Music Artist Discovery:
The Digital Road to the Top of Radio

Emily Wright

Department of Statistics

Honors Thesis

Carnegie Mellon University

Pittsburgh, PA

May 2014

Acknowledgements

I would like to acknowledge and thank my advisor, Rebecca Nugent, who has given me unwavering guidance and support in conducting my research. I was continually impressed by her level of commitment throughout the project. I look up to her greatly and aspire to one day emulate her qualities.

I would also like to acknowledge and thank Kelvin Rojas, who was responsible for automating the collection of the online metrics which greatly enhanced the quality of my research.

Lastly, I would like to thank Next Big Sound, who generously provided complete access to their extensive database of artist online metric data.

Abstract:

Radio is a powerful and influential medium with a vast and encompassing audience reaching about 244.5 million consumers a year (Nielsen, 2014). Consequently, a music artist's position on the top charts of radio is a primary measure of success in the music industry. It is of particular interest to record labels to promote and position their artists in a way to ensure they hit the top charts of radio. When identifying potential artists to join their label, record labels may also be interested in assessing how likely an artist is to hit the top charts of radio. This paper is a statistical analysis examining the relationship between an artist's online presence and their appearance on the top charts of radio. It is hypothesized that artists with higher activity online are more likely to reach radio's merit of success. First, radio data and artist online activity data were collected from two sources. The data were then linked using statistical data matching techniques to create one relational database. The online channels examined include Facebook, Wikipedia, Twitter, Youtube, Vevo and SoundCloud. The linked data were then used to fit a logistic regression model with the explanatory variables as summarized time series variables for each online medium. However, this approach was not found to adequately capture the relationships. Instead, a Cox Proportional Hazards model predicting an artist's presence on the top chart of radio was fit. The final findings show evidence urging artists and record labels to place increased attention on their use of Twitter, Vevo and SoundCloud. After establishing an online presence, the expected time period of success is between 1 and 3 years.

Section 1: Introduction

Today there are a multitude of ways that consumers can discover new music. Traditionally, radio has been, and still is, the most common channel by which consumers first hear a new artist or song (Nielsen, 2012). However, in the new age of technology, the music discovery process is further aided by online streaming services such as Pandora[1], Spotify[2], and Last.fm[3]. In essence, each service is simply a platform for consumers to stream music. On a more complex level, each service can be differentiated by how they aid the consumer in the music discovery process. Pandora uses content based algorithms which analyze the attributes of musical content and suggest songs with similar attributes (Pacula). Spotify and Last.fm implement collaborative filtering algorithms which suggest new music based on what friends or similar listeners are also listening to (Bernhardsson, 2013).

Examining all paths of music discovery as a whole, each medium can be classified by the required level of user engagement. Spotify and Last.fm require the most engagement, Pandora requires moderate to little engagement, and radio requires the least engagement (*"The Zero Button Music Player"*, 2014). The listeners who actively engage in music discovery are most likely to use online streaming services and are known as Savants and Enthusiasts. These listeners are estimated to be about 35% of the U.S. population (Orpheus, 2011). The other 65% are Casuals and Indifferents who put minimum effort into finding new music. Not surprisingly, the Casuals and Indifferents, who are the majority of consumers, still discover new music passively through radio. Therefore, in order for music artists to reach considerable stardom, an artist must be played frequently on radio. The most successful artists are often those who are on the top charts of radio with the most airplays per week.

In an effort to increase exposure, artists now often choose to promote and share their music through online means. Sharing music online is an easy and low cost way for artists to promote themselves and does not require affiliation with a record label. Some of the most

[1] Pandora is an online streaming and music recommendation service launched in 2005 with more than 250 million users. The site can be access at http://www.pandora.com/

[2] Spotify is an music streaming software service launched in 2008 and in 2013, reached about 24 million active users. The site can be accessed at https://www.spotify.com/us/

[3] Last.fm is an online music recommendation service found in 2002 with a about 30 million users. The site can bse accessed at http://www.last.fm/

common online avenues chosen by artists include Facebook[4], Wikipedia[5], Twitter[6], Youtube[7], Vevo[8], and SoundCloud[9]. Once an independent artist's online presence is established, record labels may come into the picture leading to increased exposure and opportunity to be aired on radio. As author Bob Lefsetz from Variety - an entertainment news source states, "Radio is rocket fuel that can propel that which is already successful [online] into the stratosphere [of mass discovery]," (Lefsetz, 2013). Therefore we hypothesize that there may be a relationship between artists' online activity and their likelihood of being ranked on the top charts of radio. Capturing and understanding this relationship will potentially aid record labels in choosing which artists to sign and how to best to promote them. Further independent artists can also learn how best to promote themselves.

In order to understand this relationship between an artist's online presence and their likelihood of being ranked on a radio chart, we collected data from two sources pertaining to artist online presence and radio airplay. The online avenues examined were Facebook, Wikipedia, Twitter, Youtube, Vevo and SoundCloud. The data were then linked using statistical data matching techniques to create one relational database. The linked data were then used to fit a logistic regression model with the explanatory variables as summarized time series variables. However, logistic regression was a poor choice for modeling longitudinal data. Instead, a Cox Proportional Hazards model was fit with an artist's appearance on the top charts of radio as the event of interest.

The complete analysis is detailed in the following sections. First, *Section 2: Creating the Artist Relational Database* details how the data were collected, cleaned, and statistically linked to create one large relational database. *Section 3: Exploration of Online Metric Time Series Variables* summarizes the main characteristics of the online metric data. *Section 4: Logistic*

[4] Facebook is a highly popular online social network which launched in 2004 and to date has about 1.2 billion active users. The site can be accessed at https://www.facebook.com

[5] Wikipedia is an online and publicly edited encyclopedia launched in 2001 and today receives about 85 million unique visitors per month. The sire can be accessed at https://www.wikipedia.org

[6] Twitter is a social network of status updates which launched in 2006 and today has over 200 million users. The site can be accessed at https://twitter.com/

[7] Youtube is a video sharing platform launched in 2005 and currently receives over 2 billion views per day. The sire can be accessed at https://www.youtube.com/

[8] Vevo is an all-premium music video and entertainment platform launched in 2009 and today has about 5 billion views per month. The sire can be accessed at http://www.vevo.com/

[9] SoundCloud is an online audio distribution platform launched in 2007 and today reaches about 200 million unique users. The sire can be accessed at https://soundcloud.com/

Regression – Using Summarizing Online Metric Time Series details how the explanatory variables were transformed into summarizing time series variables suitable for logistic regression. The results are then presented and discussed in regards to final modeling choices. *Section 5: Imputation of Missing Data* explains several methods for replacing missing data and the estimated accuracy. *Section 6: Cox Proportional Hazards Model* details the final modeling choices and results. Finally, *Section 7: Discussion* explores the limitations of the data, and discusses future improvements for further research.

Section 2: Creating the Artist Relational Database

In order to address our hypothesis, we needed radio and online activity data for a set of artists. Currently there does not exist a publically available data source with both. Therefore, we collected the radio data and online activity data from two separate sources and linked them to create one large relational database suited for this analysis. First, radio airplay data was collected from Digital Radio Tracker[10] via weekly online reports of national radio airplay. Next, data pertaining to artists' online activity were generously provided by the company, Next Big Sound[11], whom offers analytics of the online music industry. Among other analytics, the primary goal of Next Big Sound is to predict which artists are to become popular i.e., "The Next Big Sound". Combining the two datasets allowed us to identify which of the artists predicted by Next Big Sound to become successful actually made it on a top radio chart. For reference, these artists will be referred to as "discovered" artists throughout the paper. The following sections will outline how the data were collected, cleaned, formatted, and statistically linked to create the final relational database.

Section 2.1: Radio Airplay Data

Weekly radio airplay data from 5,000+ terrestrial radio channels and online radio channels were collected from Digital Radio Tracker[10] and were web-scraped using Excel. The radio airplay data is structured in separate charts for each genre. The genres include the Top 200 songs, the Top 50 Independent artist songs, the Top 50 Pop songs, the Top 50 R&B and Hip-Hop songs, the

[10] http://digitalradiotracker.com/chart.html
[11] https://www.nextbigsound.com

Top 50 Rock songs, the Top 50 Country songs, the Top 50 Americana songs, and the Top 50 Adult Contemporary songs. The Top 200 songs and the Top 50 Independent artist song charts span the entirety of 2013. However, all other genres were not reported until July of 2013. The radio charts for each genre consist of seven fields: the artist's rank on the chart determined by airplay, the artist's name, the song title, the number of airplays[12] within the week, and the date of the radio chart. In total there are 19,550 observations across the 260 weekly charts. A sample of the raw radio data from the Top 200 genre is provided in Table 1.

Table 1: Top 200 Radio Chart

Rank	Artist	Song Title	Airplay	Date
1	RIHANNA	DIAMONDS	5768	1/5/13
2	BRUNO MARS	LOCKED OUT OF HEAVEN	4500	1/5/13
3	FLO RIDA	I CRY	3765	1/5/13
4	MAROON 5	ONE MORE NIGHT	3339	1/5/13
5	KE$HA	DIE YOUNG	2875	1/5/13
6	FUN.	SOME NIGHTS	2629	1/5/13
8	P!NK	TRY	2124	1/5/13
35	SMOKE	I AIN'T HIDING (W/ T-PAIN)	1167	1/5/13
123	MEKA ARPEGE & BARSHAUN	KEYS OF HOPE	707	1/5/13

The original format of the radio data did not allow for accurate linkage with the Next Big Sound predictions due to several complications. First, all unique artists could not easily be identified. There were no artist ids, and the text strings describing an artist varied. For instance, in one record, the artist text string may have an accent such as "Emeli Sandé" and in another record the same artist may be listed as "Emeli Sande" without an accent. Further, there was no uniform format for listing multiple artists per song. Sometimes an additional artist is listed in the song title such as "I AINT HIDING (W/ T-PAIN)" where T-Pain is the featured artist. Other

[12] Radio airplays are the number of times an artist is played over all radio stations

times, multiple artists are listed under the primary artist column such as "MEKA ARPEGE & BARSHAUN" where Meka Arpege and Barshaun are two individual artists. Therefore, the artist names and song titles were parsed to extract the individual artist names. However, there was also no uniform format or abbreviation distinguishing these unique artists. For the two artists previously mentioned, "W/" and "&" both signal an additional artist. However, in another case, "FEAT." signals the featured artist Kimbra in the song title "SOMEBODY THAT I USED TO KNOW (FEAT. KIMBRA)" by Gotye. Therefore, the following abbreviations "feat," "feat.," "ft.," "ft," "&," w/", or "and" were all searched for while text string parsing. While the parsing was automated, the extraction process still required human inspection as the appearance of the abbreviation does not necessarily mean there are two distinct artists. For example, "Beyonce and Lady Gaga" are two distinct artists while "Florence and The Machine" is one artist/band.

After all artist names were extracted, the radio data was cleaned and reformatted. Each artist name was converted to all lower case letters. Doing so allowed for more accurate text string matching, as many text string matching algorithms can be case sensitive (the statistical record linkage techniques implemented will be further discussed in Section 2.3). Next, all individual radio charts for the different genres were combined into one large chart with additional information. The additional columns include: Genre, Artist 1-5, and Artist ID 1-5. Genre was determined by which radio chart the song came from, Artist 1-5 are the extracted unique artist names. Artist ID 1-5 is the ID assigned by Next Big Sound for each artist. Artists who were not part of the Next Big Sound predictions were not assigned an ID as they are not the focus of this analysis. A sample of the cleaned radio data is provided in Table 2. For the week of January 5th 2013, the top ranking artist with the highest number of airplays was "rihanna." Her song "diamonds" was played 5,768 times across 5,000+ radio stations in the U.S. Rihanna was not predicted by Next Big Sound because she was already well known at the beginning of the time period being analyzed. Therefore, she did not receive a Next Big Sound ID number. The sixth ranking artist was the band "fun." The band's song "some nights" was played 2,629 times across 5,000+ radio stations in the U.S. Fun. was predicted by Next Big Sound and has the Next Big Sound ID number 268129.

Table 2: Cleaned Radio Data

Rank	Artist 1	Artist 2	Artist 3	Artist 4	Artist 5	Song Title	Airplay	Date	Genre	Artist 1 ID	Artist 2 ID	Artist 3 ID	Artist 4 ID	Artist 5 ID
1	rihanna	NA	NA	NA	NA	diamonds	5,768	1/5/13	Top 200	NA	NA	NA	NA	NA
2	bruno mars	NA	NA	NA	NA	locked out of heaven	4,500	1/5/13	Top 200	NA	NA	NA	NA	NA
3	flo rida	NA	NA	NA	NA	i cry	3,765	1/5/13	Top 200	NA	NA	NA	NA	NA
4	maroon 5	NA	NA	NA	NA	one more night	3,339	1/5/13	Top 200	NA	NA	NA	NA	NA
5	ke$ha	NA	NA	NA	NA	die young	2,875	1/5/13	Top 200	NA	NA	NA	NA	NA
6	fun.	NA	NA	NA	NA	some nights	2,629	1/5/13	Top 200	268129	NA	NA	NA	NA
7	the lumineers	NA	NA	NA	NA	ho hey	2,379	1/5/13	Top 200	307085	NA	NA	NA	NA
8	p!nk	NA	NA	NA	NA	try	2,124	1/5/13	Top 200	NA	NA	NA	NA	NA

Section 2.2: Artist Online Presence Data

The online presence data was generously provided by Next Big Sound. As mentioned, Next Big Sound offers analytics of the online music industry. Currently they track and report the usage and activity of over 1 million artists' online sources such as Facebook, Wikipedia, and so on. Further they provide demographic information of artists' listeners and report major events such as shows or tours. The primary goal of Next Big Sound is to predict which artists are going to become popular i.e., "The Next Big Sound" based on their online activity. They publish new predictions every week based on how fast artists' online activity is increasing. The artists chosen to be analyzed for this work are an accumulation of the weekly artist predictions Next Big Sound has published from August 2010 through December 2013. The predictions were web-scraped in R[13] using the XML[14] and RCurl[15] packages written by Duncan Lang. The data consists of six fields: the artist ID number, the artist's rank on the chart determined by acceleration of online activity, the artist name, and the date the prediction was made. The Next Big Sound artist ID number is a primary key linking all the artists in the entire relational dataset. Overall, there are 3,213

[13] http://www.R-project.org/
[14] http://CRAN.R-project.org/package=XML
[15] http://CRAN.R-project.org/package=RCurl

observations. A sample of the Next Big Sound predictions is provided in Table 3. For example, during the week of December 26th, 2013 the artist with the fastest accelerating online activity was Skylar Stecker and thus was predicted to become the "Next Big Sound."

Table 3: Next Big Sound Artist Predictions

NBS ID	Rank	Artist	Date
541208	1	Skylar Stecker	12/26/13
338298	2	Nom De Strip	12/26/13
473877	3	Les Castizos	12/26/13

For easier linking with the radio play data, each Next Big Sound artist text string name was converted to all lower case letters. Next, the predictions were de-duplicated by removing artist ID numbers repeated on different weeks. The most recent prediction date for a duplicated artist was kept. Although the artists were de-duplicated based on the Next Big Sound ID, during human inspection, it was found that two Next Big Sound artists had the exact same name, "DJ Drama." Initially this was worrisome as our data matching approach relies on the artist name. Therefore, when matching it is impossible to automatically determine which "DJ Drama" from Next Big Sound belongs to a "DJ Drama" observation in the radio data. Therefore, the comparison required additional human inspection by researching the song title from the radio data to determine its correct artist. When exploring this issue, we found that Next Big Sound had incorrectly assigned two artist IDs to the sole unique artist, "DJ Drama." While this quells the worries of not being able to automatically classify "discovered" artists, it does prompt further questions. In particular, how many other artists accidently received two IDs? To find these artists, the original Next Big Sound predictions were again de-duplicated by the artist ID. Duplicated artist names were then found by exact string matching among the Next Big Sound predictions. Each match was then examined by hand. In addition to "DJ Drama," one more artist, "Mavado," was found to have incorrectly received two Next Big Sound ID numbers. The duplicates were then removed and the later prediction date was kept. In our exploration, we did not find any unique artists who had exactly the same name as another. In the end, a total of 3,095 artists were collected.

The metrics for the online avenues, or channels, for each predicted artist were then collected from Next Big Sound in collaboration with Kelvin Rojas (Logic and Computation, Department of Philosophy, Carnegie Mellon University). The metrics were collected in a written Java program using the org.dom4j[16] and org.apache.http[17] packages. In the program, the artist ID number was used to download the artist profile with all metrics to an XML file. All the XML files were then converted to one large CSV file. In the first collection attempt, the top key metrics presented by Next Big Sound were collected: Facebook page likes, Wikipedia page views, Twitter followers, Youtube views, Vevo plays, and SoundCloud plays. The metrics span January 2010 to December 2013 with the date recorded as a UNIX time stamp[18]. Each daily observation reports the cumulative value to date. However, Wikipedia views were reported as net daily values and were converted to the cumulative values. Finally, due to technical glitches data was only available for 2,933 of the total 3,095 Next Big Sound artists.

Table 4: Next Big Sound Online Metric Data

ID	Artist	UNIX Time Stamp	Facebook Page Likes	Wikipedia Page Views	Twitter Followers	Youtube Plays	Vevo Plays	SoundCloud Plays
2375	10 Years	1263945600	NA	452	NA	NA	NA	NA
301919	Damares	1305676800	NA	97	NA	921	NA	NA
294946	Restart	1349913600	NA	389	684,223	76,264,445	NA	NA

Section 2.3: Data Matching and Classification

Once each source was cleaned, we then linked the two data sources in order to identify the "discovered" artists. Again discovered artists are artists who where predicted by Next Big Sound to become popular and were also aired on one of the top charts of radio. Even after data cleaning, the process of identifying the discovered artists proved to be more challenging than one would initially assume. Simple exact string matching on the artist names was inaccurate because it failed to acknowledge typographical errors. For example, exact matching fails to match the

[16] http://dom4j.sourceforge.net/dom4j-1.6.1/apidocs/org/dom4j/package-summary.html
[17] http://hc.apache.org/httpcomponents-core-ga/httpcore/apidocs/org/apache/http/package-summary.html
[18] A unix time stamp is the number of second since Jan 1st, 1970

artist names “fun.” and “fun” or “p¡nk” and “pink”. Therefore, fuzzy data matching, in which two text strings are not exactly the same but are very similar, was required to find the discovered artists. While matching could have been done by hand, it is not realistic to hand match all 2.8 million comparisons (≈ 2,933 artist predictions x 949 unique radio artist text string names, the radio artist names were reduced based on exact matching to reduce number of necessary comparisons). Instead, we implemented statistical record linkage techniques. One commonly used method is an unsupervised approach that estimates the probability of a link between two records using a ratio of agreement patterns across fields (Fellgi, Sunter, 69). Because there are 2.8 million comparisons we instead used a supervised learning approach to minimize the required computation. This decision required labeled data, fit to a binary classification model, to predict the probability of a link. The classification models evaluated were logistic regression and a classification tree.

Section 2.3 A: Text String Similarity Scores

For each comparison we calculated and compared two text string similarity scores: the Jaro-Winkler and the Levenshtein which are among the most common and well established text string similarity scores currently used in the field. The Jaro-Winkler similarity score is between 0 and 1 with higher values corresponding to more similar text strings. An exact match receives a value of 1. The calculation, “accounts for the lengths of the two strings and partially accounts for the types of errors –insertions, omissions, or transpositions – that human beings typically make when constructing alphanumeric strings,” (Herzog 131, 2007). An insertion is when a character is added, an omission is when a character is deleted, and a transposition is when two characters are switched. For example, when comparing the two strings “music” and “musci” there is one transposition switching the “i” and the “c.” The score is calculated as follows:

$$\Phi_j(s_1, s_2) = \frac{W_1 c}{L_1} + \frac{W_2 c}{L_2} + \frac{W_t(c-\tau)}{c}$$

where s_1 is the first text string, s_2 is the second text string, W_1 is the weight assigned to the first string, W_2 is the weight assigned to the second string, W_t is the weight assigned to the transpositions, c is the number of characters that the two strings have in common, L_1 is the length

of the first string, L_2 is the length of the second string, and τ is the number of characters that are transposed. The weights W_1, W_2, and W_t must sum to one and have default values of 1/3 each. If $c = 0$, (i.e., no characters in common) then the Jaro-Winkler score is zero (Herzog 132, 2007). A small sample of the artist comparisons and their Jaro-Winkler scores is provided in Table 5, and the distribution of all the Jaro-Winkler scores is shown in Figure 1.

The Levenshtein similarity score is a transformed edit distance which minimizes the number of edits to match one string to another by insertions, omissions, or substitutions (Bilenko, 2003). The calculated score is between 0 and 1 with higher values corresponding to more similar text strings. An exact match receives a value of 1. Mathematically, the Levenshtein distance is defined as:

$$D(s,t,i,j) = min \begin{cases} D(s,t,i-1,j-1) & if\ s_i = t_j, and\ you\ copy\ s_i\ to\ t_j \\ D(s,t,i-1,j-1)+1 & if\ you\ subsitute\ t_j\ for\ s_i \\ D(s,t,i,j-1)+1 & if\ you\ insert\ the\ letter\ t_j \\ D(s,t,i-1,j)+1 & if\ you\ delete\ the\ letter\ s_i \end{cases}$$

where s is the first string, t is the second string, i is the ith letter in the first string, and j is the jth letter in the second string (Bilenko, 2003). This distance is then transformed into a similarity score by subtracting the distance normalized by the length of the longest string from 1 as shown below. A small sample of the artist comparisons and their Levenshtein scores is provided in Table 5 and the distribution of Levensthein scores is shown in Figure 1.

