001.

[18] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* New York: Springer, 2013.

[19] Kampstra, Peter (2008). *Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. Journal of Statistical Software, Code Snippets* 28(1). 1-9. `http://www.jstatsoft.org/v28/c01/`

[20] Oliver, Dean. 4 Aug 2011. "Guide to the Total Quarterback Rating." *ESPN.* ESPN Internet Ventures. 14 Mar. 2016. `http://espn.go.com/nfl/story/_/id/6833215/explaining-statistics-total-quarterback-rating`

[21] Ramsay, J. O., Hadley Wickham, Spencer Graves and Giles Hooker (2014). *fda: Functional Data Analysis.* R package version 2.4.4. `https://CRAN.R-project.org/package=fda`

[22] R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

[23] Rudis, Bob. 17 September 2014. "Migrating Table-oriented Web Scraping Code to Rvest W/XPath & CSS Selector Examples." *Rbloggers.* R-bloggers, 14 Mar. 2016. `http://www.r-bloggers.com/migrating-table-oriented-web-scraping-code-to-rvest-wxpath-css-selector-examples/`

[24] Smith, Michael David. 19 November 2015. “Charlie Batchs 186-yard, Two-pick Game Has ESPNs Best QBR ever.” *ProFootballTalk.* NBC Sports. 14 Mar. 2016. http://profootballtalk.nbcsports.com/2015/11/19/charlie-batchs-186-yard-two-pick-game-has-espns-best-qbr-ever/

[25] Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S. Fourth Edition.* Springer, New York. ISBN 0-387-95457-0

[26] Wickham, Hadley (2015). *rvest: Easily Harvest (Scrape) Web Pages.* R package version 0.3.1. https://CRAN.R-project.org/package=rvest

[27] Wikipedia contributors. “Metric (mathematics).” *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Metric_(mathematics)