Music Gives Wings to the Mind[*]
An Examination of Popular Music and the Nature of Billboard Hot 100 Charts

Timothy Crawford, Katherine McLaughlin, Lung Fai Ng

Statistics 202A, Fall 2011

1 Introduction

Music has long been a constant in human existence, present in different forms for different cultures, but nonetheless an integral part of civilization. In modern times, technological advances have made the recording and transmission of music easy, as well as providing a plethora of data. Popular music is a constantly adapting field, subject to the ebbs and flows of current taste and the prominent figures that in large part create what we can colloquially refer to as "pop culture." The *Billboard Hot 100* is one metric for determining popular songs. Each week, Billboard tabulates radio audience impressions measured by Nielsen Broadcast Data Systems, sales data compiled by Nielsen SoundScan, and playlists from select non–monitored radio stations to determine the 100 most popular singles and tracks. Songs that *chart* for a given week obtain a ranking between 1 and 100, with 1 being the most popular song for that week. Herein, we present informal visual analyses and more rigorous kernel density and regression estimates for several song popularity measures derived from the Billboard Hot 100 data and more general information about the song.

2 About the Data

Our data were compiled as part of the Whitburn Project [1], an effort to preserve and archive popular songs since the inception of the Billboard Hot 100. The dataset consists of 36,928 unique songs, by 8,298 unique artists, from 1890 through April 26, 2008. The very first song that charted was "The Thunderer" by The U.S. Marine Band in 1890, and the most recent song was "He Said, She Said" by Ashley Tisdale in 2008. The artist with the most songs to chart was Bing Crosby with 834. Bing Crosby also had the most songs to reach number one with 33. The song that charted for the most weeks was "How Do I Live" by LeAnn Rimes, which charted for 69 weeks.

For each song to chart, our dataset had measurements for the following parameters: the title of the song; the artist that recorded the song; the genre of the song; the year the song charted; the number of weeks the song charted for; the peak position the song obtained on the charts[2]; the length of the song; the beats per minute of the song; weekly observations on the song's chart position for the first ten weeks; and a mysterious value called Temp.1 that the creators of our dataset called "Lancefer's Scoring System" and used as a measure of the song's popularity. (In Section 4, we detail an attempt to predict Temp.1 using a general additive model to determine how the creators of the dataset constructed this variable.)

Before we began our analysis, we restricted our dataset by removing missing values. First, we removed any rows that had at least one missing value for year, artist, track, time, beats per minute, weeks charted, chart high, Temp.1, and genre. Next we removed rows that had nonsensical values of these parameters, which were determined to be "0" for weeks charted; "0" or "—" for chart high; and "#REF!" for Temp.1. Then time, originally given in the form "minutes:seconds," was converted into seconds. We kept rows that had missing values in the ten vari-

[*] Plato

[1] Baio, Andy. The Whitburn Project: 120 Years of Music Chart History. 15 May 2008. Updated 17 May 2008. `http://waxy.org/2008/05/the_whitburn_project`

[2] It should be noted that when we refer to a "high" peak position herein, we are referring to a more popular song, i.e., a song that reached the top 10 on the Billboard Hot 100 would have a "high" peak position, even though the actual value of the position is low, i.e., 1–10.

ables measuring the song's chart position in each of the first ten weeks because this was expected – many songs charted for only a few weeks and only more popular songs remained on the Billboard Hot 100 for extended periods of time. These NA's are themselves indicative of the popularity of the song so we retained them in the dataset. However, songs that had a missing value for week 1 chart position were also removed because this represents an impossibility. After these restrictions, our final dataset consisted of 18,869 unique songs by 4,670 unique artists.

We first performed some informal analyses to examine our data and pick out relationships between variables that were particularly interesting for more formal analysis. This exploratory data analysis is given below in Section 3, and the formal analysis follows in Section 4.

3 Exploratory Data Analysis

Figure 1 shows four plots that represent our first very rough attempt at summarizing the data. Figure 1A is a histogram of year. In the restricted dataset with 18,869 observations, only songs since 1957 remain – those recorded before this time have missing values. Each bin contains the same number of years, and we can see that older years tended to have more unique songs. However, since each year has the same number of weeks, we can infer that more recent songs seem to stay on the charts for longer, whereas there was more turnover of songs on the Billboard Hot 100 further in the past. Figure 1B is a barplot of genre. The six most common genres are given (rock, country, vocal, rap, r&b, and jazz), and the rest of the genres are grouped into the "other" category. Rock was by far the most common genre of song. Figure 1C is a plot of time of song versus year. There is a slight positive association, so it seems that songs have been getting longer over time. Figure 1D is a plot of chart high versus number of weeks charted for each song. This plot displays a logical trend: songs that charted for more weeks also attained a better peak position, both of which indicate their popularity. This relationship will be examined in more detail in Section 4.2.

