Evolution of Popular Music

Data Analysis for Music in the Billboard Hot 100 from 1960-2010

William A. Toth

Report for the final project Data Science: COM SCI X 450.1

May 2020

Contents

1 Introduction

1.1 Project Description

The goal of this project is to investigate the evolution of music from the mid 20th to early 21st century. To do so, I will study data from 17,000+ tracks in the Billboard Hot 100 from 1960 to 2010 by examining their style and quantitative musical properties, such harmonies, chord changes, and timbres. This will allow me to determine which quantitative musical features contributed to each shift in style, musical era, and artist popularity. Then, I will train machine learning algorithms to predict both the musical style and era of a given track based upon its quantitative musical features. By doing all of this, we can visualize and understand shifts in the musical zeitgeist in the past half century.

1.2 The Data Set

The data set comes from the Figshare open access repository and can be found here. The original data set has 17,000+ rows and 293 columns, although for my purposes, I only imported and will use 41 of these columns. The data set was created for and used in The evolution of popular music: USA 1960-2010, by Matthias Mauch, Robert M. MacCallum, Mark Levy, and Armand M. Leroi, and because much of the data set is described and synthesized in the original research paper, to understand the data set, we will need to frequently refer to it. Thus, I will use the term original research article throughout my work when I need to refer their paper to explain anything. The columns which I have imported are described below.

Note: the 17,000+ tracks in the data set do not necessarily represent the music that was most released in a given time period, but rather the music that was most successful (i.e. made it to the the Billboard Hot 100)

public_id: unique ID of the recording
artist_name: name of the artist
artist_name_clean: artist name, all uppercase, no spaces, no featured secondary artists
track_name: name of the track
first_entry: date of first entry into Billboard Hot 100
quarter, firstyear, fiveyear, decade: conversions of first_entry to larger time periods
era: The era (1-4) that the track belongs to
cluster: cluster membership of the track (in other words, the style category), derived by k-means clustering in the original research article
hTopic_01, ... , hTopic_08: harmonic topic weights, described later in the paper
tTopic_01, ... , tTopic_08: timbral topic weights, described later in the paper
PC1, ... , PC14: principal components of the harmonic and timbral topics

2 Exploratory Data Analysis

2.1 Data Summary (excluding PC values, h and t topics)

Figure 1: Data Summary

```
 recording_id   artist_name        artist_name_clean   track_name
Min.   :    1   Length:17094       Length:17094       Length:17094
1st Qu.: 4274   Class :character   Class :character   Class :character
Median : 8548   Mode  :character   Mode  :character   Mode  :character
Mean   : 8548
3rd Qu.:12821
Max.   :17094
first_entry          quarter               year         fiveyear
Length:17094       Length:17094       Min.   :1960   Min.   :1960
Class :character   Class :charac      1st Qu.:1968   1st Qu.:1965
Mode  :character   Mode  :character   Median :1979   Median :1975
                                      Mean   :1981   Mean   :1979
                                      3rd Qu.:1994   3rd Qu.:1990
                                      Max.   :2009   Max.   :2005
    decade          era           cluster
Min.   :1960   Min.   :1.000   Min.   : 1.000
1st Qu.:1960   1st Qu.:2.000   1st Qu.: 4.000
Median :1970   Median :2.000   Median : 8.000
Mean   :1977   Mean   :2.633   Mean   : 7.323
3rd Qu.:1990   3rd Qu.:4.000   3rd Qu.:11.000
Max.   :2000   Max.   :4.000   Max.   :13.000
```

2.2 Understanding Harmonic and Timbral Topics, PC Data, and Clustering

As described by the original research article, the harmonic topics reflect relative levels of different chord changes, such as 'dominant seventh chords', and the timbral topics reflect different timbres, such as 'female voice, melodic, vocal' [1]. Each track can be described as a distribution over the 8 harmonic topics (H Topics) and 8 timbral topics (T Topics), whose levels were determined by text mining a 30 second snippet of the track. On the following page is a table of the various H and T topics and their meanings.