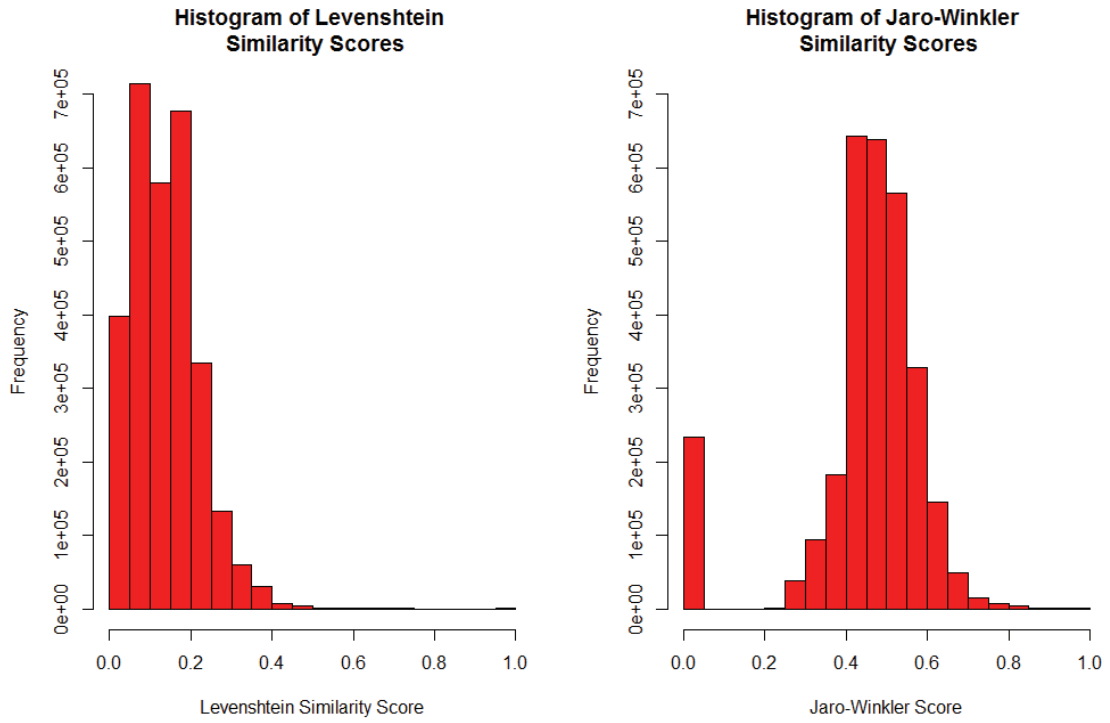
$$\text{Levenshtein Similarity } = 1 - \frac{D(s,t,i,j)}{\max(|s|,|t|)}$$

Table 5: Text String Similarity Scores

Next Big Sound Artist	Radio Artist	Levenshtein Score	Jaro-Winkler Score
doble man	hrc	0.000	0.000
dayan	keith urban	0.182	0.000
blessed by a burden	joni mitchell	0.052	0.253
zulu winter	growdy	0.090	0.419
the popopopops	3 doors down	0.214	0.539
remady	angel mary	0.400	0.605
lance herbstrong	lana del rey	0.375	0.704
the herbaliser	the grass roots	0.333	0.821
deuce	deuce.d	0.714	0.943
atlas genius	atlas genius	1.000	1.000

In Table 5, the first comparison of "doble man" and "hrc" received a zero for both scores because there are no characters in common. Comparing the two scores as the artist names become more similar, we see that the Levenshtein similarity is more strict and produces lower scores compared to the Jaro-Winkler score. For example, when comparing "the herbaliser" and "the grass roots" the two names intuitively do not appear to be similar and the Levenshtein score rightfully gives the comparison a score of 0.33. However, the Jaro-Winkler gave the same comparison a score of 0.82 which seems to be overly optimistic. For the last comparison of "atlas genius" and "atlas genius" both similarity scores produced a value of 1 since the comparison is an exact match. In Figure 1, we see that the distribution of Levenshtein similarity scores is unimodal with the mode at 0 and skewed right with small spread. In comparison, the distribution of Jaro-Winkler scores is bimodal with larger spread. The first mode occurs at 0 and the second mode occurs at 0.50.

Figure 1: Distributions of Text String Similarity Scores



Section 2.3 B: Hand-Labeled Subset of Artist Name Comparisons

After both similarity scores were calculated for the entire dataset of roughly 2.8 million comparisons, a subset of 650 comparisons was hand-labeled as training data, 0 for a nonmatch and 1 for a match. The 650 comparisons were chosen by randomly sampling from different ranges of Jaro-Winkler scores. We choose to use Jaro-Winkler score ranges rather than Levenshtein similarity score ranges because there are fewer Jaro-Winkler scores equal to zero and the Jaro-Winkler scores are more variable shown by the larger spread (Figure 1). This implies that the randomly sampled subset will be more variable and thus more representative of all the comparisons. However, instead of simply randomly sampling from the entire distribution of Jaro-Winkler scores, we purposely sampled from different ranges of Jaro-Winkler scores to capture an appropriate proportion of matches and nonmatches. Understand that in the entire dataset of comparisons, the number of matches is likely very low. At a maximum, there are 949

radio artists of the total 2,933 Next Big Sound predictions (949 is the number of unique artist text string names in the radio data). In reality, the number of unique radio artists is most likely lower than 949 since some artists are listed more than once, with varying text strings such as "Emeli Sandé" and "Emeli Sande." If this proportion of matches is replicated in a simple sample, the model may not converge because there are not enough positive matches. Therefore, using ranges of scores allowed us to have some control over the number of possible matches in the subset. The composition of the sample is shown in Table 6.

Table 6: Composition of Hand-Labeled Subset

Jaro-Winkler Score	Number of Comparisons
$JW = 0$	200
$0 < JW < 0.482$	185
$0.482 \leq JW < 0.82$	185
$0.9 \leq JW < 1$	50
$JW = 1$	30
$0 \leq JW \leq 1$	650

We sampled a total of 650 comparisons. This sample size was reasonably large to ensure reliable results and was still of a manageable size to be labeled by hand. First, 200 comparisons with a score of 0 were collected as the number of comparisons with a score of zero is prevalent in the distribution of Jaro-Winkler scores. We assumed that these comparisons are nonmatches. Next, 30 exact matches with a value of 1 were sampled to ensure a minimum number of matches as there are only 71 artist comparisons in the entire dataset which are exact matches. Then additional scores were collected by sampling 185 comparisons from above and below the median Jaro-Winkler score of 0.482 (the median Jaro-Winkler score was calculated not including the 0 or 1 values). In total there are 370 comparisons from the center of the distribution. Reasonably, this is the largest range sample as the majority of the comparisons fall here. However, our sample failed to include values higher than 0.82. Therefore, we sampled an additional 50 values with Jaro-Winkler scores between 0.9 and 1, values that indicate highly likely matches. Finally each comparison was hand labeled a match or nonmatch. The subset was found to contain 36 matches and 614 nonmatches.

Section 2.3 C: Logistic Classification Models

We first applied logistic regression models to the hand-labeled dataset. The model estimates the probability of two artists being a match by predicting the log odds. The odds of a match is the probability of two artist names being a match divided by the probability of not being a match. The log odds are modeled as a linear function of the text string similarity scores (see below).

$$\log odds = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{JaroWinkler} + \beta_2 \text{Levenshtein}$$

$$\hat{p} = \frac{e^{\beta_0 + \beta_1 \text{JaroWinkler} + \beta_2 \text{Levenshtein}}}{1 + e^{\beta_0 + \beta_1 \text{JaroWinkler} + \beta_2 \text{Levenshtein}}}$$

To start, logistic regression models with each individual similarity metric as the sole explanatory variable were fit. The outputs of each model are shown in Tables 7 and 8. The cutoff probability to classify matches was chosen by evaluating cutoff values from 0.50 to 0.95 by increments of 0.05. Each threshold was evaluated by calculating the accuracy, sensitivity, and specificity of each model, as defined below.

$$\text{Accuracy} = \frac{\text{\# of correct classifications}}{\text{total \# of comparisons}}$$

$$\text{Sensitivity} = \frac{\text{\# of comparisons correctly classified as matches}}{\text{total \# of true matches}}$$

$$\text{Specificity} = \frac{\text{total \# of comparisons correctly classified as nonmatches}}{\text{total \# of true nonmatches}}$$

The threshold choice for each model was then narrowed down further by identifying the two thresholds where there was a major change in the performance measures. Between these two thresholds, additional thresholds at increments of 0.01 were examined. For the Jaro-Winkler model, the thresholds between 0.80 and 0.85 by increments of 0.01 were evaluated. The model performed best by applying a threshold of 0.83. For the Levenshtein model, the threshold choices were again examined between 0.80 and 0.85 by increments of 0.01. The model performed best by applying a threshold of 0.85. Both models performed well with 99.7% accuracy, 94.4% sensitivity, and 100% specificity. It is not surprising that the models performed similarly since the two text string similarity scores have a correlation of 0.82. Further, the Levenshtein cutoff probability may be slightly lower than the Jaro-Winkler cutoff since the scores also tend to be lower (Figure 1).

Table 7: Univariate Logistic Classification – Levenshtein Similarity

$$\hat{p} = \frac{e^{-23.13+28.05*\text{Lev}}}{1+e^{-23.13+28.05*\text{Lev}}}$$

	Coefficient	Standard Error	P-value
Intercept	-23.13	8.67	0.00006
Levenshtein	28.05	10.43	0.00007

Table 8: Univariate Logistic Classification – Jaro-Winkler Similarity

$$\hat{p} = \frac{e^{-123.98+130.14*\text{JW}}}{1+e^{-123.98+130.14*\text{JW}}}$$

	Coefficient	Standard Error	P-value
Intercept	-123.98	37.73	0.00102
Jaro-Winkler	130.14	39.93	0.00112

Given the high correlation of 0.82 between the Jaro-Winkler score and the Levenshtein score the two scores may be multicollinear. As such we are interested in understanding how the model will be impacted by using both of the similarity scores at once. The results of this model are shown below in Table 9. Additionally, a heat map showing the relationship between the Jaro-Winkler score, Levenshtein similarity score, and the probability of the comparison being a match is shown in Figure 2. Dark maroon areas have high probability of a comparison being a match while dark blue areas have low probability of being a match. From the heat map we see that comparisons with a Jaro-Winkler similarity score of about 0.90 or higher are most likely to be a match. Interestingly, for the entire range of Levenshtein scores there is no distinct cut off where comparisons are definitely a match. Accordingly, the Jaro-Winkler score is statistically significant at the 10% level while the Levenshtein score is insignificant. Next, the threshold was again chosen by evaluating cutoff values from 0.50 to 0.95 by increments of 0.05 and further looking at thresholds between 0.80 and 0.85 by increments of 0.01. A threshold of 0.83 was found to balance the three performance measures resulting in an accuracy of 99.7%, a sensitivity of 94.4%, and a specificity of 100%. Notice the results are the same as the univariate models.

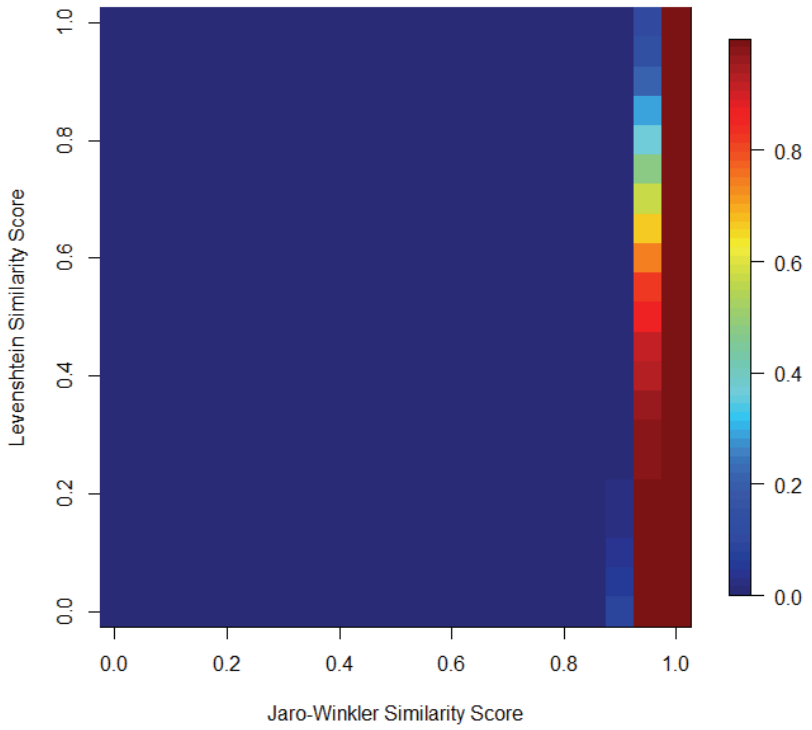
Table 9: Bivariate Logistic Classification – Jaro-Winkler and Levenshtein

$$\hat{p} = \frac{e^{-151.424+165.609\text{JW}-8.026\text{Lev}}}{1+e^{-151.424+165.609\text{JW}-8.026\text{Lev}}}$$

	Coefficient	Standard Error	P-value
Intercept	-151.424	80.304	0.0593
Jaro-Winkler	165.609	98.439	0.0925
Levenshtein	-8.026	18.923	0.6714

Figure 2: Heat Map of Bivariate Logistic Regression Fitted Probabilities

$$\hat{p} = \frac{e^{-151.424+\mathbf{1}65.609\text{JW}-8.026\text{Lev}}}{1+e^{-151.424+\mathbf{1}65.609\text{JW}-8.026\text{Lev}}}$$



Examining the results further we find that all three models classified the same 616 nonmatches and 34 matches in the hand-labeled subset with about 99.7% accuracy, 94.4% sensitivity, and 100% specificity. In total, there were two false negatives and zero false positives. Specifically, the models failed to match "ty dolla $ign" with "ty dolla " which received a Jaro-Winkler score of 0.94 and a Levenshtein score of about 0.69. The comparison received probabilities of 0.14 for the univariate Levenshtein model, 0.02 for the univariate Jaro-Winkler model, and 0.17 for the bivariate model. The error is understandable as an entire word is missing between the two names. Realistically, this type of error can only be overcome by improved data quality.

Next, the models failed to match "fun." with "fun" which received a Jaro-Winkler score of 0.94 and a Levenshtein score of 0.75. The comparison received probabilities of 0.11 for the univariate Levenshtein model, 0.19 for the univariate Jaro-Winkler model, and 0.18 for the bivariate model. The source of the error is related to the large penalty on the scores when there are so few letters in the artist name. To explain, shorter words are more heavily impacted by small differences in characters. Therefore, similar yet shorter words can have lower similarity scores than longer words (refer to page 11 for text string similarity formulas). While this is one drawback, the models still perform well overall. All model results are summarized in Table 10.

Table 10: Logistic Regression Performance

Accuracy = 0.997, Sensitivity = 0.944, Specificity = 1

	True Matches	True Nonmatches
Linked	34	0
Unlinked	2	614

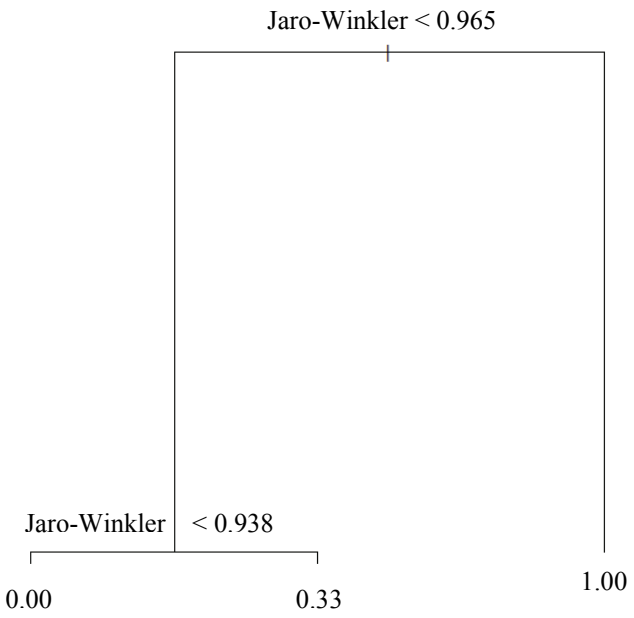
Section 2.3 D: Classification Tree

An alternative supervised record linkage technique is to use a classification tree[19]. A classification tree is a set of decision rules based on a set of chosen variables which partition the feature space using a series of if/else cutoff statements that maximize the separation between labeled matches and nonmatches. The final subgroups are then assigned a probability of their comparisons being a match. For this model, both the Jaro-Winkler and the Levenshtein similarity scores were used as explanatory variables. However, the classification tree, displayed in Figure 3, only used the Jaro-Winkler score to split the data and classify matches. When reading the classification tree each terminal node indicates the probability of comparisons being a match. Beginning at the bottom left of the diagram, comparisons with a Jaro-Winkler score below 0.938 are not predicted to be matches, comparisons with a Jaro-Winkler score greater than 0.938 and less than 0.965 have about 33% probability of being a match and comparisons with a

[19] http://cran.r-project.org/web/packages/tree/tree.pdf

Jaro-Winkler score greater than 0.965 are highly likely to be a match. Subsequently, for this analysis comparisons with a Jaro-Winkler score greater than 0.965 were classified as matches.

Figure 3: Classification Tree – Levenshtein and Jaro-Winkler



A high Jaro-Winkler score cut-off value is not surprising since typographical errors were minimal given the quality of the data from both sources. Mostly, the differences in text strings were quite small such as the extra space in "love rance" compared to "loverance." In the end, the model classified 616 nonmatches and 34 matches in the hand-labeled subset with about 99.7% accuracy, 94.4% sensitivity, and 100% specificity as shown in Table 11. Note, the results mimic that of the logistic models, failing to accurately classify the same two artist names.

Table 11: Classification Tree Performance

Accuracy = 0.997, Sensitivity = 0.944, Specificity = 1

	True Matches	True Nonmatches
Linked	34	0
Unlinked	2	614

Section 2.3E: Final Classification Model

Overall, the performances of the logistic regression model and the classification tree were exactly the same. We chose to use the classification tree to classify the entire comparison dataset due to the simplicity of the model which requires less computation than the logistic regression model. In the end 71 of the 2,933 Next Big Sound Artists were found to be aired on the top charts of radio. However, keep in mind that we are assuming the hand-labeled subset is indeed representative of the entire 2.8 million comparison dataset. Further, because the hand-labeled subset contains a higher proportion of matches compared to the entire dataset, there may be a small bias overstating the probability of comparisons being a match. At the same time, we decisively chose to only classify artist comparisons as matches, if the probability of being match was very high, in order to limit the number of false positives.

Section 3: Exploration of Online Metric Time Series Variables

After identifying which Next Big Sound predicted artists made it to the top charts of radio, we were then interested in discovering patterns among the online metrics of the "radio artists" compared to "nonradio artists". Potentially these patterns can be used to predict which artists will be aired on the top charts of radio in the future. For easy reference, the radio information was added to the Next Big Sound artist predictions as shown in Table 12. Specifically, a radio indicator column which is 0 for never appearing on a top radio chart and 1 for appearing on a top radio chart was added. Next, columns of the month, the day, and the year of the first appearance on a top radio chart were added.

Table 12: Next Big Sound Predictions with Radio Information

ID	Artist	Rank	Date	Radio Indicator	Radio Month	Radio Day	Radio Year
9297	jarren benton	6	12/26/13	0	NA	NA	NA
341198	sam smith	4	4/8/13	1	9	7	2013
336698	trinidad james	1	12/13/12	1	1	5	2013
319292	david tort	7	9/13/12	0	NA	NA	NA

In our initial data collection of the online metrics, the only data available were net daily values, i.e. the daily change in total. For this analysis the metrics of a radio artist, Bastille, and a nonradio artist, Big Boi were examined. Bastille is an English rock band formed in 2010 and currently is signed to the record label, Virgin Records. Big Boi is an American rapper who was previously a member of the well known band, Outkast. Big Boi began to pursue a solo career in 2003 and currently is working under the Def Jam Records label. As an initial step, the metrics were plotted and examined. The summary statistics for each artist are shown in Tables 13 and 14 respectively and Figures 4 and 5 show the plotted net daily online metrics for each artist. In the plots, each vertical line indicates a significant point in time for the artist. The red line indicates the date the artist was predicted by Next Big Sound and the green line indicates the date the artist first appeared on a top radio chart.