Next we consider boxplots of four music parameters by genre in Figure 2. Figure 2A looks at beats per minute by genre. There does not appear to be too much difference among the genres, with the exception of rap, which seems to have a lower number of beats per minute. Rock also has a larger variance than the other genres, but this is likely due to the larger number of songs that fall into this category. Figure 2B shows year by genre. This plot shows clear trends in the popularity of various genres of music over time. Vocal and jazz songs were especially popular in the 1960s and 1970s, while rap is a more recent phenomenon. Rock, country, and R&B maintained their popularity over time. Figure 2C shows chart high by genre. All genres span the full 100 possible chart high positions, with rap and rock in general having reaching higher chart values, while jazz seems to have lower chart values overall. Figure 2D shows number of weeks charted by genre. There are many outliers on the high end of weeks charted in this plot, representing songs that were particularly popular. Rap songs appear to chart longer than songs of other genres, but as we noted in other informal analyses, rap songs are also a more recent development and more recent songs tend to chart for longer than older songs, so there may be some confounding at work here.

3.1 Artist Spotlight: LeAnn Rimes

To conclude our exploratory data analysis, we picked several noteworthy artists to specifically discuss there songs.

As mentioned above, LeAnn Rimes' song "How Do I Live" was the longest–charting song at 69 weeks, so we chose to explore her music in more detail. All her songs that reached the Billboard Top 100 can be seen in Figure 3, where we trace their chart positions at each of the first ten weeks that they charted. It should be noted that there are no missing values in this data, i.e., all of LeAnn Rimes' songs that reached the Billboard Top 100 stayed there for at least 10 weeks. This is a feat that not many artists can claim.

3.2 Artist Battles

We decided to end our informal analysis of the data with two "artist battles" that pitted musicians from our childhoods against each other: Christina Aguilera versus Britney Spears and The Backstreet Boys versus N'Sync to attempt to determine who truly was more popular.

To do this, we created a popularity measure for each year. For year i, let the artist have had J songs chart on the Billboard Hot 100. Then the popularity measure is defined as

$$\text{pop}_i = \sum_{j=1}^{J} \sum_{k=1}^{10} 101 - \text{chart pos of song } j \text{ for week } k$$

We subtract each chart position from 101 to counteract the fact that a low–valued chart position (i.e., 1–10) represents a more popular song, and gives the songs that did not chart during a particular week a value of 0 so they do not add to the sum.

For example, if an artist had two songs chart for year i, each of which had weekly rankings for the first ten weeks it charted of $10, 9 \ldots, 2, 1$, then its popularity measure is $\text{pop}_i = 2 \cdot [(101 - 10) + (101 - 9) + \cdots + (101 - 1) = 1910]$.

This is just one possible way to consider popularity, and a further analysis could consider other metrics or other factors in the analysis.

Figure 4 shows plots of year versus this popularity measure for the two artist battles we considered. The plot on the left depicts Christina Aguilera versus Britney Spears. It appears that when the two artists first appeared on the scene, Britney Spears gained more immediate popularity. However, both artists' popularity declined around the year 2002, with Spears' decline more drastic. Christina Aguilera had another peak in popularity around 2003, while Spears' second peak was around 2004.

The plot on the right shows the artist battle for The Backstreet Boys versus N'Sync. This distribution is not bi–modal, but instead each group maintained their popularity over a number of years. It seems that The Backstreet Boys were more popular earlier, but that N'Sync was able to maintain their popularity late.

4 Analyzing the Data using `R` and `C`

We now conduct a more formal analysis of some of the relationships we observed in Section 2 using kernel density and kernel regression estimates in `R` and `C`.

4.1 Kernel Density Estimates

One general observation is that songs that attain a higher position on the Billboard Hot 100 charts also tend to last for longer than songs that did not reach as high of a peak position. To solidify this idea, we split our data set into two parts: those songs that charted for their first five weeks, and those that did not. For each of these two subsets, we performed kernel density estimation for the week 1 chart position of the songs. This is a good indicator of the initial popularity of the song, and we were curious if a song's initial popularity was related to its longevity.

Figure 5 shows these kernel density estimates. Figure 5A directly compares these two categories of songs. We can see that songs that did last for their first five weeks on the charts also tended to begin their first week at a higher chart position (lower ranking value) than those that did not last their first five weeks. Figures 5B and 5C give the 95% confidence bounds for these two kernel density estimates, for songs that did last their first five weeks (B) and for songs that did not last their first five weeks (C). Most songs seemed to debut within the 60–100 range among those that lasted their first five weeks, while those that did not seemed to premiere more in the 80–100 range.

To determine if songs that charted for their first five weeks are really different from those that did not, we can compute a t–test on our kernel density estimates of week 1 chart position between the two subsets. This gives a t–value of -58.8974, and with 8928 degrees of freedom, the P–value is $< 2.2\text{e}{-16}$. However, t–tests rely on normality assumptions which may not be valid in this case.