Topic	Meaning
H1	dominant 7th chords
H2	natural minor
H3	minor 7th chords
H4	standard diatonic
H5	no chords
H6	stepwise chord changes
H7	ambiguous tonality
H8	major chords, no changes
T1	drums, aggressive, percussive
T2	calm, quiet, mellow
T3	energetic, speech, bright
T4	piano, orchestra, harmonic
T5	guitar, loud, energetic
T6	/ay/, male voice, vocal
T7	/oh/, rounded mellow
T8	female voice, melodic, vocal

The PC data refers to the principal components of these timbral and harmonic topics. We are not concerned with this data column—to investigate the PC values would be beyond the scope of a class project—but rather with the cluster column, which was derived from the PC values through k-means clustering [1]. Each cluster number represents a style of music. For example, a 9 in the cluster column refers to 'classic rock/country/rock/singer-songwriter'. In this way, the cluster column uses the data from 30 H topic, T topic, and PC columns to classify each song under certain styles, and thus, when I use the word 'cluster' in this paper, I am referring to the music styles associated with it. Below is a table of the styles associated with each cluster.

Cluster Number	Styles
1	northern soul/soul/hip hop/dance
2	hip hop/rap/gangsta rap/old school
3	easy listening/country/love song/piano
4	funk/blues/jazz/soul
5	rock/classic rock/pop/new wave
6	female-vocal/pop/R'n'B/Motown
7	country/classic country/folk/rockabilly
8	dance/new wave/pop/electronic
9	classic rock/country/rock/singer-songwriter
10	love song/slow jams/soul/funk
11	funk/blues/dance/blues rock
12	soul/R'n'B/funk/disco
13	rock/hard rock/alternative/classic-rock

2.3 Data Transformation and Visualization

At this point in the data analysis process, it would make sense to deal with all missing values, but in our case, this is unnecessary. In all of the 700,000+ entries, there is not a single missing value. Luckily, the data set requires little munging in general; the column names are straight forward and there are no missing values. However, there are a few transformations that prove beneficial.

2.3.1 Clusters by Decade

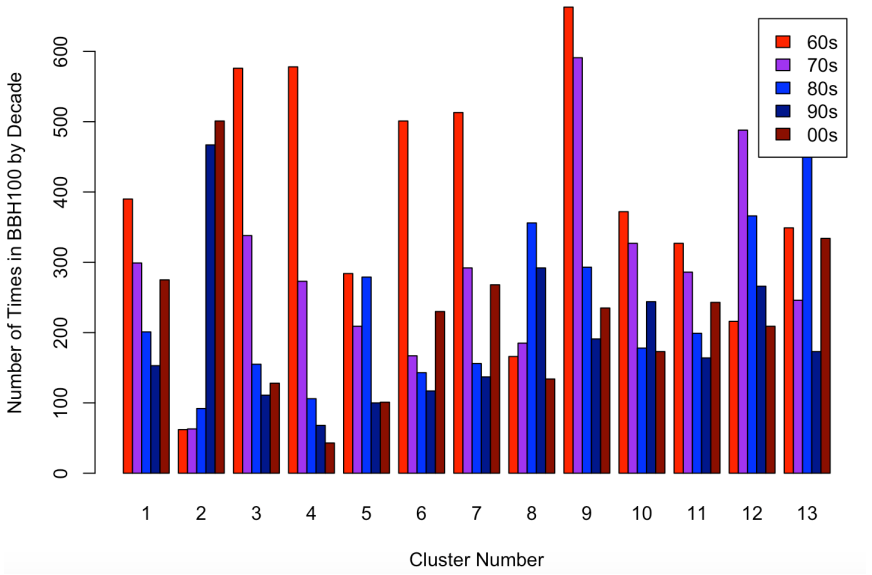
The first thing we can do to transform this data is changing the class of the cluster column. Because the numbers in this column are not quantities and only represent words/styles (i.e. 12 being one higher than 11 and two higher than 10 means nothing), we can change the class of the cluster column to factor, which is much more useful. For example, we can calculate the frequency of each cluster number (a.k.a) style in the entire data set, with 9 (classic rock/country/rock/singer-songwriter) being the most frequent (1973 tracks) and 5 (/pop/new wave/rock) being the least frequent (973 tracks).