Table 13: Summary Statistics of Bastille Net Daily Online Metrics

	Min	Max	Mean	Std. Dev
Facebook Page Likes	7	5,558	1,089	45.9
Wikipedia Page Views	0	101,770	4,156	238.5
Twitter Followers	-1	3,110	395.2	17.0
Youtube Video Views	0	96,920	16,110	562.4
Vevo Video Views	1,277	1,096,000	217,200	7,539.0
SoundCloud Plays	7,727	83,530	21,130	684.8

Table 14: Summary Statistics of Big Boi Net Daily Online Metrics

	Min	Max	Mean	Std. Dev
Facebook Page Likes	-8	1,444	187.4	5.2
Wikipedia Page Views	75	12,820	1,710	26.2
Twitter Followers	-101,600	102,400	670.7	116.3
Youtube Video Views	2,746	33,700	7,335	234.8
Vevo Video Views	125	1,537,000	17,630	1,596.3
SoundCloud Plays	20	236,900	5,907	322.4

Figure 4: Bastille Net Daily Online Metrics

- April 27, 2012: First song release "Overjoyed"
- June 14, 2012: Predicted by Next Big Sound
- June 29, 2012: Vevo Video of "Bad Blood" released
- August 20, 2012: "Bad Blood" digitally released
- February, 2013: "Pompeii" song released
- May 25, 2013: First appearance on Top 200 radio song chart

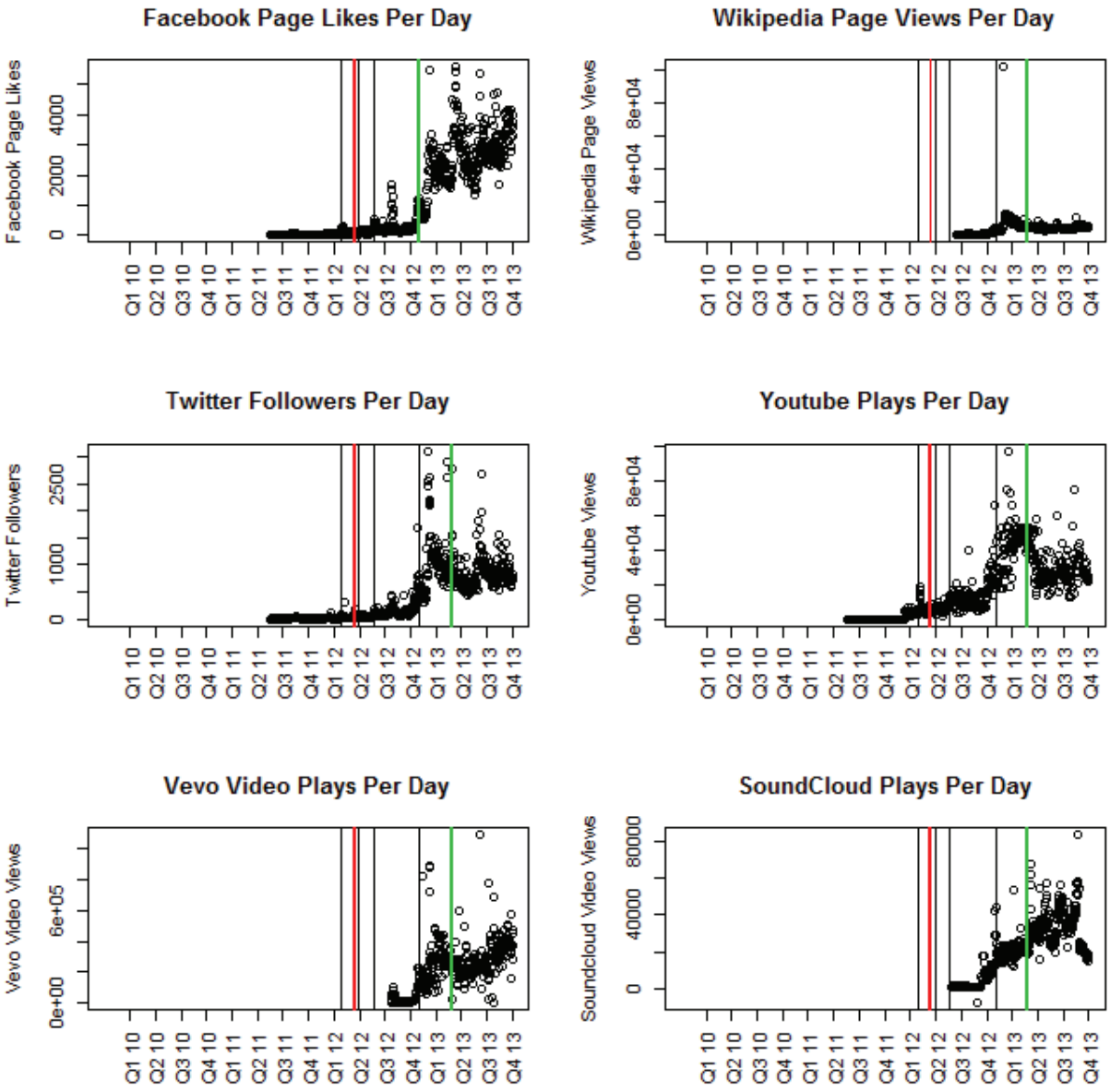
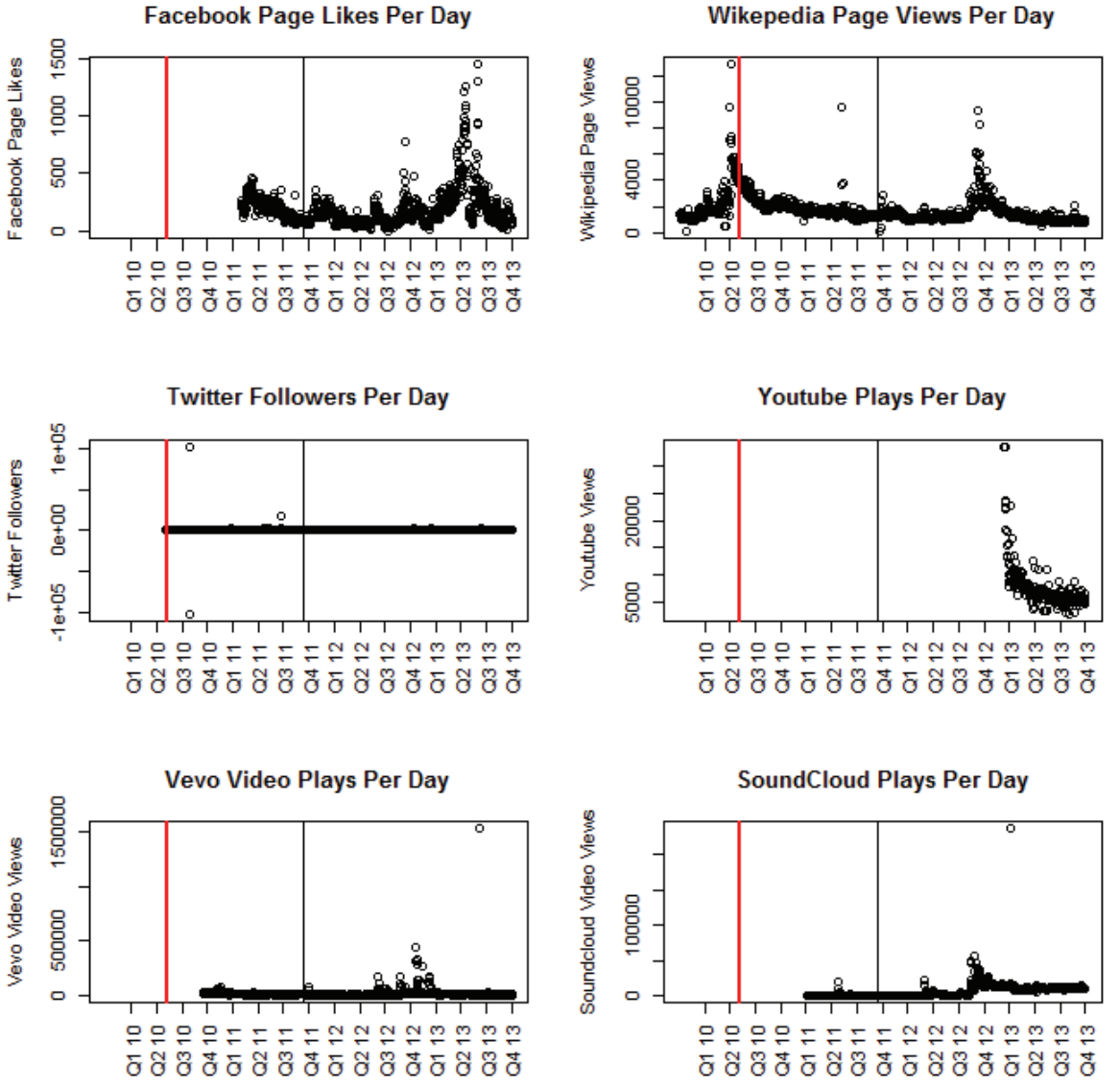


Figure 5: Scatter Plots of Big Boi Net Daily Online Metrics

August 5, 2010: Predicted by Next Big Sound

December 11, 2011: Release of second album "Vicious and Dangerous Rumors"



When examining the plots for Bastille, we see shortly before being aired on the Top 200 songs radio chart, there are large increases in Twitter followers, Youtube views, Vevo plays, and SoundCloud plays and a small increase in Wikipedia page views. Further, after being aired on a top radio chart, there are large increases in Facebook page likes. For Big Boi, while there are

increases in Facebook page likes and Wikipedia page views the increases are not as large as Bastille. Further, Big Boi's Twitter Followers, Vevo plays, and SoundCloud plays are stagnant, and his Youtube plays actually decrease. Overall, the relationship between an artist's activity online and being aired on the top charts of radio is not obvious and is likely complicated.

Examining all the plots we see the time series are left-truncated for both artists. This truncation is particularly noticeable in Big Boi's plot of Youtube plays. There are two possible sources for this truncation; either the online source is not in existence or the source is not yet being tracked by Next Big Sound. However, it is not possible to automatically determine the cause of the truncation. While the truncation may be concerning, another pressing problem is the potential bias of missing data values. The online metric data is not necessarily missing completely at random. In order for an artist's metrics to be tracked by Next Big Sound, a user must link the source. For example, if there is no information about an artist's Facebook page, a user must enter the link of the artist's Facebook page for Next Big Sound to then analyze. Given this process, the missing data values are not completely random. Instead, more popular artists are more likely to have data as users are more interested in their history and future.

A bivariate analysis was also conducted to better understand how each metric relates to the others. Below, Figures 6 and 7 display the pairs plots for our two artists. The top panels display the scatter plots of each variable pair and the bottom panels display the correlation. The red stars indicate the level of statistical significance for the correlation. The correlation was calculated using only the complete pair-wise observations (i.e. pairs with an NA value were removed). For Bastille, we see that Facebook is strongly correlated with all other metrics. Most notably, Facebook and Twitter have a positive correlation of 0.80. Youtube and Twitter are strongly associated with a positive correlation of 0.79. On the flip side, Wikipedia has the weakest association with all other metrics. Interestingly for Big Boi, the behavior of the metrics is very different. In this case, none of the metrics are strongly correlated. If anything, there is a moderate correlation between Youtube and Wikipedia and Youtube and Twitter with positive correlations of 0.59 and 0.57 respectively. It may be the case, that artists who are about to hit the top charts of radio, see an increase in almost every online medium rather than just one or a few.

Figure 6: Relationships Among Bastille Net Daily Online Metrics

statistical significance: *** 1%, ** 5%, *10%, >10%

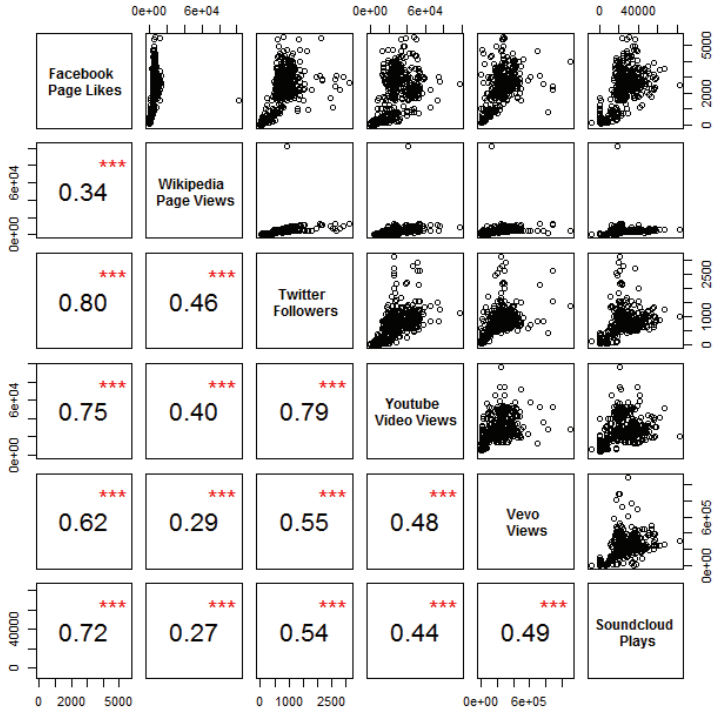
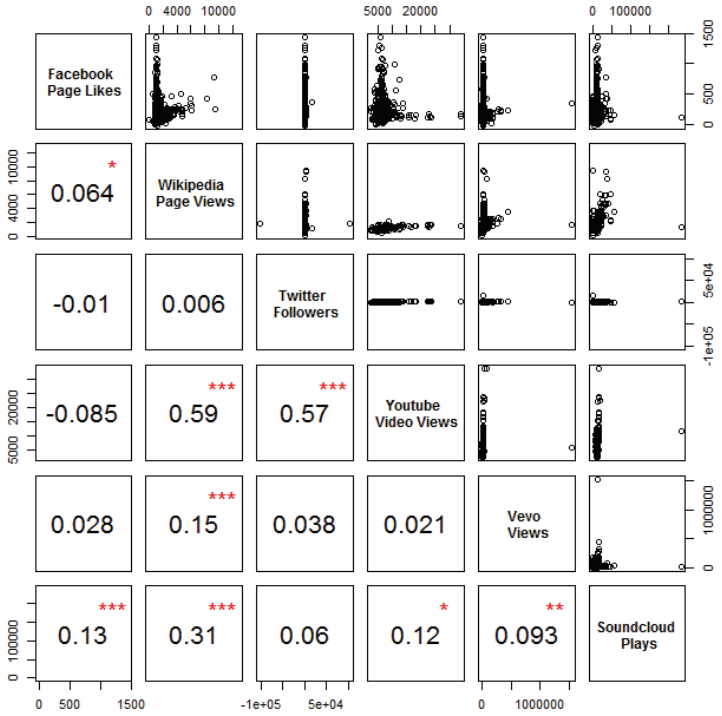


Figure 7: Relationships Among Big Boi Net Daily Online Metrics

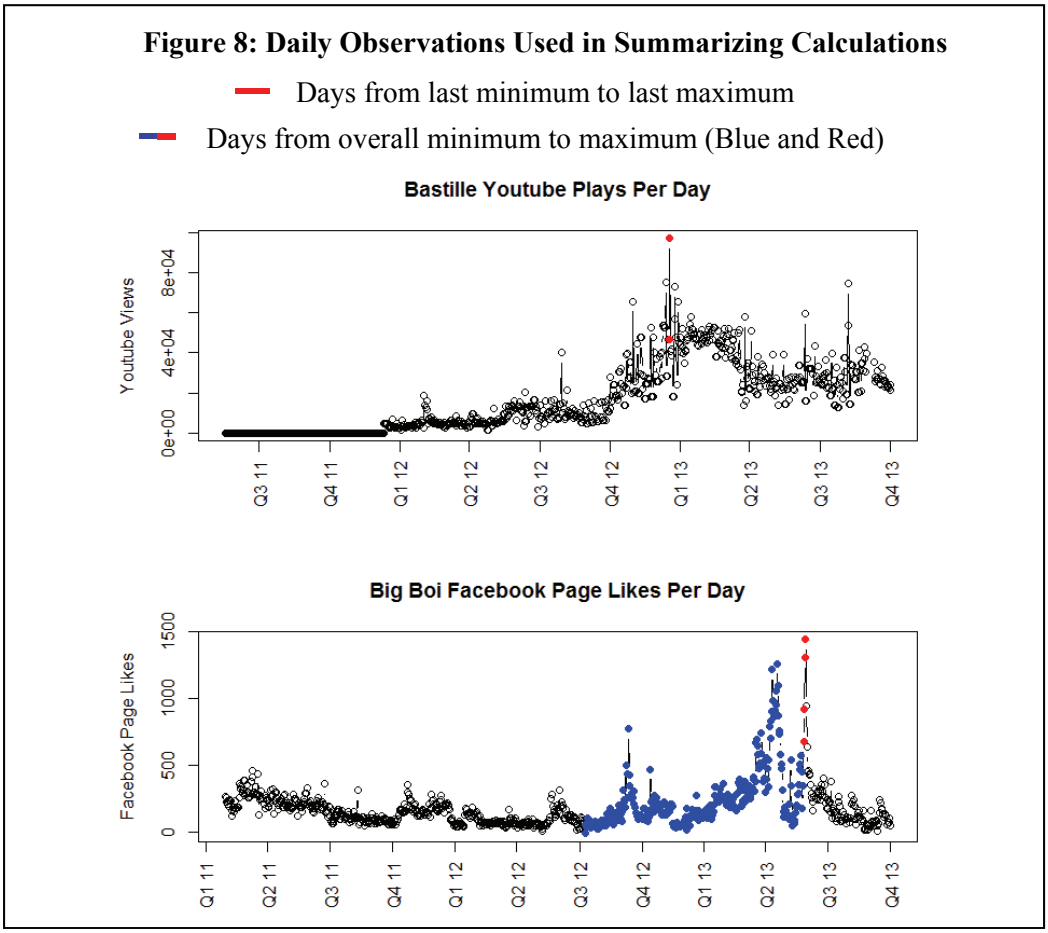
statistical significance: *** 1%, ** 5%, *10%, >10%



Section 4: Logistic Regression – Using Summarizing Online Metric Time Series

As a first attempt at modeling the data, logistic regression models were fit with a binary outcome, of 0 for no radio chart appearance and 1 for radio chart appearance. Using each daily value of an online metric would result in a model where the number of predictors would greatly exceed that number of observations and be impossible to fit. Moreover, that strategy ignores the time sequence associated with each metric. Therefore, several "summary" variables were created in an effort to extract interesting and possibly influential aspects of the time series. Each summary variable is defined below and calculated for all pre-radio chart, time periods. That is, for radio artists, the days prior to an artist appearing on a top radio chart were used in the calculations. For artists who have not appeared on those charts, all of their days in the time period examined were used. Refer to the respective artist plots in Figure 8 for all examples listed in the definitions below.

Figure 8: Daily Observations Used in Summarizing Calculations



Average: Total sum of the net daily values divided by a unit of time

A) Scaled by days

B) Scaled by weeks

C) Scaled by months

Ex: On average Bastille gained 238 Twitter followers per day prior to appearing on the top charts of radio (All data points).

Maximum Increase:

A) Maximum net daily value

Ex: Big Boi reached a maximum gain of 1,444 in Facebook page likes on Aug. 28th, 2013 (Last red data point).

Aggregate Before Peak:

A) Sum of net daily values between the last minimum and maximum

Ex: Big Boi's has a total increase in Facebook page likes of 4,880 from the last minimum on Aug. 23rd, 2013 to the last maximum on Aug. 28th, 2013 9 (Red data points).

B) Sum of net daily values from beginning to the maximum

Ex: Big Boi had 34,335 Facebook page likes from April 29th, 2011 to August 28th, 2013 (All data points prior to last red data point).

C) Sum of net daily values from overall minimum prior to the maximum

Ex: Big Boi had 8,656 Facebook page likes from the overall minimum on October 8th, 2013 to the maximum on August 28th, 2013 (Blue and red data points).

Peak Slope:

A) Slope of the values between the last minimum and last maximum

Ex: Bastille's slope in Youtube views of the last minimum on March 17th, 2013 to the last maximum on March 18th, 2013 is 16,685 (Red data points).

Percentage Increase Over Time:

A) Percentage increase of the values between the last minimum and last maximum, divided by number of days. Percentage changes involving negative and positive values were calculated by adding the absolute value of the minimum plus one to the maximum and minimum values

Ex: Bastille's percentage increase per day was 79.7% from Sept 2nd, 2013 to Sept 4th, 2013 (Red data points).

Rank:

A) Rank of Next Big Sound prediction where 1 is the highest rank possible

Ex: Referring back to Table 12, David Tort has a rank of 7

After calculating all of the collapsed data points for each artist it was found that the variables did not capture the information as intended. The aggregate before peak, slope, and percentage variables picked up only very small ranges of data to collapse (e.g. 2 or 3 days). This happened frequently because the data are daily values with high volatility. Therefore, the indentified minimum and maximum used in the calculations may just be small "blips" in the data rather than true changes in the overall trend. Further, quite often there were large amounts of missing data values which prevented the variables from being calculated. Accordingly, only artists with complete observations with a minimum of fifty days were examined. A higher minimum number of days removed too many radio artists from the analysis. In the future, it may be a good idea to convert the data to weekly values for increased stability. However, because the values are summarized it does not make a difference in this modeling technique. In future modeling, we can impute the missing values which will be discussed further in Section 5.

The summary measures are likely related (over time) since they are trying to capture how an artist is gaining online attention. The correlations between each variable pair were calculated (shown in Table 15). The lower panels show the correlation and the upper panels display the statistical significance with more "*" symbols indicating stronger significance. Overall, there is not a large concern of multicollinearity among the variables. The highly correlated variables are bolded. For example, average day and average week are exactly collinear with a correlation of 1.

Table 15: Correlation Matrix of Net Daily Summary Variables

statistical significance: *** 1%, ** 5%, *10%, >10%

	Avg. Day	Avg. Week	Avg. Month	Max Increase	Agg. Peak A	Agg. Peak B	Agg. Peak C	Slope	Percentage	Rank
Avg. Day	1.00	***	***	***	***	***	***	***	***	
Avg. Week	**1.00**	1.00	***	***	***	***	***	***	***	
Avg. Month	**1.00**	**1.00**	1.00	***	***	***	***	***	***	
Max Increase	0.40	0.40	0.40	1.00	***	***	***	***	***	
Agg. Peak A	0.50	0.50	0.50	**0.92**	1.00	***	***	***	***	
Agg. Peak B	**0.81**	**0.81**	**0.81**	0.44	0.52	1.00	***	***	***	
Agg. Peak C	**0.79**	**0.79**	**0.79**	0.50	0.60	**0.97**	1.00	***		
Slope	0.26	0.26	0.26	**0.88**	0.63	0.30	0.32	1.00	***	
Percentage	0..35	0..35	0..35	0.07	0.08	0.11	0.01	0.06	1.00	
Rank	-0.009	-0.009	-0.009	-0.03	-0.04	-0.03	-0.03	-0.02	0.02	1.00

Moving forward, the logistic regression models were fit using the summarizing variables calculated over the net daily values. Note that we also tried building models using summaries of the absolute daily values. However, doing so failed to return statistically significant and predictive results (See Appendix A). The univariate and multivariate models for each online avenue are shown in Tables 16-21. Models with statistically significant relationships at the 10% level are bolded. Reported are the number of radio artists and total artists used in the analysis, the variable, the univariate coefficient, the univariate p-value, the probability threshold which maximized accuracy, the accuracy, sensitivity, and specificity as defined in Section 2.3 C. The last two columns show the coefficients and p-values of the multivariate model. The last row shows the number of radio and nonradio artists, threshold, accuracy, sensitivity, and specificity for the multivariate model. The possible thresholds evaluated ranged from of 0.50 to 0.90 by increments of 0.05. We also defined the accuracy variance by recording the accuracy at each threshold level from 0.50 to 0.90 and then calculated the variance of the accuracies. Lower variance implies higher stability. The accuracy variance was <0.0001 for all models (not shown in the tables).