Therefore, we used a nonparametric test, the Wilcoxon Rank-Sum test, to compare the week 1 chart positions among the two subsets. This is an appropriate test because the week 1 chart positions are themselves ranks, so we are not losing any information about the distance between our observations by ranking the values, as dictated by the Wilcoxon Rank-Sum test. This test gives the value of W as 12276987, which results in a P–value $< 2.2\text{e}{-16}$. This agrees with the t–test we conducted and our intuition that songs that lasted five weeks on the Billboard Hot 100 chart should have different debut positions than those that did not.

4.2 Kernel Regression Estimates

Another way to examine whether or not songs that lasted longer on the Billboard Hot 100 charts also attained a higher position is to directly compare the number of weeks a song charted and its chart high position using kernel regression estimates.

This kernel regression is shown in Figure 6, along with 95% bootstrap confidence bounds for the estimate. This plot demonstrates that songs that lasted longer on the charts tended to reach higher on the charts (i.e., they had a smaller value of peak chart position).

It is interesting to note that the kernel regression estimates shown in Figure 6 look to roughly follow the line $y = 1/x$. We therefore transformed our data and used a linear model to look at this relationship. The regression summary for this method is

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.0992     0.2222   113.0   <2e-16 ***
I(1/CH)     110.2997     0.9186   120.1   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 23 on 18867 degrees of freedom
Multiple R-squared: 0.4332,     Adjusted R-squared: 0.4331
F-statistic: 1.442e+04 on 1 and 18867 DF,  p-value: < 2.2e-16
```

The inverse of chart position is highly significant, and the adjusted R–squared value for the model is 0.4331, indicating a moderately good fit.

4.3 General Additive Model

Finally, we decided to fit a general additive model to attempt to predict Temp.1, the mysterious measure used by the creators of the dataset to measure popularity. In our attempt to reverse engineer this value, we considered four parameters: the number of weeks a song charted for, the chart high that the song obtained, the time of the song in seconds, and the beats per minute of the song. These initial parameters were selected based on our informal analysis, which led us to believe that they were good predictors of a song's popularity.

The process for this general additive model is shown in Figure 7, which gives multiple kernel regression estimates to predict Temp.1 along with 95% bootstrap confidence intervals. The predictors are (in order): number of weeks charted, chart high, time of song (seconds), and beats per minute. The residuals from the previous estimate are used as the new y values sequentially. The plots should be read down and across: the top left plot shows the kernel regression estimate of Temp.1 using weeks charted as a predictor. The kernel regression estimate is the solid red line, and the 95% bootstrap confidence intervals are shown as dashed lines. The residuals from this plot are shown in the histogram below it (bottom-left). They do not look approximately normal yet, so we continue adding parameters to the model. These residuals are used as the y value in the second prediction, where we use the chart high as a predictor. This process continues, using the time of the song in seconds as the third predictor, until we include beats per minute in the model as the fourth and final predictor. Adding this term does not seem to make the residuals appear any more normal (the two bottom right plots look the same), so we conclude that beats per minute does not help with the prediction of Temp.1 and drop it from the model. The residuals after using the first three predictors do appear roughly normal, so it seems that our model is valid.

Therefore the final general additive model to predict Temp.1 should include the number of weeks a song charted for, its high position on the chart, and the length of the song in seconds. We can check the result of this model selection using the `gam` function from the `mgcv` package in `R`. Both combinations of predictors (the three final predictors with and without beats per minute) both produce models that have an adjusted R–squared value of 0.854, indicating that they explain 84.5% of the observed deviance. This indicates a good fit, and adding beats per minute to the model contributes nothing. For this final model, we have the following regression summary for the parametric coefficients:

```
Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 271.91371    9.42191   28.86   <2e-16
CH           65.22941    0.41637  156.66   <2e-16
High         -6.04475    0.09168  -65.93   <2e-16
sec          -0.56007    0.03974  -14.09   <2e-16
```

Note that all three of the predictors are significant at the 1% level, with very small P–values. Thus we can do a good job of reverse engineering the mysterious Temp.1 measure of popularity using the parameters we were given in the dataset. In particular, the number of weeks a song charted for and the peak value that it reached on the Billboard Hot 100 charts seem to be good predictors.

5 Interpeting the Results

Based on our analyses, it appears that there are noticeable trends in popular music over time and over genre. Older songs tended to be shorter and of different types of genre than more recent songs. Interestingly, the beats per minute of the song does not seem to be changing too much on average over time.

Furthermore, we observed trends relating the number of weeks that a song charted for and its highest ranking on the Billboard Hot 100 charts. Songs that charted for longer also peaked higher, and are thus what we would consider "popular."

Our analyses that focused on particular artists only touched on the work of five musicians (LeAnn Rimes, Christina Aguilera, Britney Spears, The Backstreet Boys, and N'Sync), out of the 4,670 unique artists in the restricted dataset. Future work could focus on other artists or a further limited dataset, for example considering only songs that reached number 1 on the Billboard Hot 100 or only songs that lasted for at least ten weeks or only songs of a certain genre.