With the cluster column as a factor variable, we can determine the frequency of cluster types by decade. There are two graphics below. The first (Figure 2) is the raw table of how many of each cluster type appeared on the Billboard Hot 100 in a given decade. The second (Figure 3) is the same results in a bar plot.

Figure 2: Cluster Frequency by Decade

	1	2	3	4	5	6	7	8	9	10	11	12	13
1960	390	62	576	578	284	501	513	166	663	372	327	216	349
1970	299	63	338	273	209	167	292	185	591	327	286	488	246
1980	201	92	155	106	279	143	156	356	293	178	199	366	452
1990	153	467	111	68	100	117	137	292	191	244	164	266	173
2000	275	501	128	43	101	230	268	134	235	173	243	209	334

Figure 3: Bar Plot for Cluster Frequency by Decade

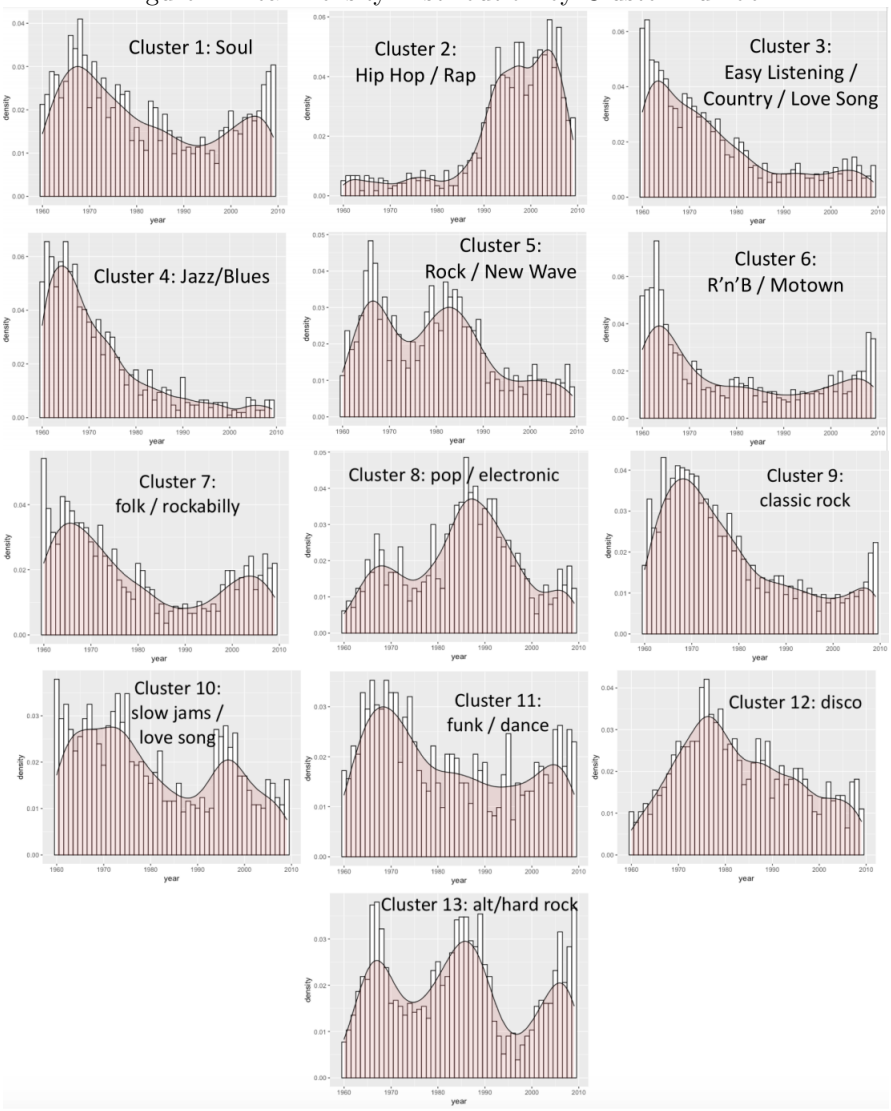


What is so interesting about visualizing the data in this way is that we can see how the trends in music styles changed over 50 years. The cluster that draws my attention most is cluster 2 (rap/hip hop). In the 60s and 70s, it is by far the least frequent cluster, and by the 2000s, it is by far the most frequent. Looking to the bar plot (Figure 3), we can visually grasp the explosion that rap music (cluster 2) experienced over time. Conversely, cluster 4 (blues/jazz/soul/funk) experienced a different fate. Being the second most frequent of the 13 style clusters in the 60s, by the 2000s, it had plummeted to the least frequent. On the bar plot, we can see this decline in popularity of jazz music, and more generally how the musical zeitgeist has shifted over the past fifty years.

2.3.2 Further Cluster Analysis

After completing the clusters by decade analysis, I realized there was much more potential for the cluster column. For each of the 13 cluster numbers, I decided to subset a new data frame with one column: year. This allows us to plot a histogram for each, showing how many times each style of music entered the BBH100 each year. Next, we can plot year density distribution for each of these, as shown in Figure 4 on the following page, which ultimately shows how the popularity of each music style varied over time with results far more in depth than those of section 2.3.1.

Figure 4: Year Density Distribution by Cluster Number