Table 16: Facebook Logistic Regression Models

Statistically significant univariate models and multivariate variables at the 10% level are bolded

Radio/Total # of Artists	Variable	Univariate Coefficient	Univariate P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Adjusted Coefficient	Adjusted P-Value
66/2,857	Avg. Day	$-6.3*10^{-8}$	0.94	0.50	0.98	0	1	**$5.03*10^{-4}$**	**0.071**
66/2,857	Avg. Week	$-9.0*10^{-7}$	0.94	0.50	0.98	0	1	NA	NA
66/2,857	Avg. Month	$-2.1*10^{-7}$	0.94	0.50	0.98	0	1	NA	NA
66/2,857	**Max Inc.**	**$-3.3*10^{-5}$**	**0.087**	**0.50**	0.98	**0**	**1**	$5.7*10^{-5}$	0.64
66/2,683	Agg Peak A	$-4.4*10^{-6}$	0.11	0.50	0.98	0	1	$-5.7*10^{-5}$	0.21
66/2,857	Agg Peak B	-0.41	0.99	0.50	0.98	0	1	$7.5*10^{-7}$	0.62
66/2,857	Agg Peak C	$6.1*10^{-9}$	0.96	0.50	0.98	0	1	$-8.5*10^{-7}$	0.60
66/2,683	Slope	-0.00035	0.081	0.50	0.98	0	1	$-4.7*10^{-4}$	0.38
66/2,683	Percentage	-0.0037	0.29	0.50	0.98	0	1	$-1.6*10^{-3}$	0.59
66/2,683	Rank	-0.0086	0.68	0.50	0.98	0	1	$-1.7*10^{-2}$	0.41
66/2,683	Multivariate	-	-	0.50	0.98	0	1	-	-

Table 17: Wikipedia Logistic Regression Models

Statistically significant univariate models and multivariate variables at the 10% level are bolded

Radio/Total # of Artists	Variable	Univariate Coefficient	Univariate P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Adjusted Coefficient	Adjusted P-Value
63/1,482	Avg. Day	$-2.1*10^{-5}$	0.75	0.50	0.96	0	1	$-7.7*10^{-5}$	0.75
63/1,482	Avg. Week	$-3.0*10^{-6}$	0.75	0.50	0.96	0	1	NA	NA
63/1,482	Avg. Month	$-6.9*10^{-7}$	0.75	0.50	0.96	0	1	NA	NA
63/1,482	Max Inc.	$-1.0*10^{-5}$	0.30	0.50	0.96	0	1	$2.7*10^{-5}$	0.65
63/1,478	Agg Peak A	$-4.5*10^{-6}$	0.34	0.50	0.96	0	1	$-1.2*10^{-5}$	0.53
63/1,482	Agg Peak B	$4.1*10^{-8}$	0.66	0.50	0.96	0	1	$1.8*10^{-7}$	0.61
63/1,478	Agg Peak C	$-4.5*10^{-8}$	0.66	0.50	0.96	0	1	$2.7*10^{-7}$	0.47
63/1,478	Slope	$-9.0*10^{-5}$	0.25	0.50	0.96	0	1	$-2.8*10^{-4}$	0.36
63/1,482	Percentage	-0.0041	0.69	0.50	0.96	0	1	$1.4*10^{-3}$	0.84
63/1,482	**Rank**	**0.046**	**0.026**	**0.50**	**0.96**	**0**	**1**	**$4.5*10^{-2}$**	**0.029**
63/1,478	Multivariate	-	-	0.50	0.96	0	1	-	-

Table 18: Twitter Logistic Regression Models

No univariate models or multivariate variables were statistically significant at the 10% level

Radio/Total # of Artists	Variable	Univariate Coefficient	Univariate P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Adjusted Coefficient	Adjusted P-Value
66/2,555	Avg. Day	$1.5*10^{-5}$	0.92	0.50	0.97	0	1	$-4.1*10^{-4}$	0.28
66/2,555	Avg. Week	$2.1*10^{-6}$	0.92	0.50	0.97	0	1	NA	NA
66/2,555	Avg. Month	$5.0*10^{-7}$	0.92	0.50	0.97	0	1	NA	NA
66/2,555	Max Inc.	$-8.7*10^{-6}$	0.44	0.50	0.97	0	1	$-3.1*10^{-6}$	0.96
66/2,400	Agg Peak A	$-3.1*10^{-7}$	0.96	0.50	0.97	0	1	$2.0*10^{-5}$	0.29
66/2,555	Agg Peak B	$1.8*10^{-7}$	0.46	0.50	0.97	0	1	$3.3*10^{-7}$	0.71
66/2,555	Agg Peak C	$2.5*10^{-7}$	0.36	0.50	0.97	0	1	$4.1*10^{-7}$	0.65
66/2,400	Slope	$-5.9*10^{-5}$	0.40	0.50	0.97	0	1	$-1.9*10^{-4}$	0.48
66/2,400	Percentage	$-3.7*10^{-5}$	0.86	0.50	0.97	0	1	$1.0*10^{-4}$	0.63
66/2,555	Rank	-0.016	0.45	0.50	0.97	0	1	-0.011	0.63
64/2,400	Multivariate	-	-	0.50	0.97	0	1	-	-

Table 19: Youtube Logistic Regression Models

No univariate models or multivariate variables were statistically significant at the 10% level

Radio/Total # of Artists	Variable	Univariate Coefficient	Univariate P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Adjusted Coefficient	Adjusted P-Value
61/2,222	Avg. Day	$-3.4*10^{-6}$	0.12	0.50	0.97	0	1	$-7.3*10^{-6}$	0.36
61/2,222	Avg. Week	$-4.8*10^{-7}$	0.12	0.50	0.97	0	1	NA	NA
61/2,222	Avg. Month	$-1.1*10^{-7}$	0.12	0.50	0.97	0	1	NA	NA
61/2,222	Max Inc.	$-7.8*10^{-9}$	0.64	0.50	0.97	0	1	$-1.3*10^{-6}$	0.34
60/1,1974	Agg Peak A	$-3.7*10^{-7}$	0.24	0.50	0.97	0	1	$5.7*10^{-7}$	0.40
61/2,222	Agg Peak B	$-3.6*10^{-9}$	0.30	0.50	0.97	0	1	$-1.4*10^{-7}$	0.31
61/2,222	Agg Peak C	$-2.5*10^{-9}$	0.56	0.50	0.97	0	1	$1.5*10^{-7}$	0.27
60/1,1974	Slope	$-3.8*10^{-9}$	0.85	0.50	0.97	0	1	$2.2*10^{-6}$	0.29
60/1,1974	Percentage	-0.0026	0.26	0.50	0.97	0	1	$-2.2*10^{-3}$	0.30
61/2,222	Rank	-0.0010	0.64	0.50	0.97	0	1	-0.013	0.57
60/1,974	Multivariate	-	-	0.75	0.97	0	1	-	-

Table 20: Vevo Logistic Regression Models

Statistically significant univariate models and multivariate variables at the 10% level are bolded

Radio/Total # of Artists	Variable	Univariate Coefficient	Univariate P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Adjusted Coefficient	Adjusted P-Value
49/746	Avg. Day	$9.2*10^{-7}$	0.28	0.50	0.94	0	1	**$5.2*10^{-6}$**	**0.04**
49/746	Avg. Week	$1.3*10^{-7}$	0.28	0.50	0.94	0	1	NA	NA
49/746	Avg. Month	$3.1*10^{-8}$	0.28	0.50	0.94	0	1	NA	NA
49/746	Max Inc.	$-4.9*10^{-11}$	0.996	0.50	0.94	0	1	**$1.6*10^{-6}$**	**0.066**
49/725	Agg Peak A	$-3.7*10^{-10}$	0.97	0.50	0.94	0	1	**$-1.6*10^{-6}$**	**0.061**
49/746	Agg Peak B	$1.9*10^{-9}$	0.42	0.50	0.94	0	1	$-1.3*10^{-9}$	0.88
49/746	Agg Peak C	$1.7*10^{-9}$	0.57	0.50	0.94	0	1	$-3.0*10^{-9}$	0.76
49/725	Slope	$-1.3*10^{-10}$	0.997	0.50	0.94	0	1	$5.6*10^{-8}$	0.85
49/725	**Percentage**	**$1.6*10^{-6}$**	**0.091**	**0.90**	**0.94**	**0**	**1**	$9.3*10^{-7}$	0.36
49/746	Rank	-0.031	0.18	0.50	0.94	0	1	$-3.4*10^{-2}$	0.15
49/725	Multivariate	-	-	0.90	0.94	0	1	-	-

Table 21: SoundCloud Logistic Regression Models

Statistically significant univariate models and multivariate variables at the 10% level are bolded

Radio/Total # of Artists	Variable	Univariate Coefficient	Univariate P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Adjusted Coefficient	Adjusted P-Value
44/1,804	Avg. Day	$-1.4*10^{-6}$	0.90	0.50	0.98	0	1	$6.8*10^{-6}$	0.80
44/1,804	Avg. Week	$2.0*10^{-7}$	0.90	0.50	0.98	0	1	NA	NA
44/1,804	Avg. Month	$-4.7*10^{-8}$	0.90	0.50	0.98	0	1	NA	NA
44/1,804	Max Inc.	$-6.2*10^{-7}$	0.11	0.50	0.98	0	1	**$-1.0*10^{-5}$**	**0.013**
44/1,740	Agg Peak A	$-4.7*10^{-7}$	0.22	0.50	0.98	0	1	**$6.6*10^{-6}$**	**0.023**
44/1,804	Agg Peak B	$-1.7*10^{-8}$	0.61	0.50	0.98	0	1	$4.7*10^{-8}$	0.91
44/1,804	Agg Peak C	$-1.1*10^{-8}$	0.77	0.50	0.98	0	1	$-1.3*10^{-8}$	0.97
44/1,740	Slope	$-1.6*10^{-6}$	0.23	0.50	0.98	0	1	$10.0*10^{-6}$	0.11
44/1,740	Percentage	-0.00094	0.24	0.50	0.98	0	1	$-8.5*10^{-4}$	0.26
44/1,804	Rank	0.0056	0.83	0.50	0.98	0	1	-0.01	0.72
44/1,740	Multivariate	-	-	0.50	0.98	0	1	-	-

Overall, the models did not perform very well. While the accuracy is high for each model, the sensitivity is 0 for every model meaning no radio artists were ever accurately classified. This is most likely due to the high proportion of nonradio to radio artists. There may not be enough information to clearly identify special behavior among the radio artists.

We then explored multivariate models using all online avenues. First, the model was fit using all summary measures. Interestingly, the majority of the statistically significant variables are from Facebook. However, the number of artists dropped to 20 radio artists and 266 nonradio artists, due to incomplete artist observations which do not have data for every online avenue. Again the threshold value was optimized and was found to be 0.50. Overall, this performed only slightly better than the individual online avenue models with an accuracy of 0.92, sensitivity of 0.17, and specificity of 0.96. The results are shown in Table 22. In another attempt to model the data, the univariate statistically significant variables from Tables 16-21 were used as explanatory variables. However, this model did not perform well with 0 sensitivity. Additionally, an attempt was made to use only the statistically significant variables from the complete model in Table 22. However, the results were again poor with 0 sensitivity.

Table 22: Complete Multivariate Logistic Regression Model

20 radio artists, 266 nonradio artists

Threshold = 0.50, Accuracy = 0.92, Sensitivity = 0.17, Specificity = 0.96

Statistically significant variables at the 10% level are bolded

Online Avenue	Variable	Coefficient	P-Value
-	Intercept	-1.53	0.067
Facebook	Avg. Day	$-1.4*10^{-3}$	0.57
Wikipedia	Avg. Day	$-1.9*10^{-3}$	0.42
Twitter	**Avg. Day**	$\mathbf{-1.7*10^{-2}}$	**0.086**
Youtube	Avg. Day	$2.8*10^{-5}$	0.36
Vevo	Avg. Day	$3.1*10^{-5}$	0.46
SoundCloud	**Avg. Day**	$\mathbf{2.0*10^{-4}}$	**0.012**
Facebook	**Max Increase**	$\mathbf{2.7*10^{-3}}$	**0.057**
Wikipedia	Max Increase	$6.7*10^{-4}$	0.19
Twitter	Max Increase	2.510^{-4}	0.87

Youtube	Max Increase	$6.4*10^{-7}$	0.89
Vevo	Max Increase	$1.0*10^{-6}$	0.94
SoundCloud	Max Increase	$-1.5*10^{-6}$	0.91
Facebook	**Aggregate Before Peak A**	**$-1.7*10^{-3}$**	**0.045**
Wikipedia	Aggregate Before Peak A	$-1.7*10^{-4}$	0.28
Twitter	Aggregate Before Peak A	6.010^{-4}	0.44
Youtube	Aggregate Before Peak A	$4.0*10^{-7}$	0.89
Vevo	Aggregate Before Peak A	$6.3*10^{-6}$	0.47
SoundCloud	Aggregate Before Peak A	$3.8*10^{-6}$	0.76
Facebook	Aggregate Before Peak B	$9.0*10^{-6}$	0.63
Wikipedia	Aggregate Before Peak B	$2.8*10^{-6}$	0.32
Twitter	Aggregate Before Peak B	$-9.8*10^{-5}$	0.47
Youtube	Aggregate Before Peak B	$-5.3*10^{-8}$	0.89
Vevo	Aggregate Before Peak B	$2.8*10^{-7}$	0.45
SoundCloud	Aggregate Before Peak B	$1.6*10^{-6}$	0.27
Facebook	Aggregate Before Peak C	$4.6*10^{-5}$	0.16
Wikipedia	Aggregate Before Peak C	$-4.8*10^{-6}$	017
Twitter	Aggregate Before Peak C	$1.4*10^{-4}$	0.34
Youtube	Aggregate Before Peak C	$2.4*10^{-7}$	0.56
Vevo	Aggregate Before Peak C	$-2.7*10^{-7}$	0.56
SoundCloud	**Aggregate Before Peak C**	**$-3.2*10^{-6}$**	**0.060**
Facebook	**Peak Slope**	**$-6.8*10^{-3}$**	**0.060**
Wikipedia	Peak Slope	$-2.1*10^{-3}$	0.24
Twitter	Peak Slope	$-1.9*10^{-3}$	0.70
Youtube	Peak Slope	$6.4*10^{-7}$	0.96
Vevo	Peak Slope	$-4.5*10^{-5}$	0.32
SoundCloud	Peak Slope	$-1.0*10^{-5}$	0.40
Facebook	**Percentage Change Over Time**	**0.017**	**0.083**
Wikipedia	Percentage Change Over Time	$-1.0*10^{5}$	0.12
Twitter	Percentage Change Over Time	-0.074	0.15
Youtube	Percentage Change Over Time	-0.012	0.32
Vevo	Percentage Change Over Time	$-6.8*10^{-5}$	0.34
SoundCloud	Percentage Change Over Time	$-9.7*10^{-6}$	0.84
-	Rank	-0.049	0.45

In summation, logistic regression did not perform well for this analysis. In particular, the models failed to classify any radio artists. This may be the case because the summary variables do not capture the information as intended. Or perhaps, more information related to the time aspect is also necessary for better results. As such, other models more suited for longitudinal

data must be used for this analysis but first, we will focus on imputing the missing data before continuing with modeling.

Section 5: Imputation of Missing Data

As discussed in Section 4, in addition to the left-truncation of our time series (source unknown), there tends to be a few days missing here and there most likely due to, for example, system malfunctions. Our second attempt at Next Big Sound metric data collections was more computationally automated and gave us the daily cumulative values for our time period rather than the net daily values in the first round of collection. As such, we can collapse the data to weekly values resulting in fewer missing values and increased stability. With the cumulative values, we only need to know what the value is on the 7th day of the week rather than having to sum all 7 days of the net daily values. The weekly conversion was implemented beginning the week of December 27th, 2009 to January 2nd, 2010 and ended the week of December 22nd, 2013 to December 28th 2013. While the weekly conversion helped with stability there was still a large number of missing weeks. Specifically, of the 444,254 weekly observations only about 3.67% were entirely complete (i.e., there were no NA values for any of the online metrics). The completeness of each metric was examined individually, and the percentages of complete weekly observations are shown in Table 23 below. Notice that Vevo and SoundCloud have the fewest complete observations.

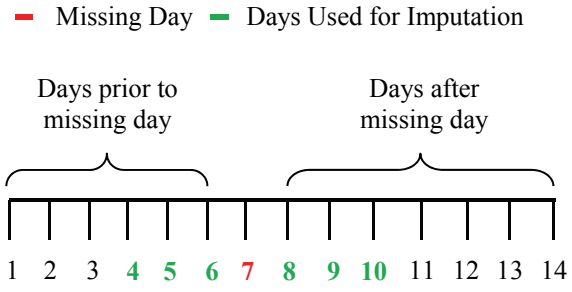
Table 23: Percentage of Complete Weekly Observations

Facebook Page Likes	76.7%
Wikipedia Views	58.2%
Twitter Followers	67.0%
Youtube Views	44.6%
Vevo Views	19.8%
SoundCloud Plays	32.4%
All	3.67%

To mitigate the missingness, one option was to use only the complete observations for the analysis. However, this greatly reduced the sample size to 16,304 daily observations belonging to 43 radio artists and 348 nonradio artists. Instead we chose to impute, or replace, some of the missing weekly values. Implicitly, we made the assumption that the missing values are missing completely at random. We assumed for example, that the system malfunctions occur randomly. However, the left-truncated data is not missing at random as previously discussed in Section 3, and was therefore, not imputed.

The first imputation approach tested was a simple system of averaging. If the day corresponding to the end of the week is missing, the closest day prior to and after the missing day are averaged. The days to be averaged must be within 3 days of the missing day. Any further, and the days begins to bleed into the prior or next week. The choice of days averaged are dependent on which days are available. Therefore, there are 9 possibilities of day combinations prior to and after the missing day in the order of (1 day prior , 1 day after), (1 day prior , 2 days after),….,(3 days prior , 3 days after). Ideally we want the averaging window to be as small as possible (1 day prior , 1 day after). For example in Figure 9, the best case scenario would be to average days 6 and 8 to replace day 7. In the case when there are only values prior to the missing day, the previous day is used to replace the missing day. If the previous day is also missing, the week receives an NA value. An analogous rule is used if there are only complete days after the missing day.

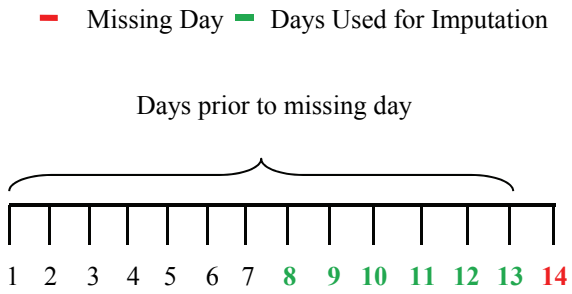
Figure 9: Average Imputation Diagram



The second imputation approach estimates missing values by applying a prior average growth rate. If the day corresponding to the end of the week is missing, the average percentage

change from day to day of the prior six days or fewer is applied to the last complete observation. For example, in Figure 10 below, if day 14 is missing the growth rates between each consecutive day in the order of (day 8 , day 9), (day 9 , day 10),…,(day 12 , day 13) are averaged and applied to day 13. Again, the days must be consecutive for the growth rate to be included. For example, if days 8, 9, 11, and 12 are the only days available then the growth rate between days 8 and 9 and the growth rate between days 11 and 12 are averaged. We do not average days 9 and 11. Ideally we want to average over all six days when possible.

Figure 10: Average Growth Imputation Diagram



To determine the appropriate imputation method, we estimated the imputation error by applying the methods to 500 randomly sampled complete weeks with known values. The percentage error was calculated by comparing the imputation estimate to the actual value. The resulting distributions of imputation errors for each online metric are shown in Figures 10-15.

The first method of averaging the absolute values generally outperformed the second method of using average growth rates. However, the second method proved more accurate and stable for Youtube. Therefore, method 1 was applied to the Facebook, Wikipedia, Twitter, Vevo, and SoundCloud variables while method 2 was applied to Youtube plays. See Appendix B for the results of the reciprocal imputation methods for each online metric.