The analyses that we presented herein represent only a small portion of the work than can be conducted to analyze music popularity data. There are other parameters and combinations of parameters in this dataset that could be utilized, in addition to the plethora or other sources about the music industry. For future research, it would be interesting to link a song's position on the Billboard Hot 100 Chart to sales numbers for the song or the revenue that the artist is making off the song. The Billboard ranking is based in part on plays of the song on radio stations, so it would be telling to examine the relationship between popularity and revenue, perhaps using genre to subdivide the songs into groups. We could also use more weeks beyond just the first ten that the song charted for. Additionally, with each new week that passes, a new Billboard Hot 100 chart ranking becomes available, adding to the wealth of data that already exists and offering more opportunities for analysis.

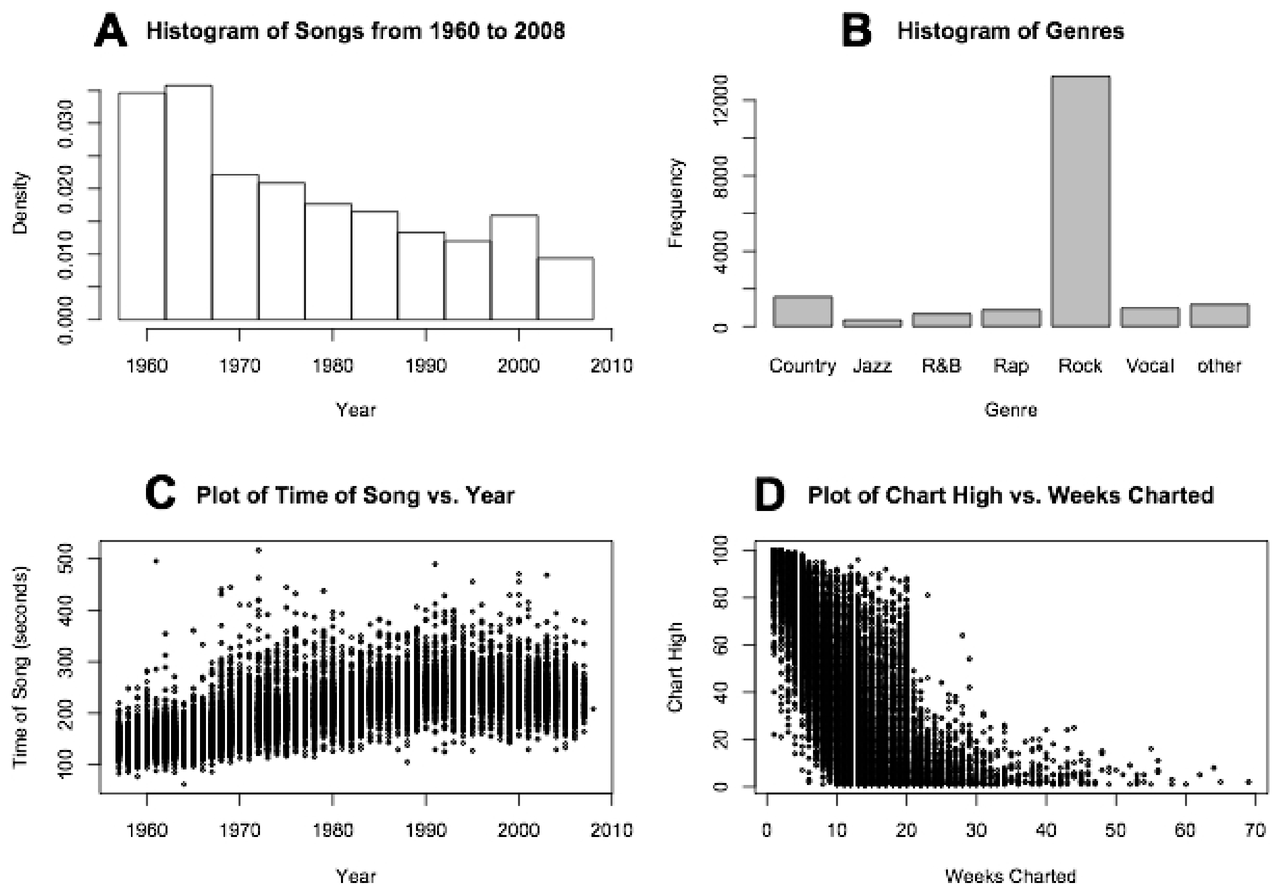

Figure 1: Preliminary exploratory data analysis. A: histogram of year. B: barplot of genre. C: plot of time of song vs. year. D: plot of chart high vs. number of weeks charted.

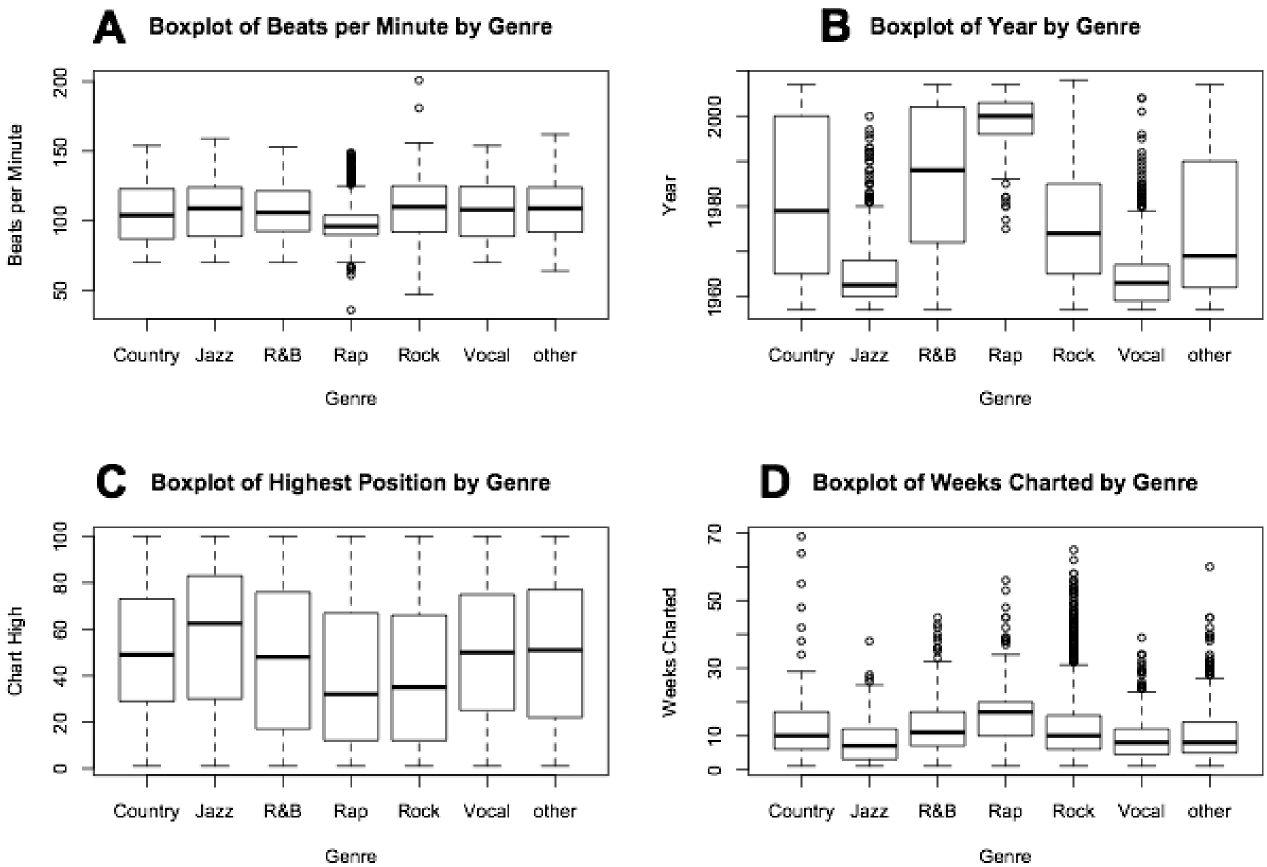

Figure 2: Boxplots of music parameters by genre. A: beats per minute. B: year. C: chart high. D: weeks charted.

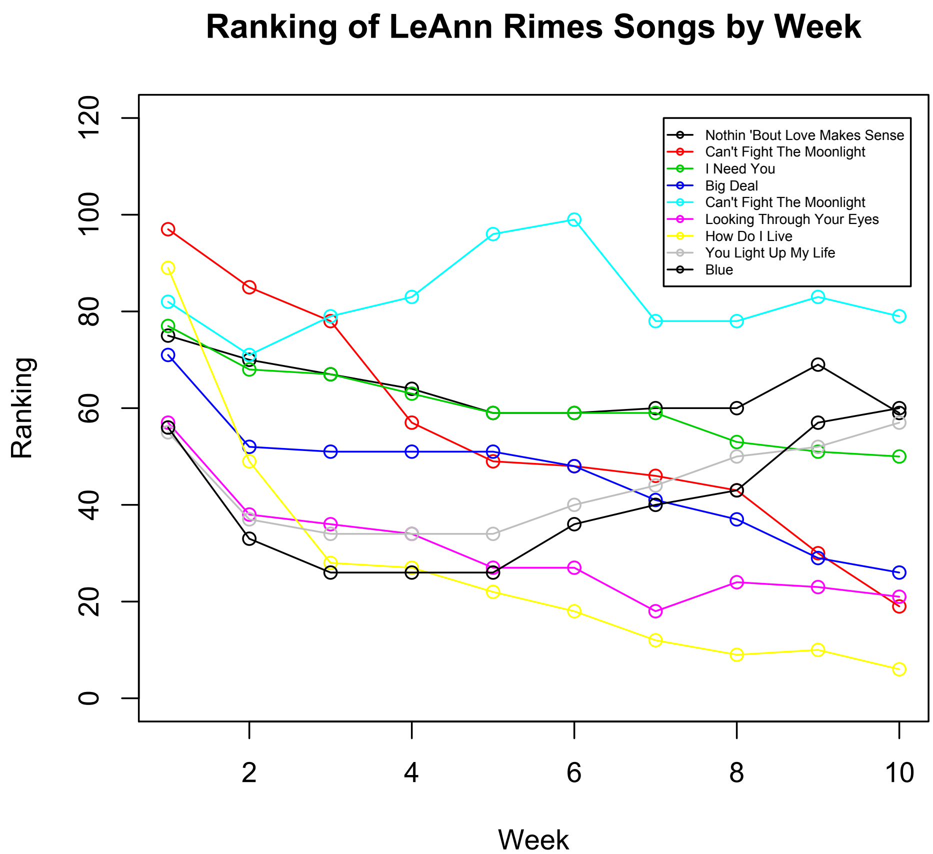

Figure 3: Plot of LeAnn Rimes' songs that reached the Billboard Hot 100 over the first ten weeks that they charted.

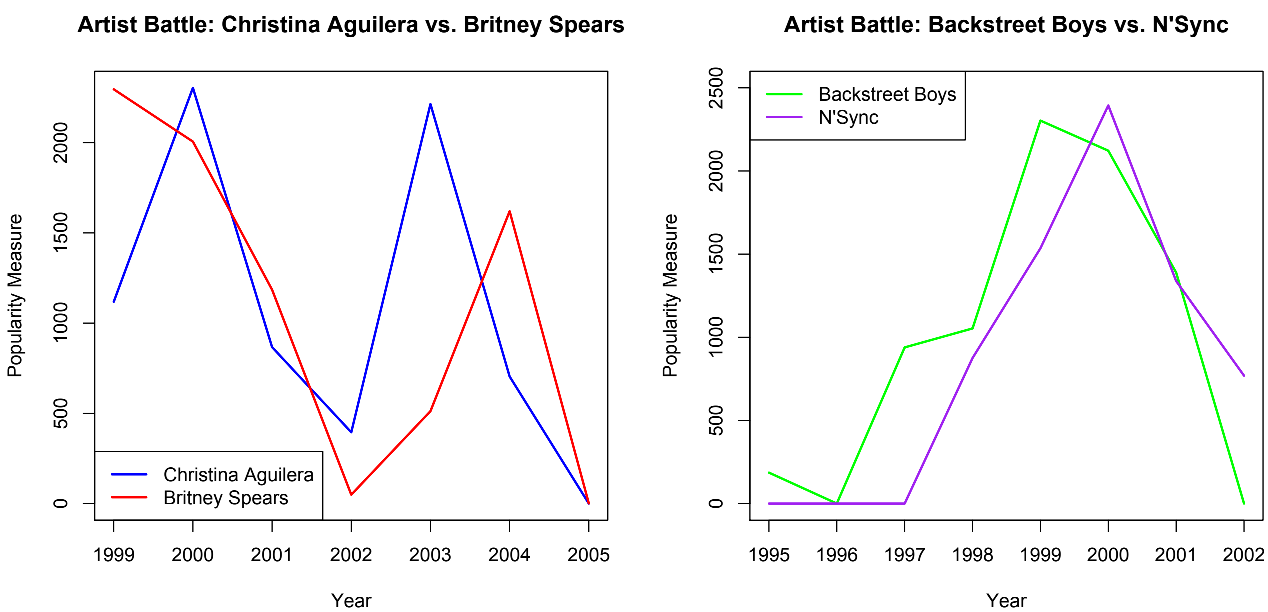

Figure 4: Plots for artist battles by year. Left: Christina Aguilera vs. Britney Spears. Right: Backstreet Boys vs. N'Sync.

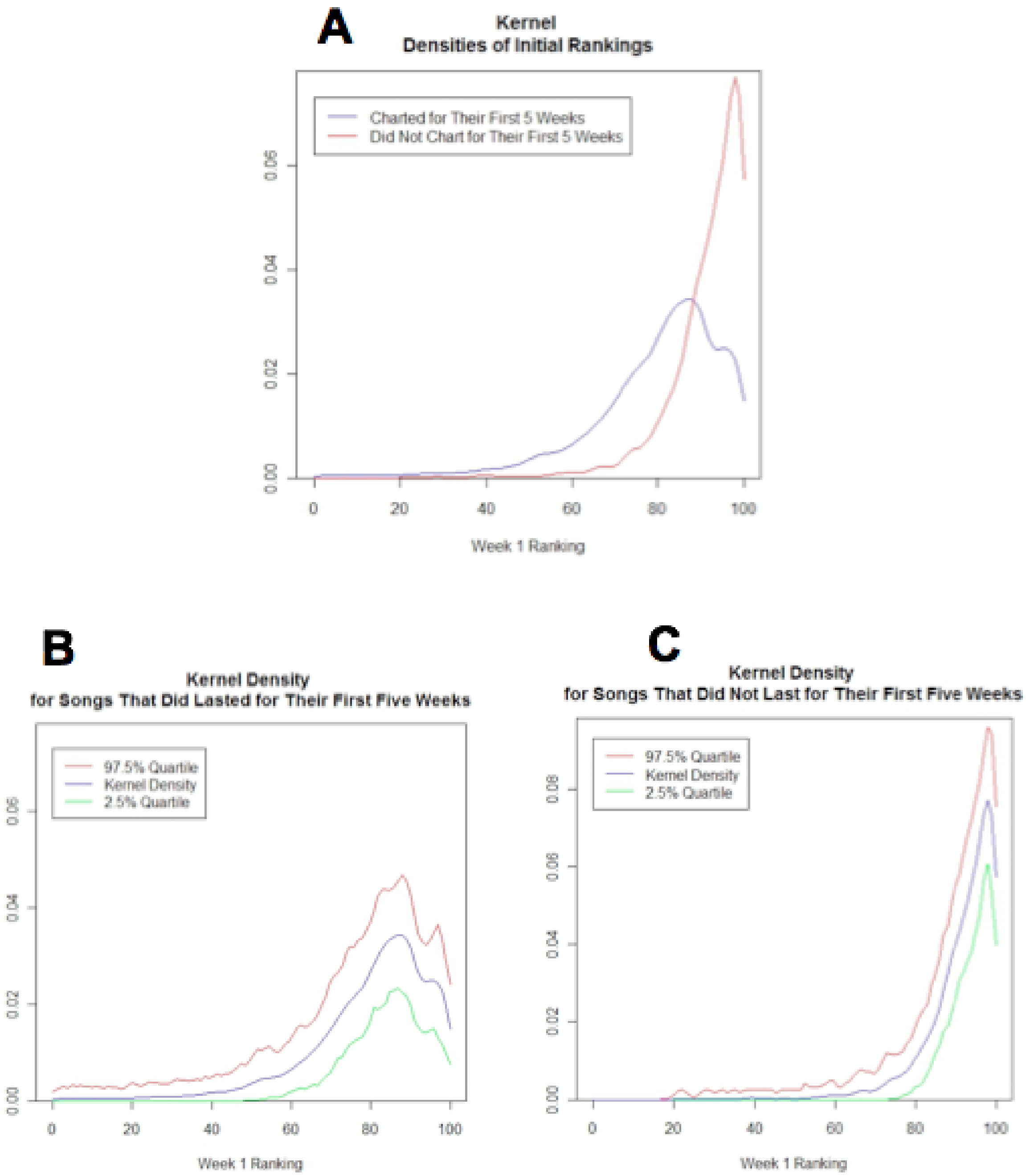