What is most interesting about the results in Figure 4 is not the major spikes for each cluster—for most of these, we knew this information prior to plotting the distributions (i.e. rap exploded over time, while rock and jazz dwindled in popularity)— but rather the smaller peaks. For example, we know soul music had it's peak circa 1970, but from the year density distribution, we see subtle reemergence into mainstream music in the 2000s. Another notable distribution is that of 13: alt/hard rock, which has three distinct peaks. Many people think of this genre as the music produced by 70s/80s bands like Aerosmith and Van

Halen, but hard rock, although more alternative this time, reemerged in the 2000s with groups like Linkin Park and Nickelback. Usually when genres of music come back, they are different. They are new. Looking at cluster 6 in Figure 4, we see a slight reemergence of R'n'B and Motown in the 2000s, but it's not the same R'n'B and Motown from the 60s. This time, it's pop artists drawing on themes from the 60s, like Rihanna and Mariah Carrey. The most prominent example of this phenomenon is the comeback of jazz, which happened after 2010 (not visible in Figure 4). Rap and pop artists, starting in around 2015, began sampling old jazz music and making their own, and if the data set included one more decade, I have no doubt we would see a reemergence of the jazz cluster, which has been dead since the 70s, not the same jazz, but 2010s-ified version of it.

2.3.3 Artist Popularity

To determine which artists appeared in the BBH100 most frequently in the 50 year range, we can convert the artist_name_clean, which is the artist name in all caps, no spaces, and not including features, to a factor variable (Note, we use this column instead of the artist_name column, as the latter includes artist features, which would create extra factor levels for each artist). This allows us to create accurate levels for a factor table (Note: a track that features an artist, but is not by the artist, will not contribute to their quantified popularity). Figure 5 below shows the 30 artists with the most BBH100 hits from 1960-2010. Each number corresponds to the name above it.