Figure 10: Facebook Estimated Imputation Error Method 1

*The before and after in the title indicate the number of days before and after the missing value

Total Range of Percentage Errors: -2.6% to 6.1%

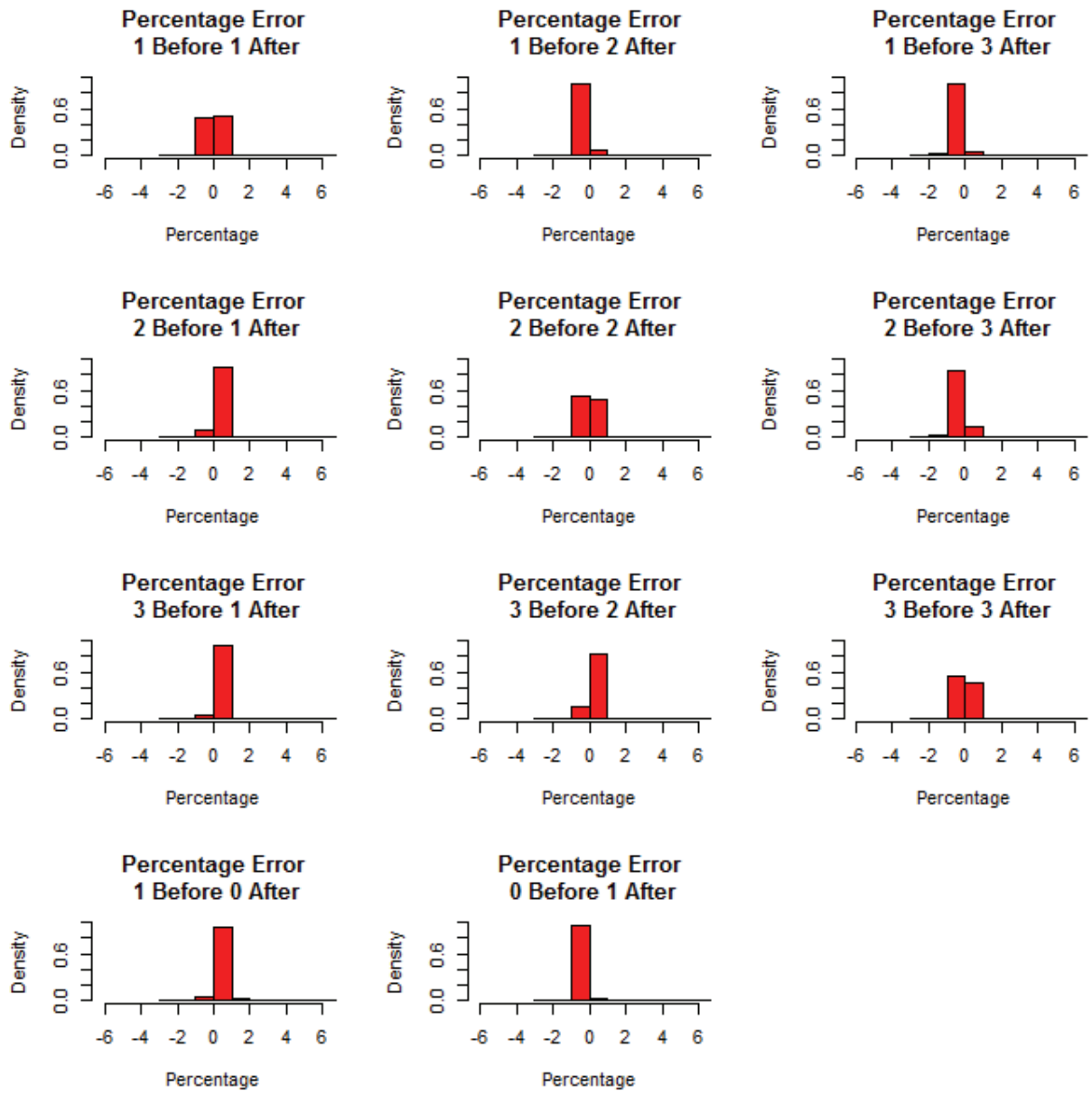


Figure 11: Wikipedia Estimated Imputation Error Method 1

Total Range of Percentage Errors: -5.68% to 7.80%

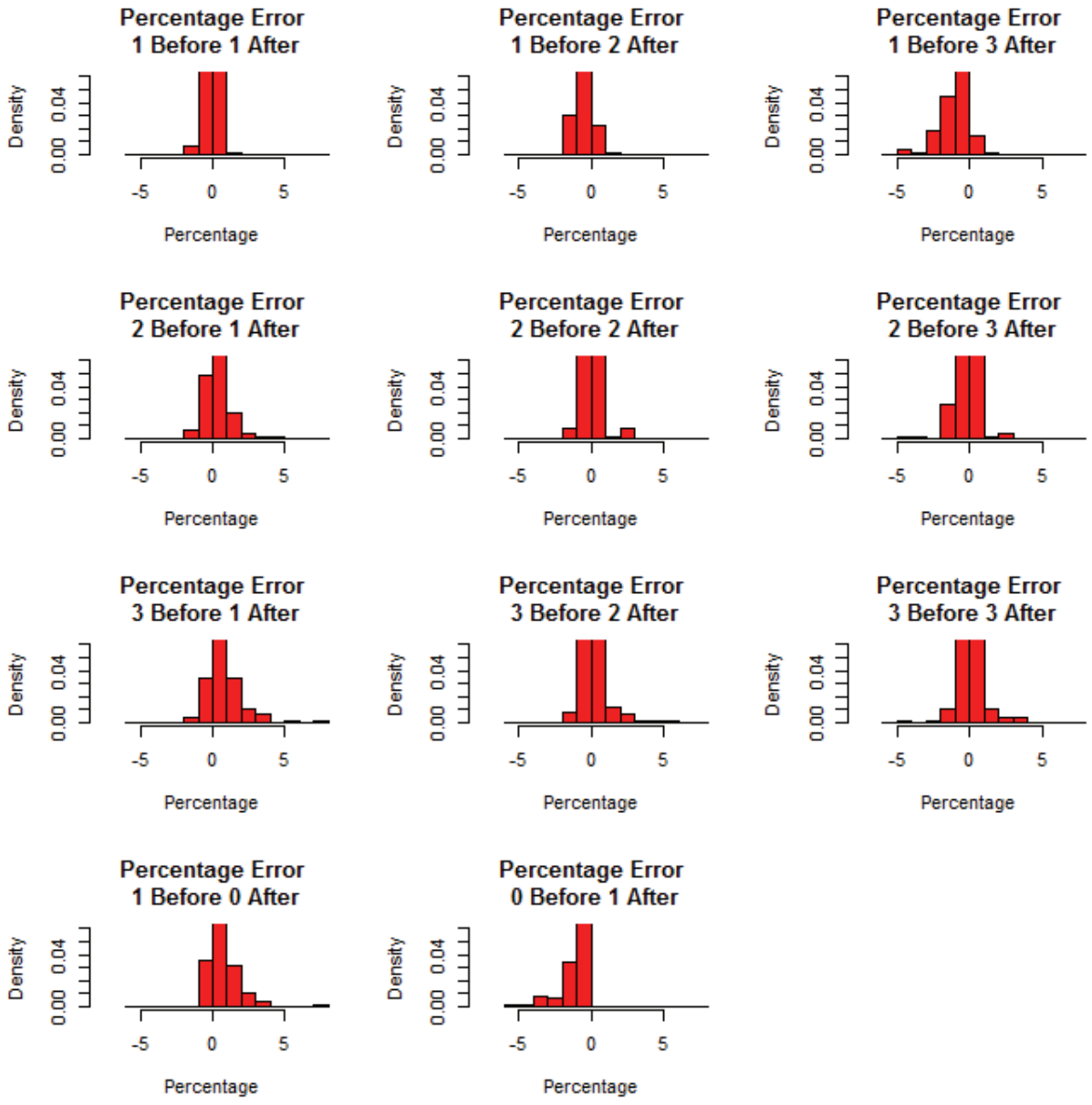


Figure 12: Twitter Estimated Imputation Error Method 1

*The before and after in the title indicate the number of days before and after the missing value

Total Range of Percentage Errors: -12.5% to 20.0%

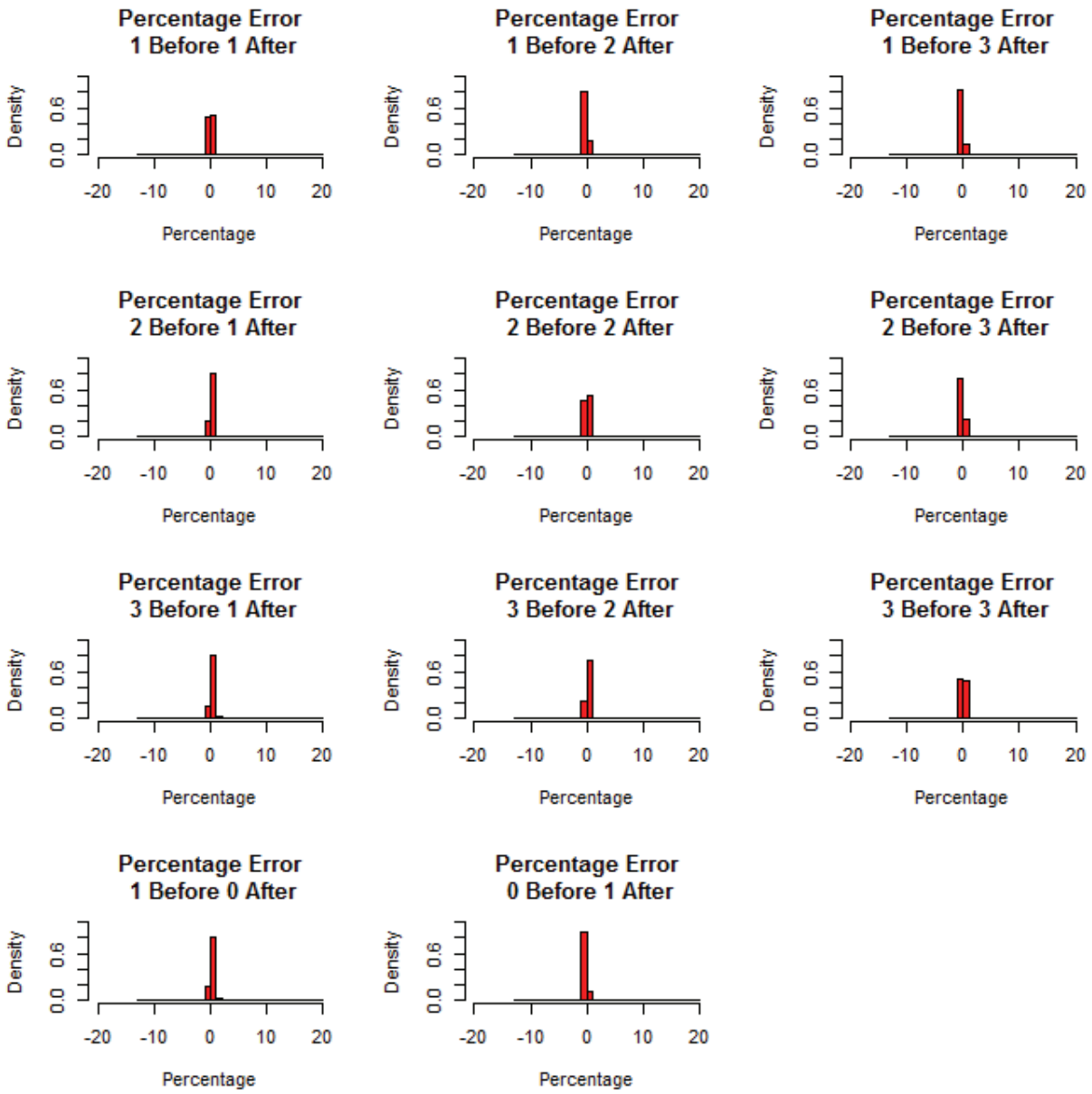


Figure 13: Youtube Estimated Imputation Error Method 2

Total Range of Percentage Errors: -7.5% to 16.9%

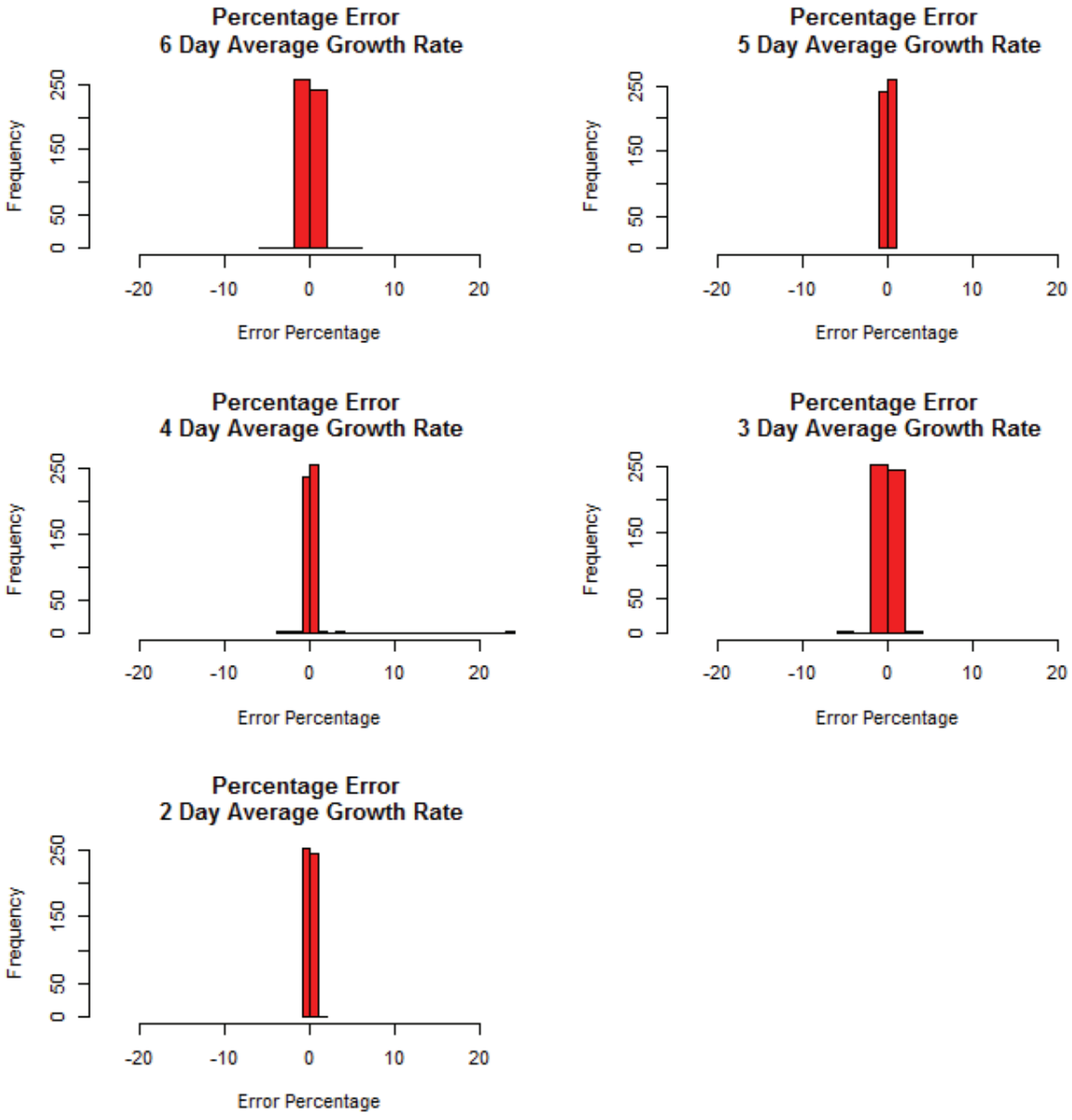


Figure 14: Vevo Estimated Imputation Error Method 1

*The before and after in the title indicate the number of days before and after the missing value

Total Range of Percentage Errors: -38.2% to 41.5%

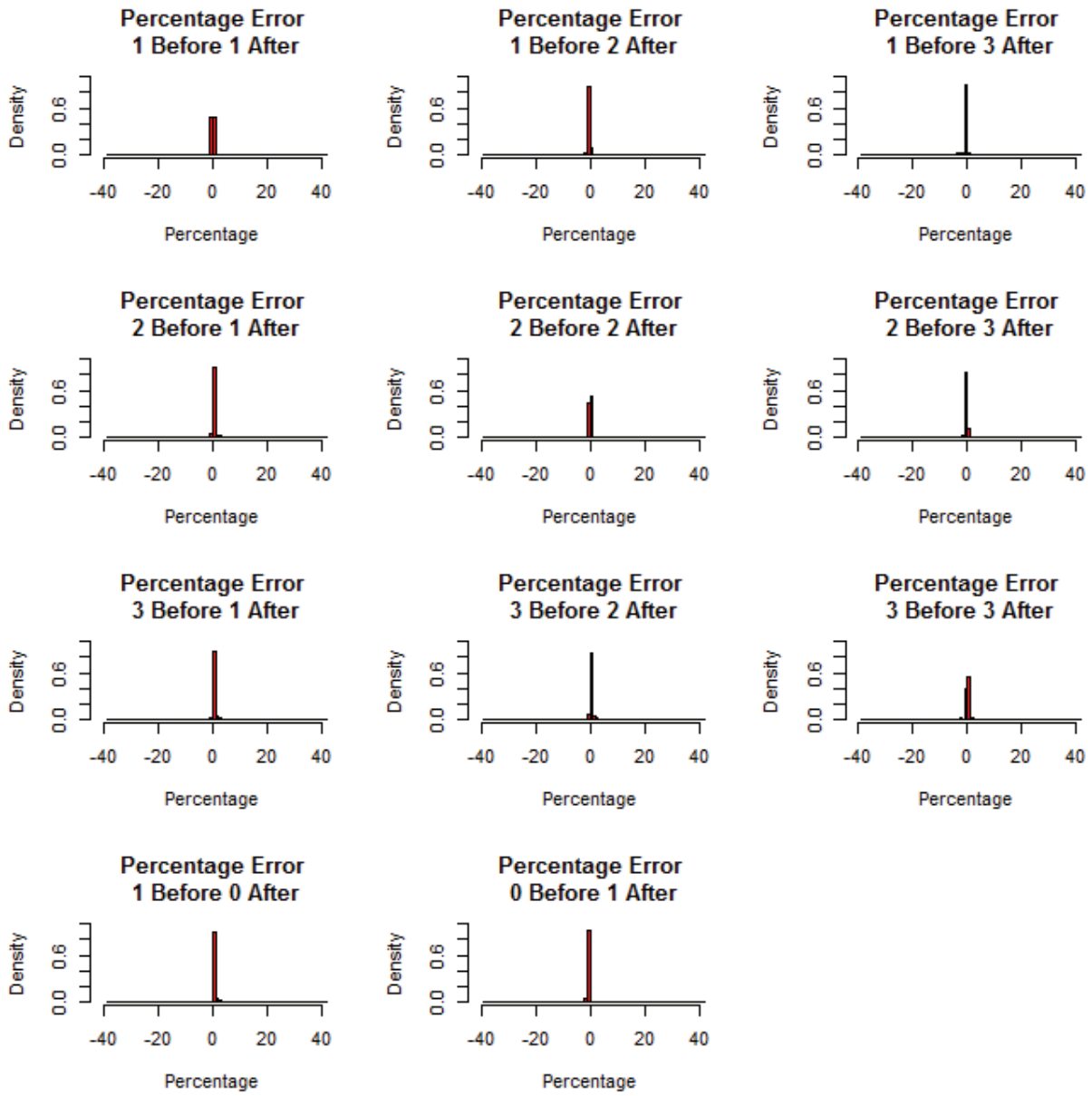
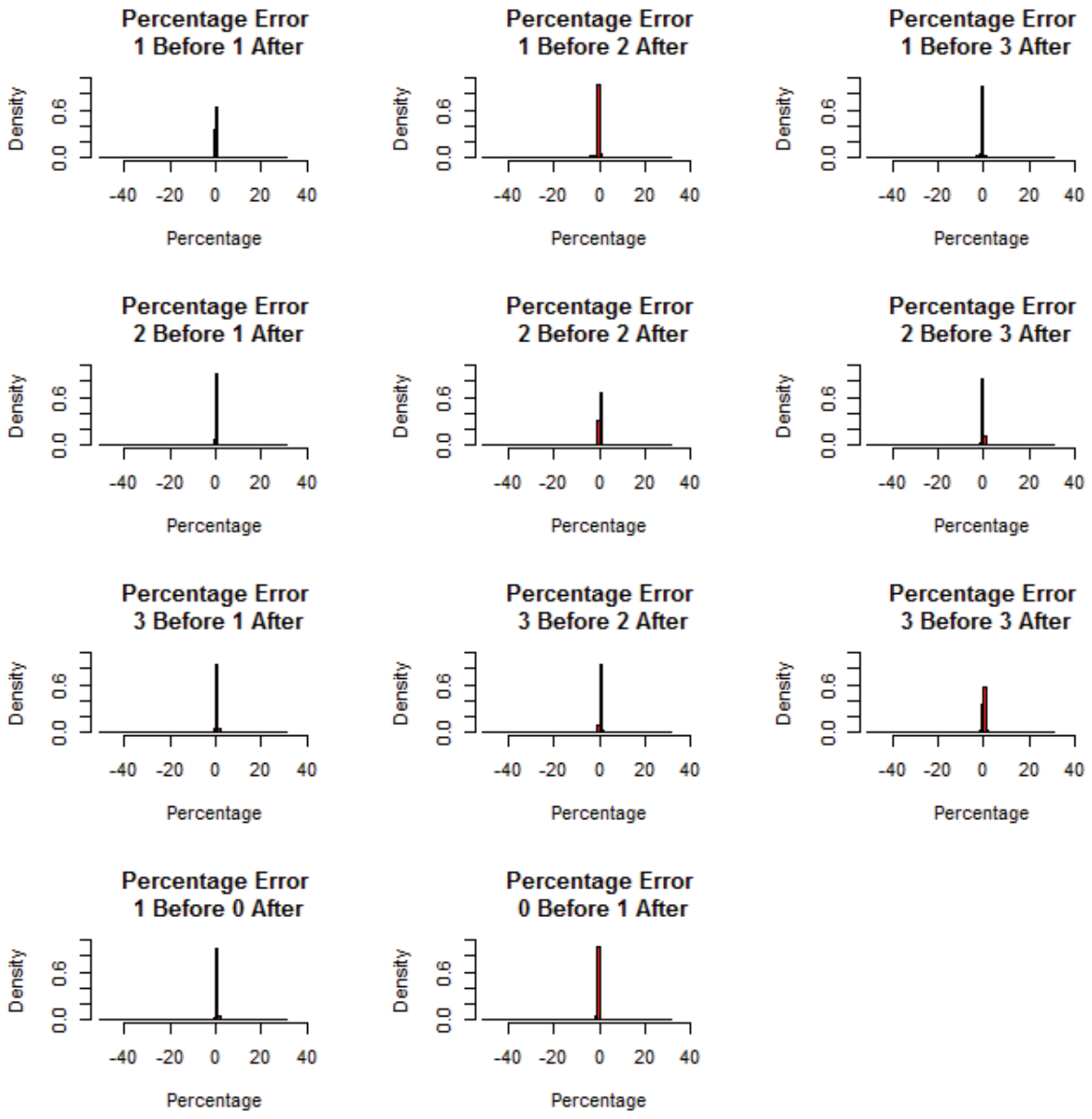


Figure 15: SoundCloud Estimated Imputation Error Method 1

*The before and after in the title indicate the number of days before and after the missing value

Total Range of Percentage Errors: -50.2% to 30.7%



For each avenue, we find that errors are consistently centered around zero. However, many of the online avenues have outlier errors or moderate spread. Imputing Facebook values performed the best with the smallest range of errors between -2.6% and 6.1% followed by Wikipedia with a range of errors between -5.9% and 7.8%. Twitter and Youtube performed moderately well with a range of errors between -12.5% - 20.0% and -7.5% - 16.9% respectively. Vevo and SoundCloud performed less than ideally with errors ranging up to 50%.