Figure 5: Kernel density plots for week 1 chart position for data split into songs that charted for their first five weeks and songs that did not. A: direct comparison of kernel density estimates for these two categories of songs. B: kernel density estimates and 95% confidence bounds for songs that did last their first five weeks. C: kernel density estimates and 95% confidence bounds for songs that did not last their first five weeks.

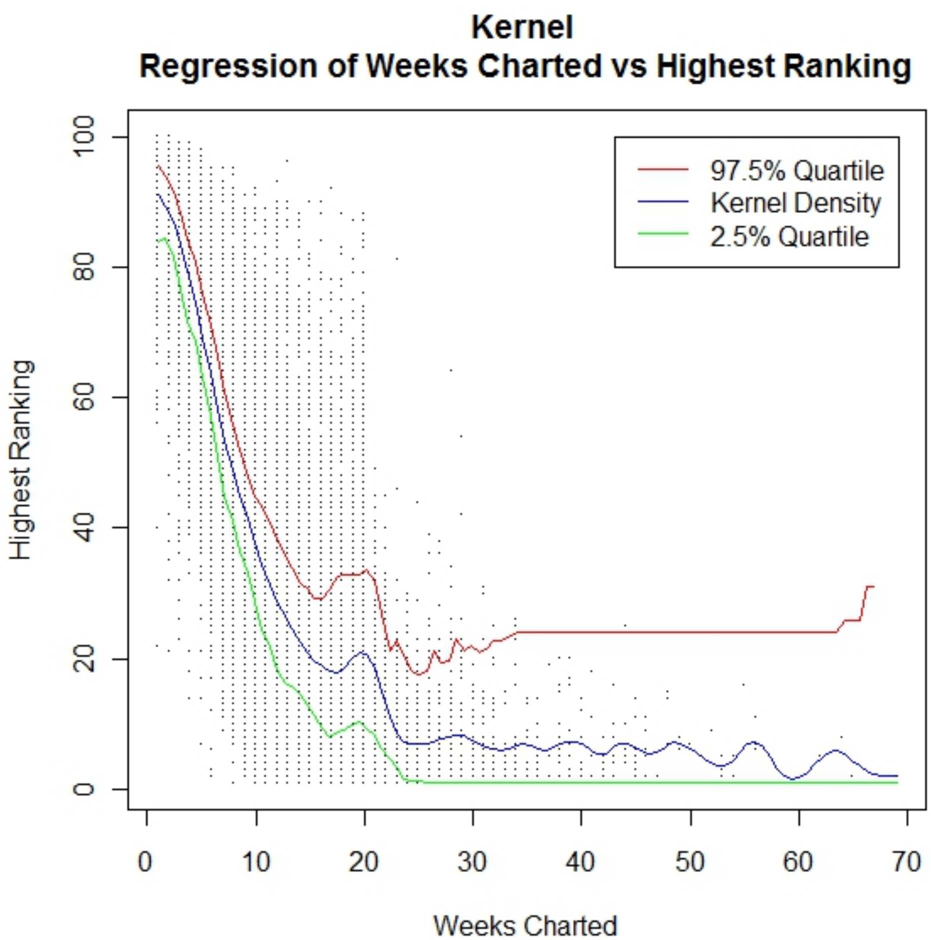

Figure 6: Kernel regression of number of weeks charted versus highest chart position.

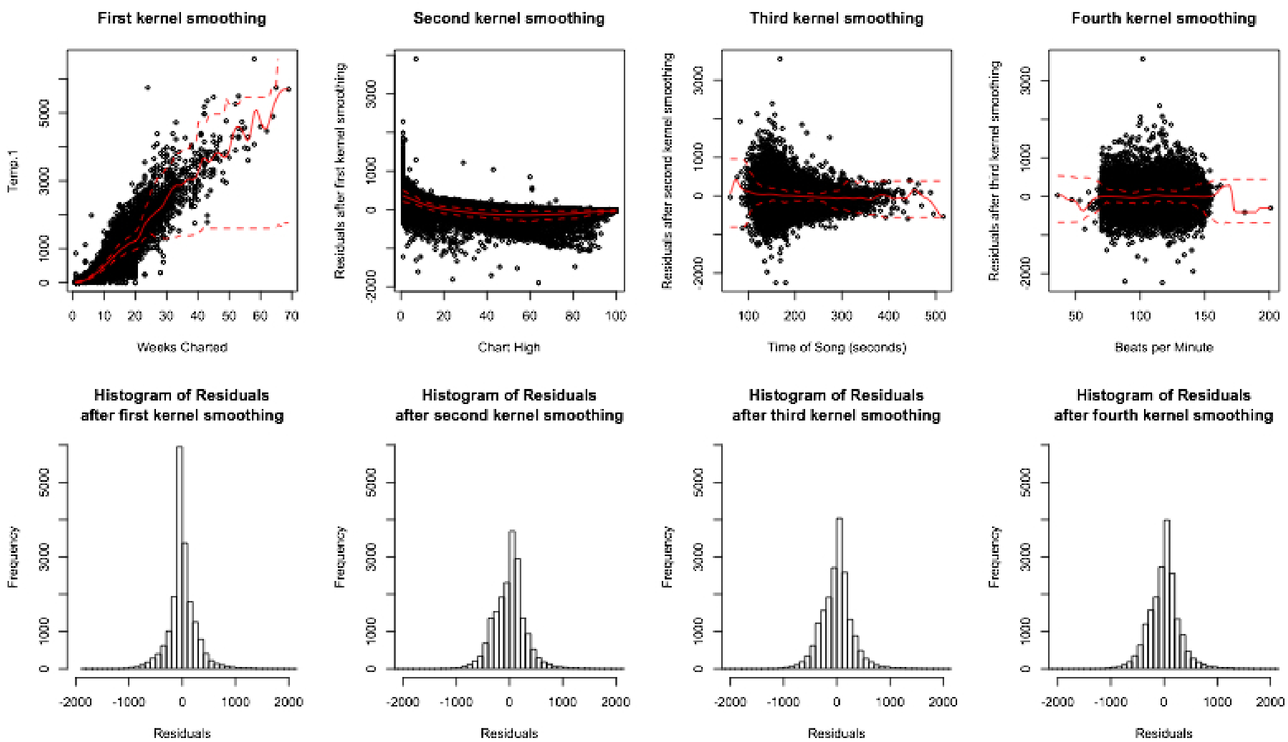

Figure 7: General additive model: multiple kernel regression estimates to predict Temp.1 along with 95% bootstrap confidence intervals. The predictors are (in order): number of weeks charted, chart high, time of song (seconds), and beats per minute.