Figure 5: 30 Artists with the Most BBH100 Hits (1960-2010)

ELVISPRESLEY	JAMESBROWN	BEATLES	ROLLINGSTONES
96	74	63	57
ELTONJOHN	ARETHAFRANKLIN	NEILDIAMOND	STEVIEWONDER
55	50	49	47
TEMPTATIONS	FOURTOPS	CHICAGO	RODSTEWART
44	42	41	41
CONNIEFRANCIS	DIONNEWARWICK	IMPRESSIONS	MIRACLES
40	40	38	38
SUPREMES	RAYCHARLES	GLADYSKNIGHTPIPS	MADONNA
38	37	35	35
MARVINGAYE	BOBBYBLAND	BARBRASTREISAND	BEEGEES
34	33	32	32
BRENDALEE	CHER	BILLYJOEL	TOMMYJAMES
32	32	31	31
BEACHBOYS	BOBBYDARIN		
30	30		

If we want to see how a given artist's popularity fared over time, we can implement the approach in section 2.3.2, where we subset a new data frame with the year column, but this time, set the artist_name_clean column to a specific value/string. We can then plot a histogram with year density distribution

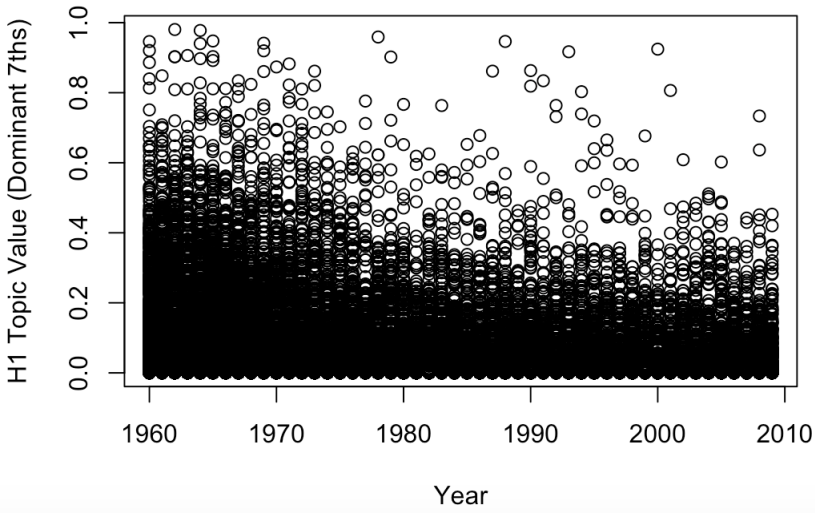
to see the artist's performance in the BBH100 over time. For our purposes, I will use the Beach Boys, but this method can be applied to any artist. In Figure 6, we can see that there are two relative maxima (1965 and 1980), revealing that despite losing popularity in the 70s, the Beach Boys had a comeback around 1980. This information that these histograms present is important because with it, we can further analyze an artists performance and try to differentiate what contributed to both peaks by comparing the harmonic and timbral topics and cluster data from the 1965 Beach Boys to the 1980 Beach Boys. We can ask questions like, "Is their reemergence into the BBH100 a result of a decrease in dominant 7th chord use (H1)?" and attempt to answer these questions. Although we will stop here these visualizations allow for much room for exploration.

Figure 6: Beach Boys BBH100 presence over time (year density distribution)

2.3.4 Harmonic and Timbral Topic Visualization

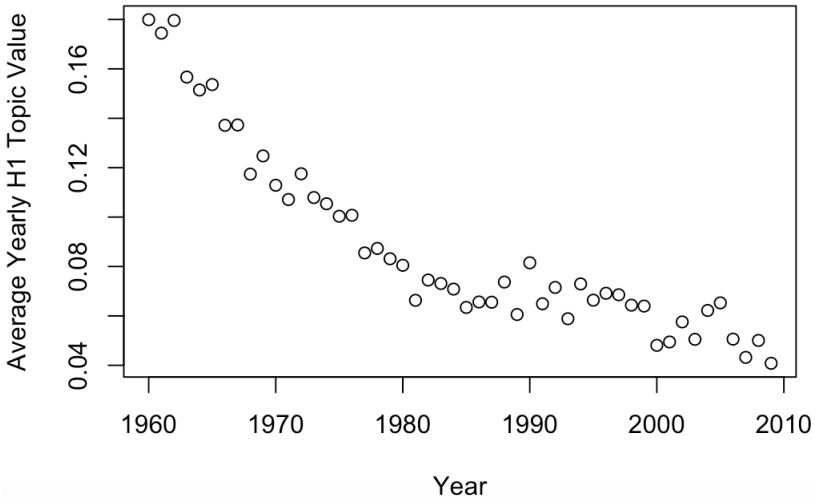
To see how quantitative musical features (H and T topics) vary over time, we can create a simple scatter plot. Figure 7 is a scatter plot of the H1 topic (Dominant 7th Chords) vs Time.