Estimating the error helped us to determine the magnitude or type of any bias in the final analysis. Given the error for Facebook was quite low and symmetric, there were minimal concerns of bias of this variable in the final model. For Twitter, Vevo, and SoundCloud, imputed with method 1, the errors were moderate and symmetric with both underestimates and overestimates. We expect the imputed values to be underestimates when more days after the missing day were used in the calculation and we expected the imputed values to be overestimates when more days before the missing day were used in the calculation. Therefore, the bias of these variables was unclear without further exploration. The error for Youtube was symmetric with a moderate level of error. While there was little concern about bias, there was more of a concern about precision. Overall, the imputation only had a marginal effect in the final analysis as the number of complete observations only slightly improved from 3.67% to 3.73%. The small increase is primarily due to the large amount of left-truncated data which was not imputed. All in all, we traded minimal bias for increased stability leading to more reliable results.

Section 6: Cox Proportional Hazards Model

In an attempt to better capture the longitudinal effects over time, a Cox Proportional Hazards model was fit to the time series of online metrics. The Cox Proportional Hazards model estimates the risk of an event occurring over time (Cox, 1972). In this case we estimated the risk that an artist is aired on the top charts of radio. The language may be counterintuitive as Cox Proportional Hazards models are traditionally used in clinical trials where the event of interest is death. However, increased risk is beneficial in our case. The estimated risk is a function of time as well as a set of covariates (formula 1 below). The covariates are an artist's online metrics over time. The estimated risk, or hazard function, at time t is converted to a proportion by

dividing by the baseline risk at time t (formula 2 below). The covariates are then modeled linearly predicting the log of the proportion (formula 3 below). Implicitly, we made the assumption that the proportional hazard is relatively constant over time. In other words, as the risk,$h(t)$, changes over time, the baseline risk, $h_0(t)$, changes in the same way over time.

$h(t)$ is the estimated risk at time t
$h_0(t)$ is the baseline risk at time t which is solely dependent on time
p is the number of covariates
β_j is the estimated coefficient and hence size of effect for the j[th] covariate

$$h(t) = h_0(t) * e^{\sum_{j=1}^{p} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p} \quad (1)$$

$$\frac{h(t)}{h_0(t)} = e^{\sum_{j=1}^{p} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p} \quad (2)$$

$$\log\left(\frac{h(t)}{h_0(t)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (3)$$

Ideally, we would observe all artists over identical time periods. However, the observed time periods are different lengths for each artist because the data is left and right censored. As we mentioned previously, the data is left-truncated, or left-censored, either because the artist is not yet active on the online channel or because Next Big Sound has not begun tracking the online source yet. The data is right-censored because the hazard or risk is observed for each subject until occurrence of the event. Subsequently, each artist time period begins when there is data available and ends once they reach the top carts of radio or till the end of the observed time period in December of 2013. Unfortunately, this may cause bias in our final findings. If the left-truncated data is missing because the online source does not exist, this is not a concern. However, if the left-truncated data is missing because Next Big Sound is not tracking it, we may be missing out on valuable information. For the right-censored data, an artist could possibly appear on a radio chart shortly after our given time period. Therefore, we may inaccurately label

some artists as never making it on a radio chart. Unfortunately, without additional information, this cannot be corrected.

The weekly data with the missing values imputed, was reformatted to a Cox Proportional Hazards model data structure as shown in Table 24. Notice each row is a week observation in which the start and stop columns indicate how many weeks have passed for the artist's specific time period. The end of an artist's observations, and thus appearance on a top radio chart, is indicated by a value of 1 in the radio column. For artists who do not appear on a radio chart, all of their weekly observations are reported with a zero in the radio column. The online metrics were scaled, for easier interpretation later on, in units of 100,000.

Table 24: Cox Proportional Hazards Model Data Structure

ID	Artist	Start	Stop	Radio	Facebook Likes (100 K)	Wikipedia Views (100 K)	Twitter Followers (100 K)	Youtube Views (100 K)	Vevo Plays (100 K)	SoundCloud Plays (100 K)
11299	Meek Mill	0	1	0	NA	0.00009	NA	NA	NA	NA
11299	Meek Mill	1	2	0	NA	0.000036	NA	NA	NA	NA
11299	Meek Mill	2	3	0	NA	0.000293	NA	NA	NA	NA
11299	Meek Mill	3	4	0	NA	0.000380	NA	NA	NA	NA
11299	Meek Mill	4	5	0	NA	0.000397	NA	NA	NA	NA
...	...	...	...	...	...	...	...	...	...	...
11299	Meek Mill	188	189	0	18.22	35.01	26.55	694.8	449.9	62.1
11299	Meek Mill	189	190	1	18.34	35.36	26.94	699.9	452.3	64.6
11429	The Maccabees	0	1	0	1.80	NA	NA	NA	NA	NA
11429	The Maccabees	1	2	0	1.83	2.0	NA	NA	NA	NA
...	...	...	...	...	...	...	...	...	...	...
11429	The Maccabees	207	208	0	2.51	2.88	0.921	42.57	99.06	NA
11429	The Maccabees	208	209	0	2.52	2.90	0.922	42.61	99.25	NA

Rather than using all 2,933 artists, we chose a subset of artists for analysis. Otherwise, there was difficulty with model convergence as there is an extremely low number of radio artists (71) compared to nonradio artists (2,862). Instead, we used all 71 radio artists and a random sample of 284 nonradio artists (four times the amount of radio artists). This proportion of nonradio artists to radio artists was still relatively large, maintaining the structure of the data, yet small enough to provide signal to predict when an artist will reach the top charts of radio.

First, univariate models of the individual online metrics were fit (Table 25). For these models, the weekly observations begin as soon as an artist has data for any one of the online metrics. Note, the number of artists analyzed for each model decreases due to missing values which could not be imputed in our current imputation scheme. Reported in the table are the ratio of radio to total artists, the coefficient, the exponential transformation of the coefficient, the p-value of the coefficient, and the assumption p-value testing if the model assumptions are met. Each univariate model did not violate the proportional hazards assumption, and each covariate is statistically significant. The estimated risk and validity of the proportional hazards assumption for each model are visually displayed in Figures 16-21. In each figure, the left plot graphs the survival curve of the estimated proportion of nonradio artists over time. Therefore, as the proportion decreases, more artists are estimated to be on the top charts of radio. Plotted around the estimated survival curve are 95% confidence bands. The right plot displays the proportional hazard over time. A spline smoother is fit to the Schoenfield residuals as a nonlinear function of time. A horizontal line indicates adherence to the proportional hazards assumption. Again, 95% confidence bands are plotted around the proportional hazard line. While some of the curves look wavy, note that the y-axis values are very small. Further, adherence to the proportional hazard assumption was statistically tested; and the result is listed as the proportional hazard assumption p-value in the last column on Table 25. A p-value above 0.05 indicates that we do not reject the null hypothesis of proportional hazards and the proportional hazard assumption is met.

For each model, at around 150 weeks, we see a large decrease in the proportion of nonradio artists and hence increase in the number of radio artists. However, Vevo behaved differently with a sharp decrease around 50 weeks. Therefore, an artist's activity may not be particularly telling until about 1 to 3 years after they establish an online presence.

Table 25: Univariate Cox Proportional Hazards Models – All Weeks

Bolded models are statistically significant at the 1% level

Radio/Total # of Artists	Variable	Univariate Coefficient	Exponential of Coefficient	Univariate P-Value	Proportional Hazard Assumption P-Value
64/343	**Facebook Page Likes**	**0.030**	**1.030**	**$3.8*10^{-10}$**	**0.16**
63/217	**Wikipedia Page Views**	**0.012**	**1.012**	**$8.3*10^{-13}$**	**0.83**
67/322	**Twitter Followers**	**0.053**	**1.05**	**$2.0*10^{-16}$**	**0.82**
62/282	**Youtube Plays**	**0.00035**	**1.00035**	**0.00071**	**0.85**
51/126	**Vevo Plays**	**0.00045**	**1.00045**	**$6.05*10^{-9}$**	**0.23**
46/228	**SoundCloud Plays**	**0.031**	**1.032**	**3.0410^{-14}**	**0.16**