Figure 7: Dominant 7th Chord Use in BBH100 Over Time



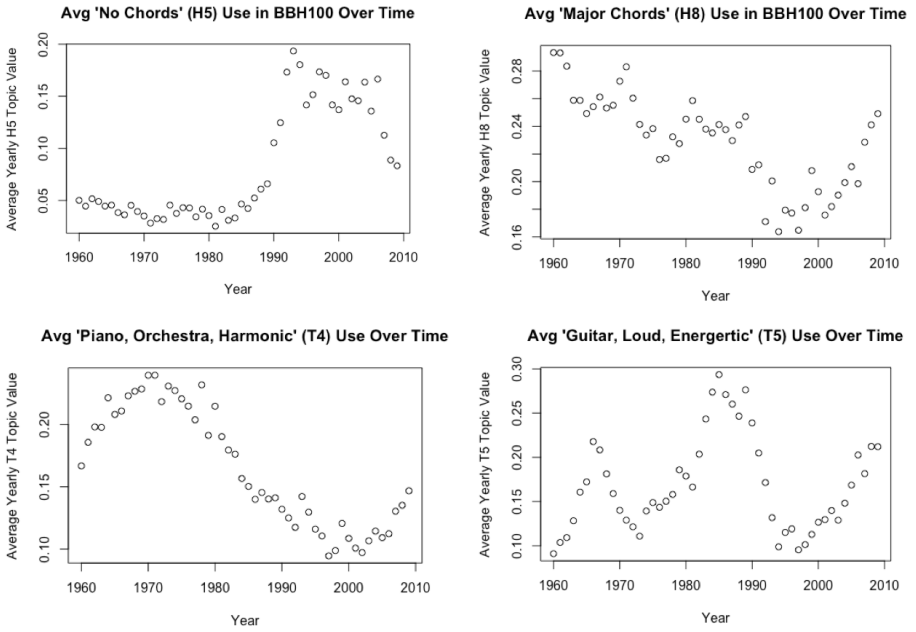
The above scatter plot, for obvious reasons, isn't very helpful to us. For this reason, we can, instead of plotting the H1 topic value for every single song, plot the average H1 value for a given year. This way, there will only be about fifty data points, each of which reflect the average (or holistic) dominant 7th chord use in popular music each year.

Figure 8: Avg Yearly Dominant 7th Chord Use in BBH100 Over Time



As we can see, there is a clear downward trend in average dominant 7th

chord use over time that looks almost like exponential decay. This shows how when the popularity of older styles, like jazz and classic rock, phased into new genres, like rap and electronic music, dominant 7th chords phased out. Not all H and T topics generated plots as interesting as this, but I have included the most interesting ones in the figure below.



Looking at the H5 'No Chords' plot, we see a major spike in the 90s and early 2000s. This is expected because the dominant music styles at this time are rap and hip hop. Earlier styles relied on chordal themes to achieve popularity, but rap, while usually involving some sort of beat, lacks the same chord changes. We can also note the decline of 'No Chords' in the mid 2000s, which is also expected as the music became less rap focused and more poppy. In addition, the T4 plot is very interesting, not because of where the highs and lows are, but because of the shape. This was the only one of the 16 topic plots that appears truly sinusoidal, and I would like to get a hold of the data from the next decade (2010s) and see if the plot continues on this trend.