Figure 16: Facebook Univariate Risk Estimate and Diagnostic

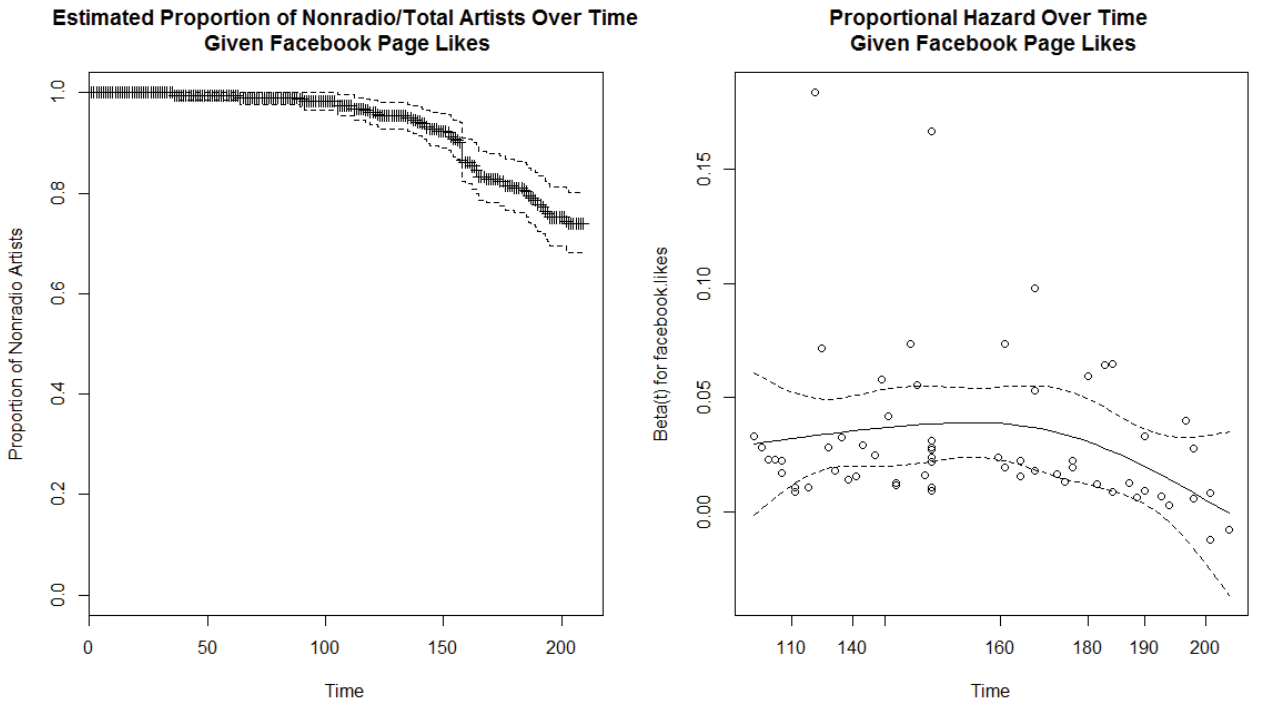


Figure 17: Wikipedia Univariate Risk Estimate and Diagnostic

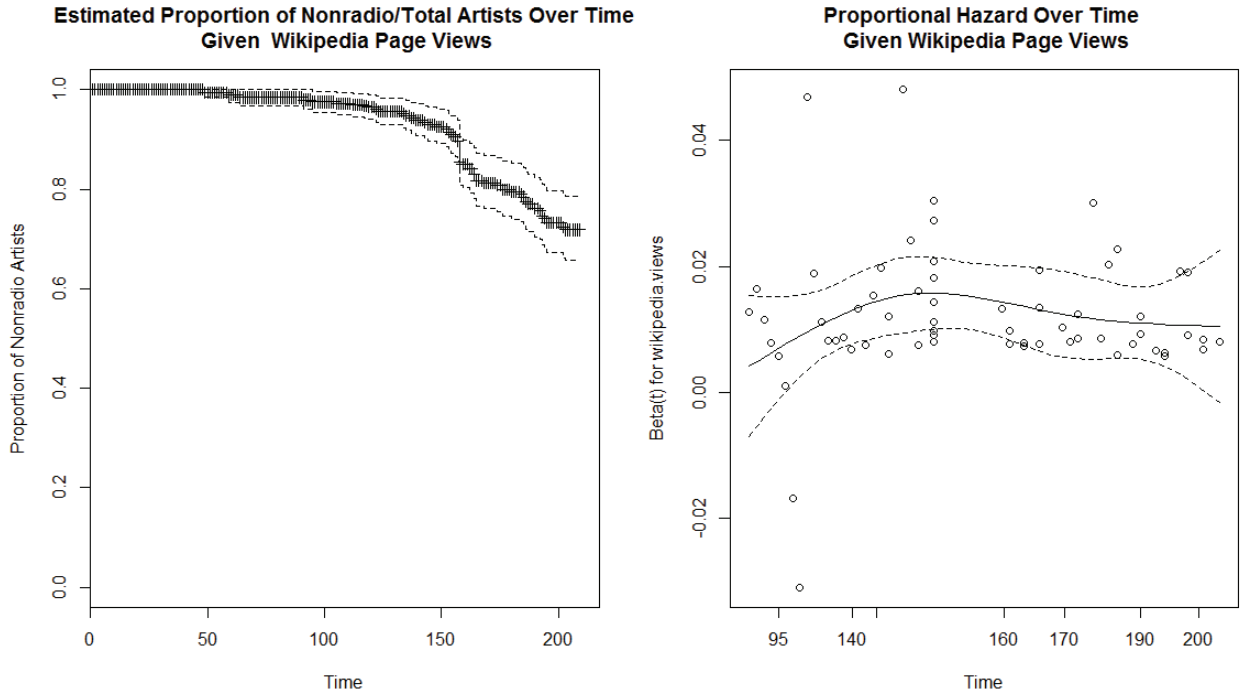


Figure 18: Twitter Univariate Risk Estimate and Diagnostic

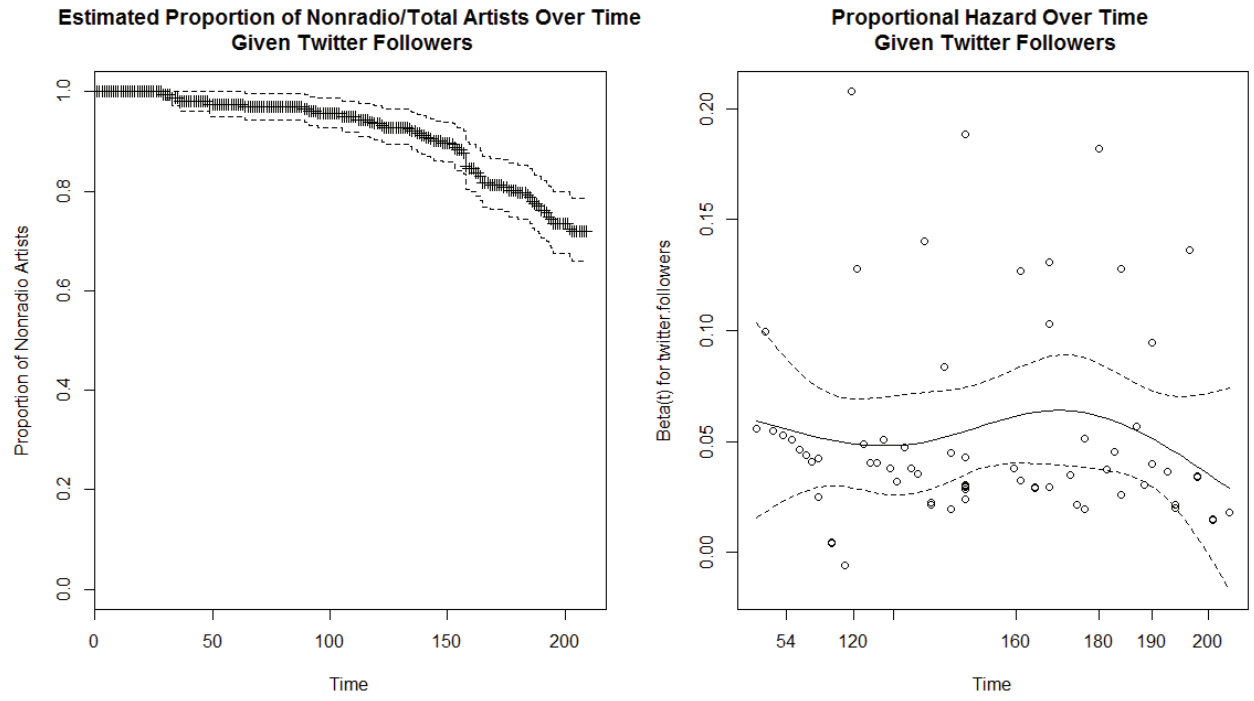


Figure 19: Youtube Univariate Risk Estimate and Diagnostic

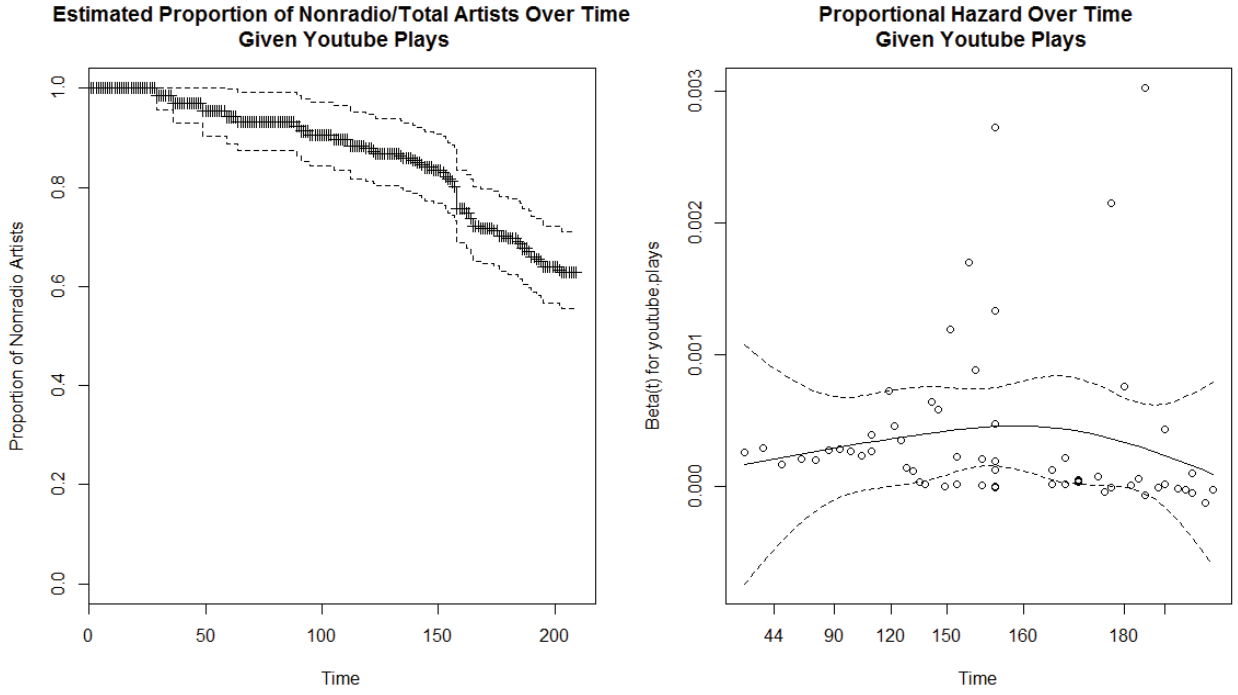


Figure 20: Vevo Univariate Risk Estimate and Diagnostic

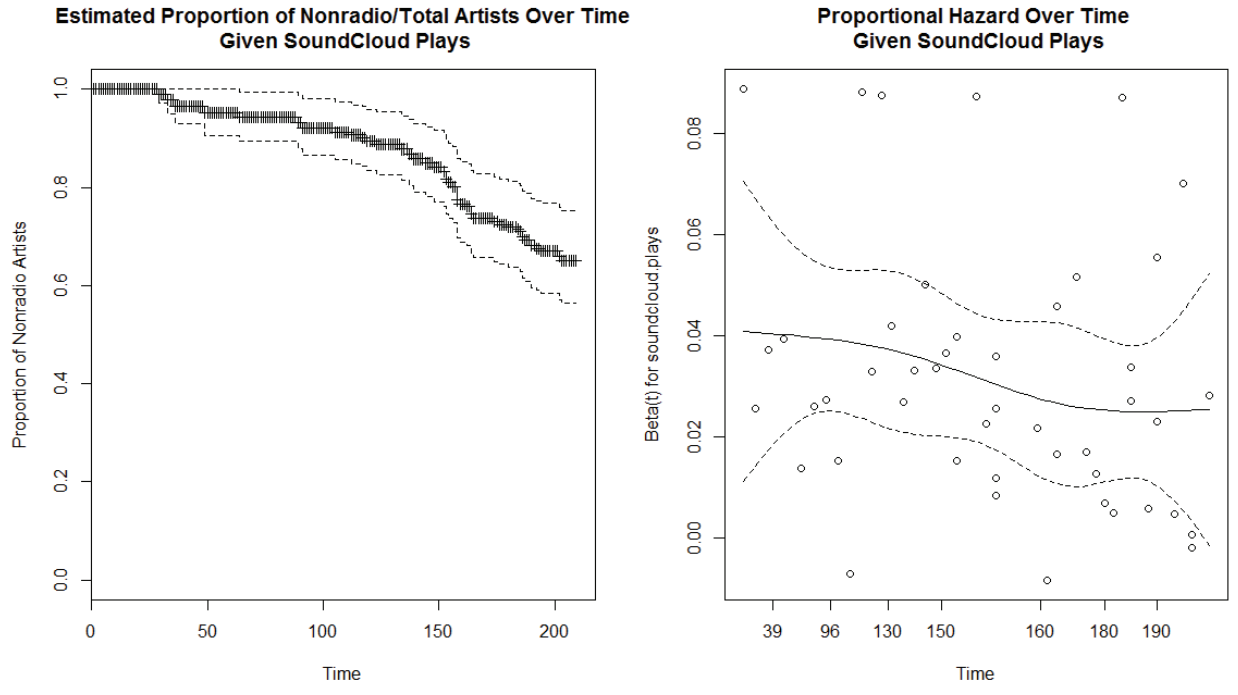
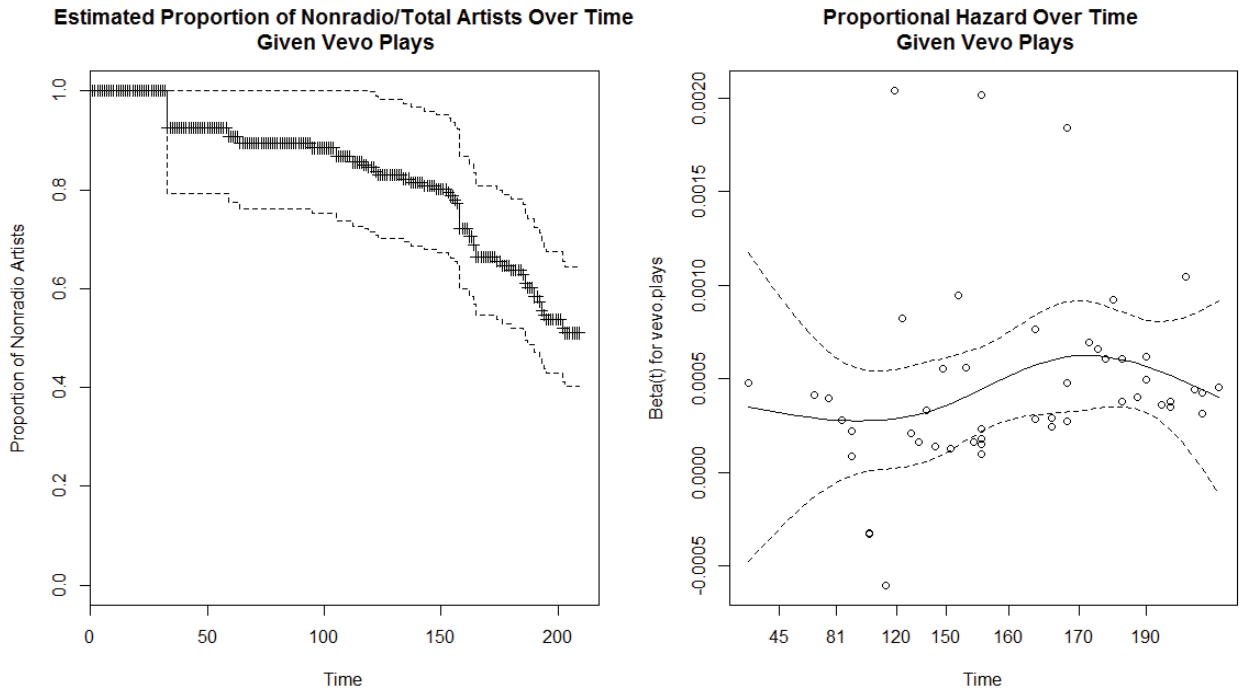


Figure 21: SoundCloud Univariate Risk Estimate and Diagnostic



The results of the multivariate model are shown in Table 26 with the survival curve of the estimated proportion of nonradio artists and corresponding diagnostic plots shown in Figures 22-23. Again the model satifies the proportional hazard assumption as the global assumption p-value and the individual assumption p-values for each metric indicate insignificance. Although all the online metrics were statistically significant in their univariate models, the only statistically significant variable in the multivariate model is SoundCloud. In this model it appers that Vevo plays may be significantly influencing the model results as we see a large change in the number of radio artists after 50 weeks, similar to the univaritae Vevo model. However notice, the number of artists analyzed dropped drastically due to missing values. Now only 22 radio artists and 38 nonradio artists were analyzed. In hopes of analyzing a larger sample size of artists, another model with all key metrics was fit excluding Vevo Plays (which has the most missing data of all the metrics). The number of radio artists increased to 31 radio artists and 111 nonradio artists. As a result, SoundCloud plays, Youtube plays, and Twitter followers were all found to have a statistically significant relaitonships at the 1% level and again we see a large drop in nonradio artists around 50 weeks (See Appendix C).

Table 26: Multivariate Cox Proportional Hazards Model - All Weeks

22 Radio Artist, 60 Total Artists

Global Proportion Hazard Assumption = 0.85

Bolded models are statistically significant at the 1% level

Variable	Coefficient	Exponential of Coefficient	P-Value	Proportional Hazard Assumption P-Value
Facebook Page Likes	-0.027	0.973	0.33	0.64
Wikipedia Page Views	-0.0053	0.995	0.40	0.35
Twitter Followers	0.031	1.031	0.13	0.53
Youtube Plays	0.00045	1.00045	0.13	0.93
Vevo Plays	0.00073	1.00073	0.12	0.49
SoundCloud Plays	**0.0205**	**1.0207**	**0.0047**	**0.73**

Figure 22: Multivariate Risk Estimate – All Weeks

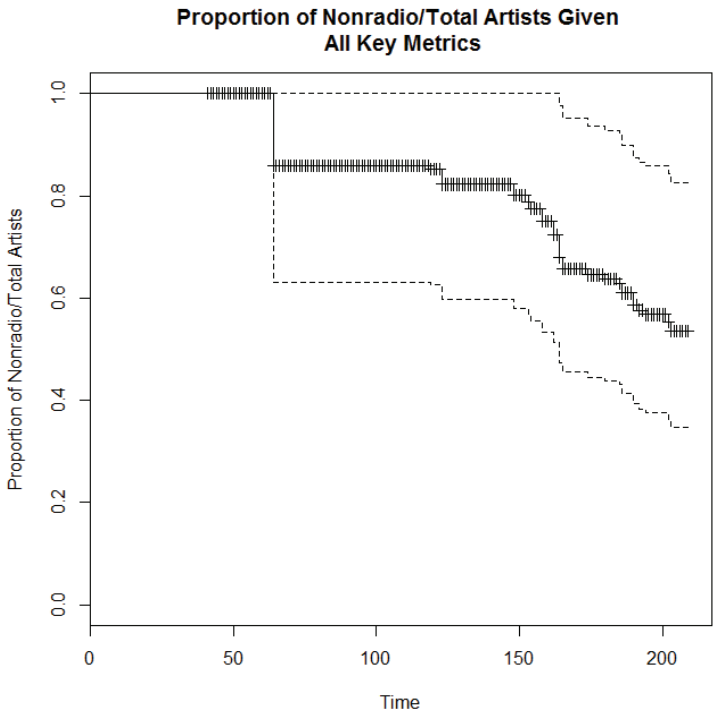
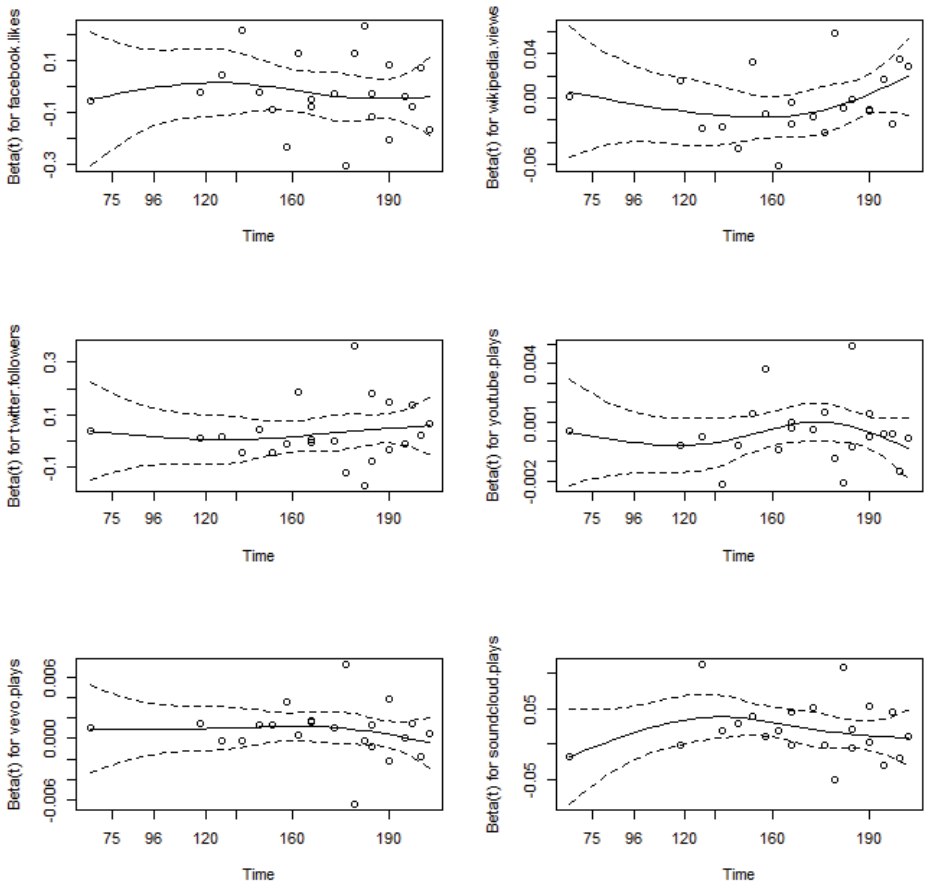


Figure 23: Multivariate Diagnostics – All Weeks



A large number of weekly observations were removed from the previous models due to missingness. Therefore, we again attempted to model the data, but this time each artist's weekly observations began once data was available for all metrics, rather than just one metric. With this approach the sample size was reduced to 23 radio artists and 290 nonradio artists. The univariate results form this approach were similar to the univariate results from the previous approach. Facebook, Youtube, and Vevo had slightly stronger statistical significance, and the statistical significance of SoundCloud decreased slightly. Models for Wikipedia and Twitter did not converge, likely due to fewer observations per artist. The multivariate model is summarized in Table 27. The estimated survival curve of the proportion of nonradio artists over time and the diagnostic plots are shown in Figures 24-25.

Table 27: Multivariate Cox Proportional Hazards Model – Complete Weeks

23 Radio Artist, 290 Total Artists

Global Proportion Hazard Assumption = 0.45

Bolded models are statistically significant at the 1% level

Variable	Coefficient	Exponential of Coefficient	P-Value	Proportional Hazard Assumption P-Value
Facebook Page Likes	$-4.4*10^{-7}$	1	0.18	0.12
Wikipedia Page Views	$-7.3*10^{-8}$	1	0.41	0.97
Twitter Followers	$\mathbf{1.0*10^{-6}}$	**1**	$\mathbf{2.6*10^{-5}}$	**0.056**
Youtube Plays	$7.9*10^{-10}$	1	0.75	0.11
Vevo Plays	$\mathbf{5.9*10^{-9}}$	**1**	**0.0088**	**0.44**
SoundCloud Plays	$\mathbf{1.5*10^{-7}}$	**1**	**0.0087**	**0.75**

Figure 24: Multivariate Risk Estimate – Complete Weeks

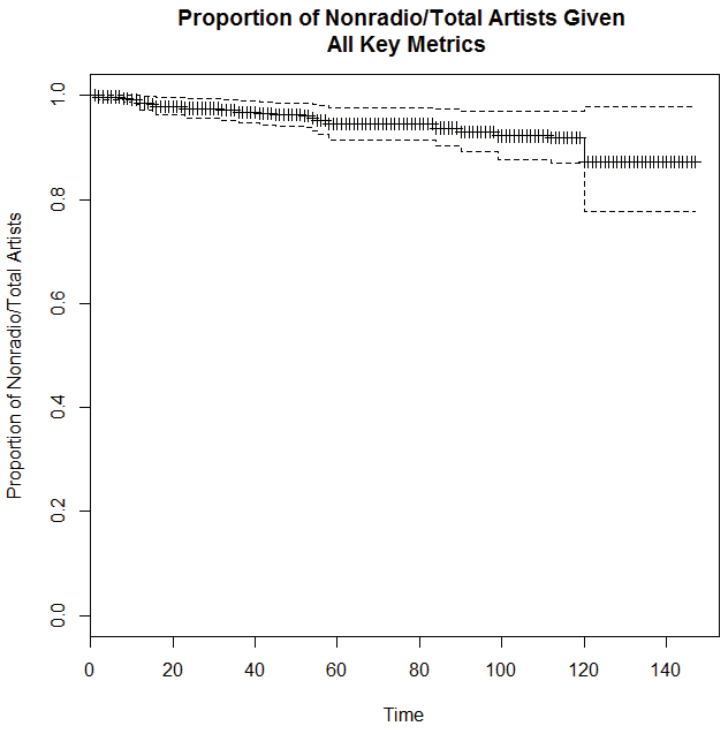
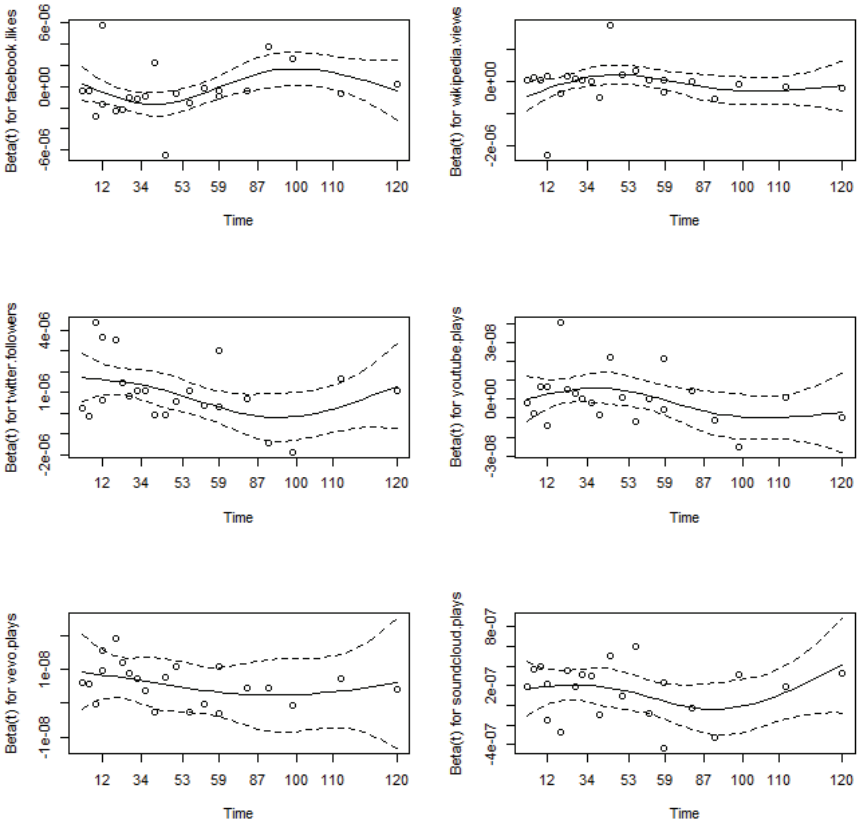


Figure 25: Multivariate Diagnostics – Complete Weeks



The model satisfies the proportional hazards assumption as the global assumption p-value is 0.45. However individually, Twitter shows some evidence of varying with time with a p-value of 0.056, thus possibly violating the assumption. Compared to the previous multivariate model, Twitter followers, Vevo plays, and SoundCloud plays are all statistically significant. Further, we see an increase in radio artists in the survival curve (Figure 24) at around 120 weeks after establishing an online presence across all mediums.

All in all, it may be in the best interest for artist and record labels to focus their attention on Twitter, Vevo, and SoundCloud when promoting themselves. Interestingly, Facebook, which is usually a major player in the online world due to its vast reach on the population, was not statistically significant. This lack of association may be due to the ubiquitous use of Facebook among artists and fans. To think about it in another way, Facebook page likes are cheap and are

therefore, not unique enough to signal any major change in the likelihood of an artist reaching the top charts of radio. Instead, we may see spikes in Facebook page likes after an artist has been successfully aired on the top charts of radio (rather than before), as was the case for Bastille (Section 3, Figure 4). Further, it appears the expected time range for artists to reach radio, once they have established an online presence, is between 1 to 3 years.

Section 7: Discussion

Let us now examine the entire analysis as a whole. In the initial steps of data collection and data cleaning, not all steps were automated and some steps required extensive human labor. In particular, it would be beneficial to create an entirely automated program to clean and identify all unique radio artists in the radio data. When creating the artist relational database, our record linkage techniques performed well with an expected accuracy of 99.7%. Given that we are only matching on the artist text string name, this is a notable success. The success of the method is most likely due to the sampling method of decisively sampling from different Jaro-Winkler ranges rather than using a simple random sample. Future work in this area may need to adopt a similar approach. In the first modeling attempt, logistic regression found some significant relationships but did not classify any radio artists. Finally, the Cox Proportional Hazard model performed well by incorporating the longitudinal aspect of the data. However, it is dependent on the imputation methodology. While the simple average imputation method performed moderately well, we may be able to improve our performance by implementing more advanced model-based imputation techniques.

The biggest limitation of the data was the number of missing data values. As we mentioned, the data are left-truncated. Unless data are collected for the entirety of every artist's career, the data will always be truncated. However, if we knew the cause of the truncation (either the online source does not exist or Next Big Sound does not yet track the source) we could better understand how the truncation biases the results. In regards to the radio play data, we do not have any information about radio airplay on satellite radio stations such as XM satellite radio. Therefore, we may be missing some artists, or the ranks of the charts may be slightly different once radio satellite airplay is included. At the same time, we do not expect the

exclusion of satellite radio to drastically affect the results as the songs and artists tend to be the same across the two radio types.

The accuracy of the results could further be improved by more accurately identifying "discovered" artists. First we could apply a more stringent data matching algorithm to identify "discovered" artists when linking the radio play data with the Next Big Sound predictions. Currently, the algorithm only uses the artist text string names which could be inaccurate as it is possible for a band/artist to have the same name as another band/artist. If we fail to match an artist and thus inaccurately classify an artist as a nonradio artist, the relationship between an artist's online activity and radio plays will be misrepresented. If we had more information to match on, such as artist songs or genre, the accuracy of the results could be greatly enhanced. The relationship may also be misrepresented because our time periods of the radio play data and the metric data are not entirely overlapping. While we have online metric data and predictions from 2010 to 2013, we only have radio play data for 2013. Therefore, if an artist reached the top charts of radio either before or after 2013, we miss this information and inaccurately classify the artist as a nonradio artist.

Given the information we do have, the analysis conducted was just the tip of the iceberg. In addition to the six online metrics examined: Facebook page likes, Wikipedia page views, Twitter followers, Youtube plays, Vevo plays, and SoundCloud plays, it would have been interesting to explore the interaction among these variables. Further, Next Big Sound also reports metrics for Last.fm, Instagram, and Tumblr. In total, there are thirty online metrics across the nine online sources (See Appendix D for the complete list of variables). We were only recently able to download these thirty metrics for all artists. At the same time, using more variables may further limit sample size as all examined artists must have complete data for all online sources. Moving beyond the online variables, one could also do a demographic analysis by type of artist such as: genre, geographic location, age of artist, length of music career, solo artist vs. band and more. Finally, in future research, we should examine other possible longitudinal modeling techniques.

References:

Bernhardsson, Erik. “Music Recommendations at Spotify.” 25 Jan. 203. SlideShare.com. < http://www.slideshare.net/erikbern/collaborative-filtering-at-spotify-16182818> 2 Mar. 2014.

Bilenko, Mikhail, Raymond Mooney, William Cohen, Pradeep Ravikumar, and Stephen Fienberg. "Adaptive Name Matching in Information Integration." *Carnegie Mellon School of Computer Science*. IEEE Computer Society, 2003. Web. 6 Jan. 2013.

Cox, D.R. 1972. “Regression Models and Life Tables (with Discussion).” Journal of the Royal Statistical Society, Series B 34: 187-220.

Herzog, Thomas N., Fritz Scheuren, and William E. Winkler. "13 Strong Comparator Metrics for Typographical Error." *Data Quality and Record Linkage Techniques*. New York: Springer, 2007. N. pag. Print.

Fellegi and Sunter. “A Theory for Record Linkage.” Journal of the American Statistical Association 64(328): pp 1,183-1,210.

Lefsetz. “Internet Killed the Radio, and Now Netizens are in Control.” Variety Media LLC, 1 Oct. 2013. Web. 26 Mar. 2014. < http://variety.com/2013/biz/news/internet-killed-the-radio-star-and-now-netizens-are-in-control-1200689229/>.

Nielsen Company, LLC. "How People Are Consuming Music." *Nielsen*. 14 Aug. 2012. Web. 01 Oct. 2013. < http://www.nielsen.com/us/en/press-room/2012/music-discovery-still-dominated-by-radio--says-nielsen-music-360.html >.

Nielsen Company, LLC. “Radio Increases Year-Over-year Reach by More than 1.2 Million, According to March 2014 RADAR Report.” *Nielsen*. 10 Mar. 2014. Web. 28 Mar. 2014. < http://www.nielsen.com/us/en/press-room/2014/radio-increases-year-over-year-reach-by-more-than-1-2-million.html?>.

Orpheus Media Research, LLC. “2011 Orpheus Media Research Consumer Survey: Executive Summary.” Orpheus Media Research, LLC, Feb. 2011. Web. 26 Mar. 2014. < http://www.cliomusic.com/wp-content/uploads/2011/04/omr.executivesummary.consumer_110324-1500.pdf>

Pacula, Maciej. *A Matrix Factorization Algorithm for Music Recommendation Using Implicit User Feedback*. MIT CSAIL, n.d. Web. <http://mpacula.com/publications/lastfm.pdf>.

"The Zero Button Music Player." *Music Machinery*. N.p., 14 Jan. 2014. Web. 28 Mar. 2014. < http://musicmachinery.com/2014/01/14/the-zero-button-music-player-2>.