Next, I would like to draw attention to the H8 and T5 plots, which in accordance with the idea of 'culture vs. counter-culture'. This is the idea that after and era of culture / positivity, there is an era of counter culture / cynicism. Looking at the H8 plot, we see a relative maximum in major chords in the 80s. This is most likely due to bands like Journey, who played very positive and uplifting music. The use of major chords rapidly declines in the 90s with the counter-culture of groups like Nirvana, Radiohead, and the Smashing Pumpkins, whose music reflected more minor and cynical themes. The major

chords increased again in 2000s, and I have no doubt, if this data set extended to 2020, that we would see another decrease (and counter culture) in the 2010s. The T5 plot reflects average 'Guitar, Loud, Energetic' timbres over time, musical features which I associate with culture eras. Looking at this T5 plot in line with the H8 plot, we can even see that the peaks and troughs occur at relatively the same places, with relative maximas in the 60s, 80s, and 2000s, and relative minimas in the 70s and 90s. Also note, my ideas about culture and counter culture are merely conjectures and are not in any way the absolute truth. I am rather trying to give meaning to the interesting data trends in the above plots.

3 Machine Learning Algorithms

3.1 Predicting Cluster with Random Decision Forests

In this section, we will predict the cluster number for a given track using 16 h and t topics as predictor variables. For classificatory modeling, we can use logistic regression or random forests. In this case, we will use random forests,an algorithm, which uses feature randomness to create an uncorrelated collection of decision trees, because the response variable has more than two levels. For this case, we will use a 60:40 train:test ratio and the algorithm default of 500 trees. The results are below.

Figure 9: Random Forest Specs and Confusion Matrix

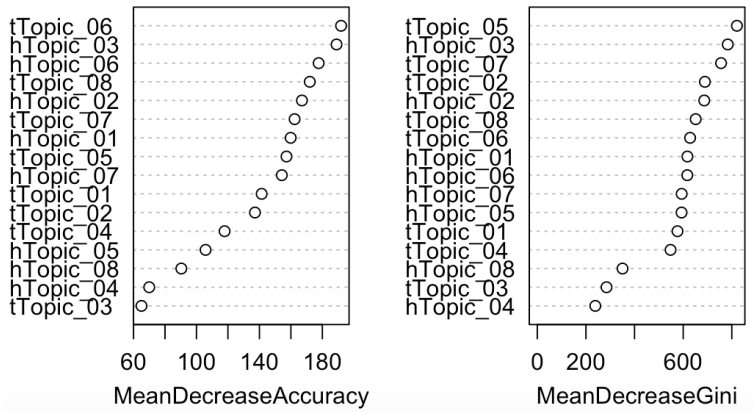
```
Call:
 randomForest(formula = cluster ~ ., data = train_rf_df1, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 9.16%
Confusion matrix:
     1   2   3   4   5   6   7   8    9  10  11  12  13 class.error
1  723   6   9   0   1   6   7   7   14  12   3   7  10  0.10186335
2    7 658   0   5   2   5   1   2    0   1   8   9   2  0.06000000
3    3   0 760   5   4   2   3   1   14  10   1   5   0  0.05940594
4    3   3   7 567   2   9  11   2   13   5   7   5   7  0.11544462
5    2   0   4   1 491   8  12   5   23   4   7   4  13  0.14459930
6    2   6   1   4   4 640   3   0   14   1   1   8   2  0.06705539
7    7   4  10   9   6   5 718   6   20   2   7   7   5  0.10918114
8    5  10   1   7   6   2   2 585   21   1   8  13   7  0.12425150
9   13   1   9   7  13  10  13  12 1072   7  12  20   6  0.10292887
10   9   2  18   2   1   2   0   0    6 702   3  13   0  0.07387863
11  13   8   5   6   1   8   9   6   11   4 630   1  16  0.12256267
12   5   8   7   2   1   5  10   9   19  10   1 861   6  0.08792373
13   2   1   0   3   5   2   2   4   16   0   5   3 910  0.04512067
```

As we can see, the algorithm was relatively good at predicting the cluster for a given track, with a misclassification error rate of 8.9 percent. This is largely due to the fact that the cluster column was indirectly derived from h and t

topics (the data set creators uses k-means clustering on PC data, and the PC data was derived from h and t topics). Below are variable importance plots for the 16 topics.

Figure 10: Variable Importance Plot



3.2 Predicting Era with Random Decision Forests

In this section, we will be predicting the musical era of a given track based on its relative levels of harmonic and timbral topics. The creators of the data set define four eras (1–4 in the data frame). These four eras are separated by three musical revolutions where there were drastic changes in the quantitative music data in 1964, 1983, and 1991. Thus, era 1 is 1960–1964, era 2 is 1964–1983, era 3 is 1983–1991, and era 4 is 1991–2010. Note, era one may possibly extend to years before 1960 and era 4 may extend to years after 2010, but this is where our data starts and ends. To predict classification, we will again implement the random forest algorithm with a 60:40 train:test ratio. The results are below.

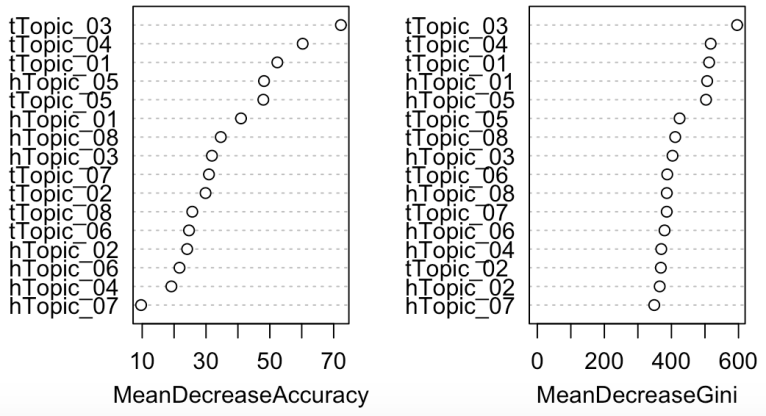
Figure 11: Random Forest Specs and Confusion Matrix

```
Call:
 randomForest(formula = era ~ ., data = train_rf_df, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 42.01%
Confusion matrix:
   1    2   3    4 class.error
1 79  915  11   56   0.9255419
2 67 4004 107  544   0.1520542
3  2  848 177  391   0.8751763
4  5 1255 108 1687   0.4477905
```

The random forest algorithm functioned unexpectedly worse than I had anticipated, with a misclassification error rate of 42.6 percent. While this algorithm succeeded in classifying the era more often than not, 42.6 percent error is certainly not good. Looking at the confusion matrix, we see that the algorithm was relatively good at classifying tracks of era 2 (1964-1983), mediocre at classifying the tracks of era 4 (1991-2010), and terrible at classifying the tracks of eras 1 (1960-1964) and 3 (1983-1991). But to say the algorithm was good at classifying tracks of era 2 would be wrong. If we look at where the mistaken tracks were placed for the other three eras, they were, for the most part, wrongly placed into era 2. This shows that the algorithm sorted a disproportionate amount of songs into era 2—it was the most common misclassification location when the actual era was 1, 2, and 3—so when it was sorting era 2 tracks, it would of course be more likely to sort them correctly. I believe the large error for era 1 comes largely from the fact that era 1 is such a small duration (only 4 years). Due to this, not only are there less era 1 tracks in the data set, but the era one topic values will be very close to those of era 2. Thus, when I removed era 1 from the available levels (i.e. subsetted the data to be from 1964 and on), the error dropped to around 36 percent. Ultimately, these results indicate that h and t topics work as predictors for musical era, but not very well. Below are variable importance plots, which tell us which h and t topics were strong and weak predictors.

Figure 12: Variable Importance Plot



References

[1] Matthias Mauch, Robert M MacCallum, Mark Levy, and Armand M Leroi. The evolution of popular music: Usa 1960–2010. *Royal Society open science*, 2(5):150081, 2015.