Appendix A: Modeling results of Initial Logistic Regression using absolute daily metric values

Table 1: Facebook Univariate Logistic Regression Models

Radio/Total # of Artists	Variable	Coefficient	P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Accuracy Variance
66/2,857	Avg. Day	$-5.2*10^{-7}$	0.27	0.50	0.97	0	1	0
66/2,857	Avg. Week	$-7.4*10^{-8}$	0.27	0.50	0.97	0	1	0
66/2,857	Avg. Month	$-1.7*10^{-8}$	0.27	0.50	0.97	0	1	0
66/2,857	Max Inc.	$-2.3*10^{-7}$	0.27	0.50	0.97	0	1	0
66/2,857	Agg Peak A	$-4.3*10^{-10}$	0.38	0.50	0.97	0	1	0
66/2,857	Agg Peak B	$-5.9*10^{-10}$	0.23	0.50	0.97	0	1	0
66/2,857	Agg Peak C	$-5.9*10^{-10}$	0.23	0.50	0.97	0	1	0
66/2,857	Slope	-0.00028	0.21	0.50	0.97	0	1	0
66/2,857	Percentage	-0.013	0.91	0.50	0.97	0	1	0
66/2,857	Rank	-0.0086	0.68	0.50	0.97	0	1	0

Table 2: Wikipedia Univariate Logistic Regression Models

Radio/Total # of Artists	Variable	Coefficient	P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Accuracy Variance
63/1,549	Avg. Day	$1.1*10^{-8}$	0.90	0.50	0.96	0	1	0
63/1,549	Avg. Week	$1.6*10^{-9}$	0.90	0.50	0.96	0	1	0
63/1,549	Avg. Month	$3.7*10^{-10}$	0.90	0.50	0.96	0	1	0
63/1,549	Max Inc.	$4.0*10^{-9}$	0.90	0.50	0.96	0	1	0
63/1,549	Agg Peak A	$4.2*10^{-12}$	0.95	0.50	0.96	0	1	0
63/1,549	Agg Peak B	$4.3*10^{-12}$	0.95	0.50	0.96	0	1	0
63/1,549	Agg Peak C	$4.3*10^{-12}$	0.95	0.50	0.96	0	1	0
63/1,549	Slope	$3.5*10^{-6}$	0.94	0.50	0.96	0	1	0
63/1,549	Percentage	-0.00018	0.57	0.50	0.96	0	1	0
63/1,549	Rank	-0.046	0.027	0.50	0.96	0	1	0

Table 3: Twitter Univariate Logistic Regression Models

Radio/Total # of Artists	Variable	Coefficient	P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Accuracy Variance
66/2,623	Avg. Day	$1.05*10^{-7}$	0.71	0.50	0.97	0	1	0
66/2,623	Avg. Week	$1.5*10^{-8}$	0.71	0.50	0.97	0	1	0
66/2,623	Avg. Month	$3.5*10^{-9}$	0.71	0.50	0.97	0	1	0
66/2,623	Max Inc.	$4.6*10^{-8}$	0.78	0.50	0.97	0	1	0
66/2,623	Agg Peak A	$1.1*10^{-10}$	0.74	0.50	0.97	0	1	0
66/2,623	Agg Peak B	$1.7*10^{-10}$	0.57	0.50	0.97	0	1	0
66/2,623	Agg Peak C	$1.7*10^{-10}$	0.57	0.50	0.97	0	1	0
66/2,623	Slope	$4.7*10^{-5}$	0.40	0.50	0.97	0	1	0
66/2,623	Percentage	0.00083	0.93	0.50	0.97	0	1	0
66/2,623	Rank	-0.016	0.45	0.50	0.97	0	1	0

Table 4: Youtube Univariate Logistic Regression Models

Radio/Total # of Artists	Variable	Coefficient	P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Accuracy Variance
61/2,283	Avg. Day	$-1.3*10^{-8}$	0.24	0.50	0.97	0	1	0
61/2,283	Avg. Week	$-1.9*10^{-9}$	0.24	0.50	0.97	0	1	0
61/2,283	Avg. Month	$-4.4*10^{-10}$	0.24	0.50	0.97	0	1	0
61/2,283	Max Inc.	$-9.3*10^{-9}$	0.23	0.50	0.97	0	1	0
61/2,283	Agg Peak A	$-2.4*10^{-11}$	0.22	0.50	0.97	0	1	0
61/2,283	Agg Peak B	$-1.5*10^{-11}$	0.23	0.50	0.97	0	1	0
61/2,283	Agg Peak C	$-2.4*10^{-11}$	0.20	0.50	0.97	0	1	0
61/2,283	Slope	$-1.0*10^{-5}$	0.21	0.50	0.97	0	1	0
61/2,283	Percentage	-0.0071	0.77	0.50	0.97	0	1	0
61/2,283	Rank	-0.010	0.64	0.50	0.97	0	1	0

Table 5: SoundCloud Univariate Logistic Regression Models

Radio/Total # of Artists	Variable	Coefficient	P-Value	Radio Threshold	Accuracy	Sensitivity	Specificity	Accuracy Variance
45/1,849	Avg. Day	$-2.5*10^{-5}$	0.20	0.50	0.98	0	1	0
45/1,849	Avg. Week	$-3.5*10^{-8}$	0.20	0.50	0.98	0	1	0
45/1,849	Avg. Month	$-8.2*10^{-9}$	0.20	0.50	0.98	0	1	0
45/1,849	Max Inc.	$-4.6*10^{-8}$	0.40	0.50	0.98	0	1	0
45/1,849	Agg Peak A	$-7.6*10^{-10}$	0.17	0.50	0.98	0	1	0
45/1,849	Agg Peak B	$-5.9*10^{-10}$	0.16	0.50	0.98	0	1	0
45/1,849	Agg Peak C	$-5.9*10^{-10}$	0.17	0.50	0.98	0	1	0
45/1,849	Slope	$-3.1*10^{-6}$	0.80	0.50	0.98	0	1	0
45/1,849	Percentage	-0.041	0.76	0.50	0.98	0	1	0
45/1,849	Rank	0.0033	0.90	0.50	0.98	0	1	0

No models were fit for Youtube because there were no radio artists in the sample.

Appendix B: Estimation of imputation error for methods not chosen for final imputation

Figure1: Facebook Estimated Imputation Error Method 2

Total Range of Percentage Errors: -15.1% to 9.9%

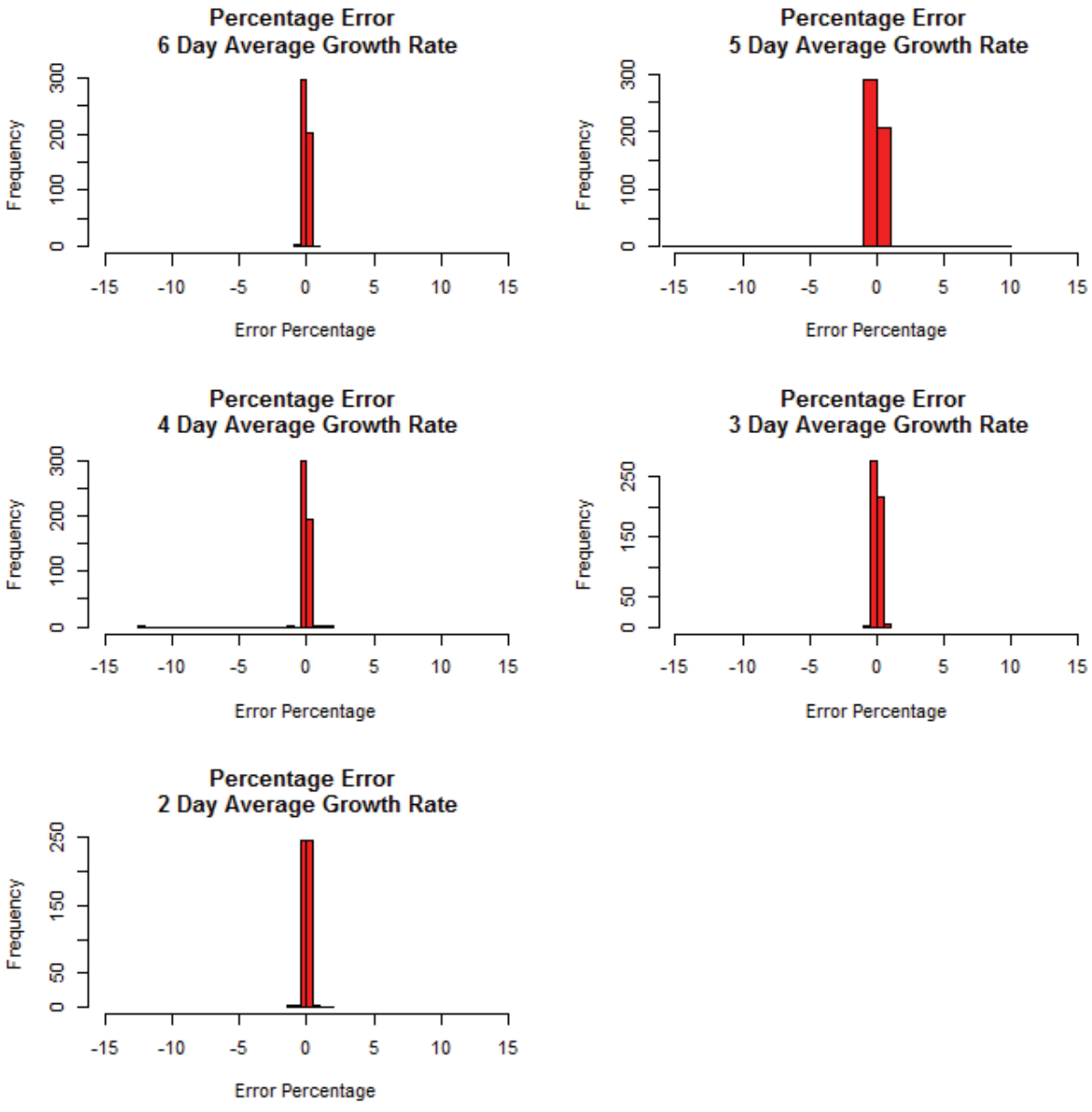


Figure 10: Wikipedia Estimated Imputation Error Method 2

Total Range of Percentage Errors: -34.6% to 3.97%

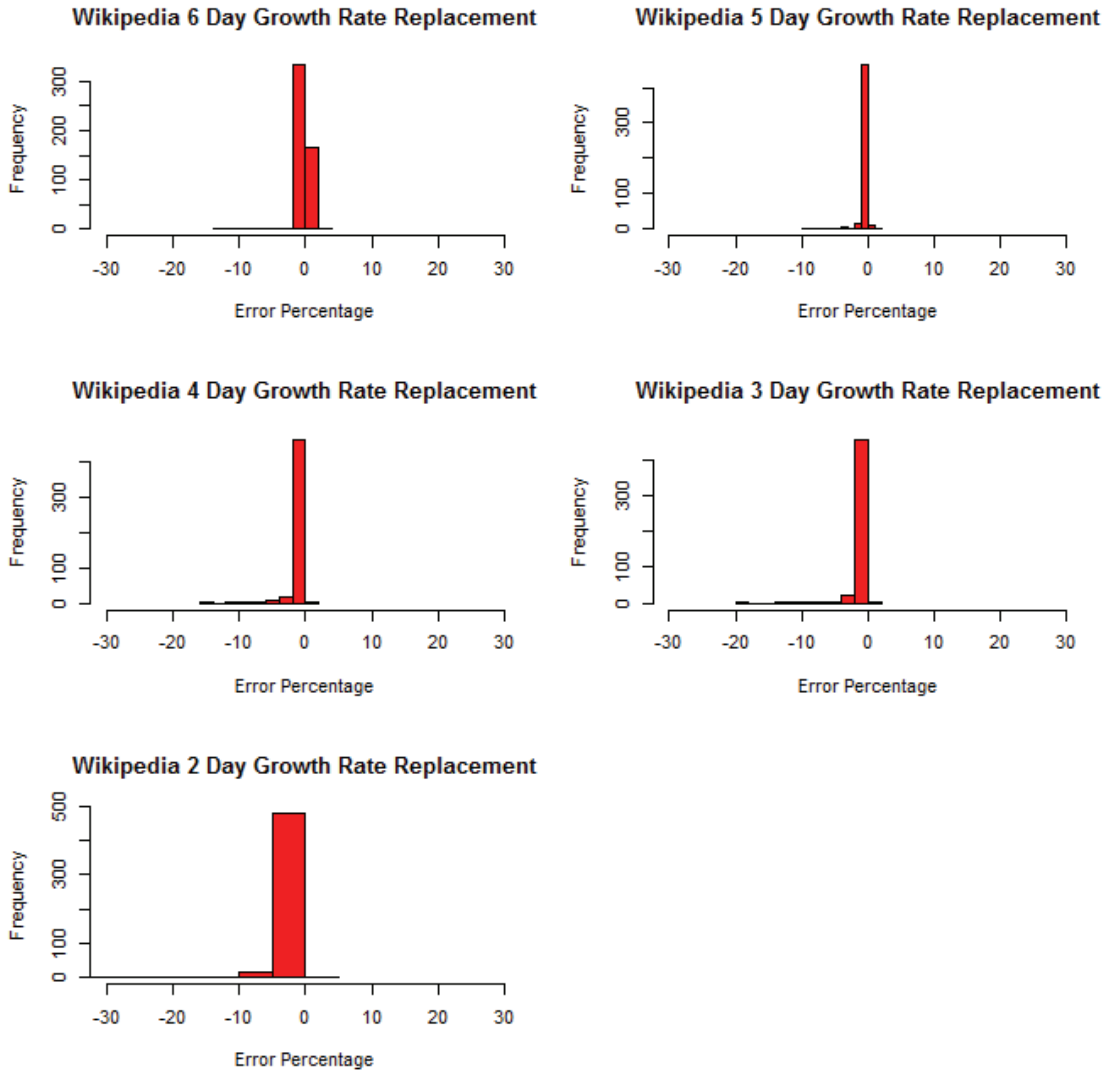


Figure 3: Twitter Estimated Imputation Error Method 2

Total Range of Percentage Errors: -1,5392.5% to 9.7%

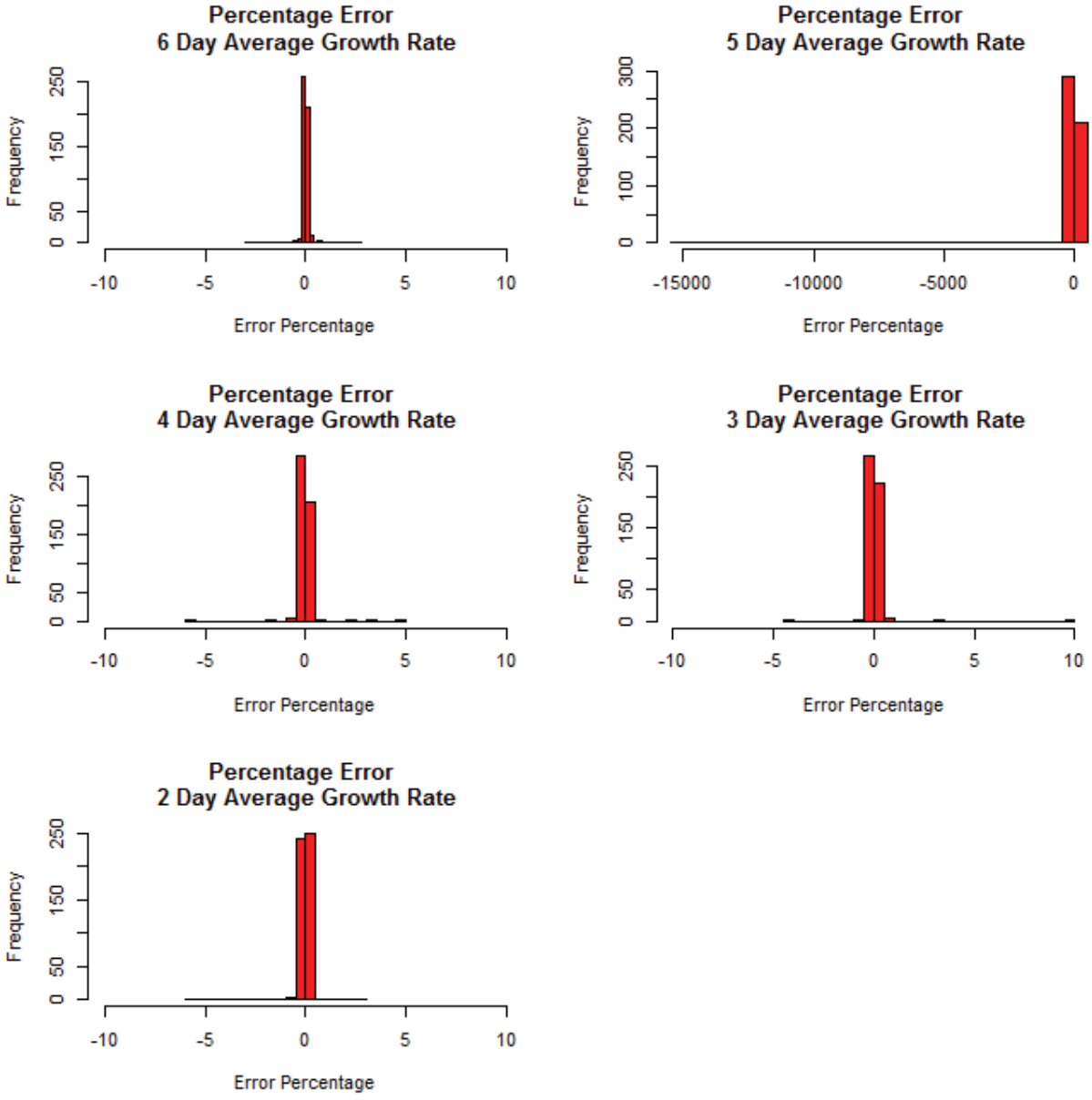


Figure 4: Youtube Estimated Imputation Error Method 1

*The before and after in the title indicate the number of days before and after the missing value

Total Range of Percentage Errors: -21.6% to 41.6%

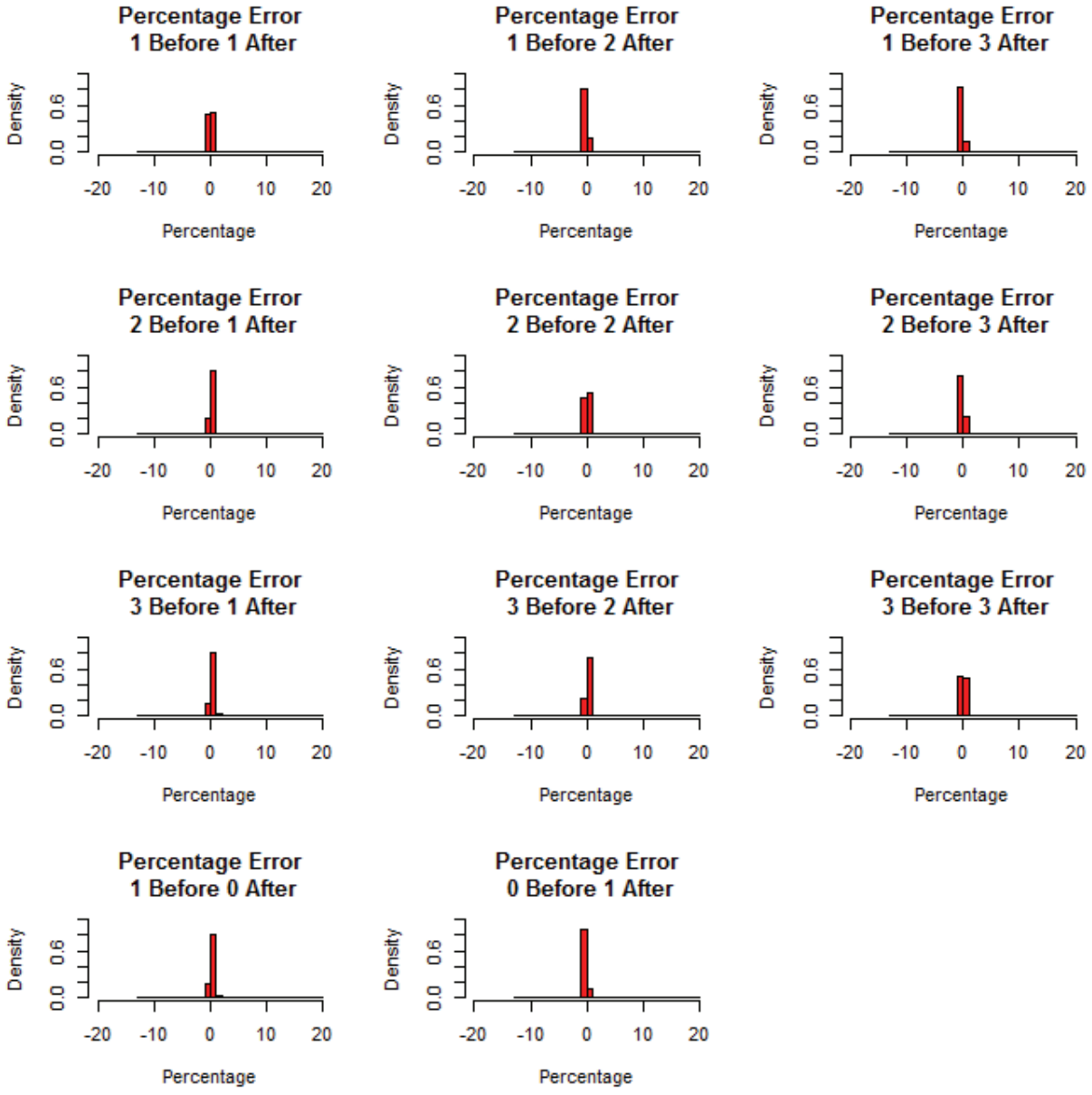


Figure 5: Vevo Estimated Imputation Error Method 2

Total Range of Percentage Errors: -3,942.3 % to 57.8%

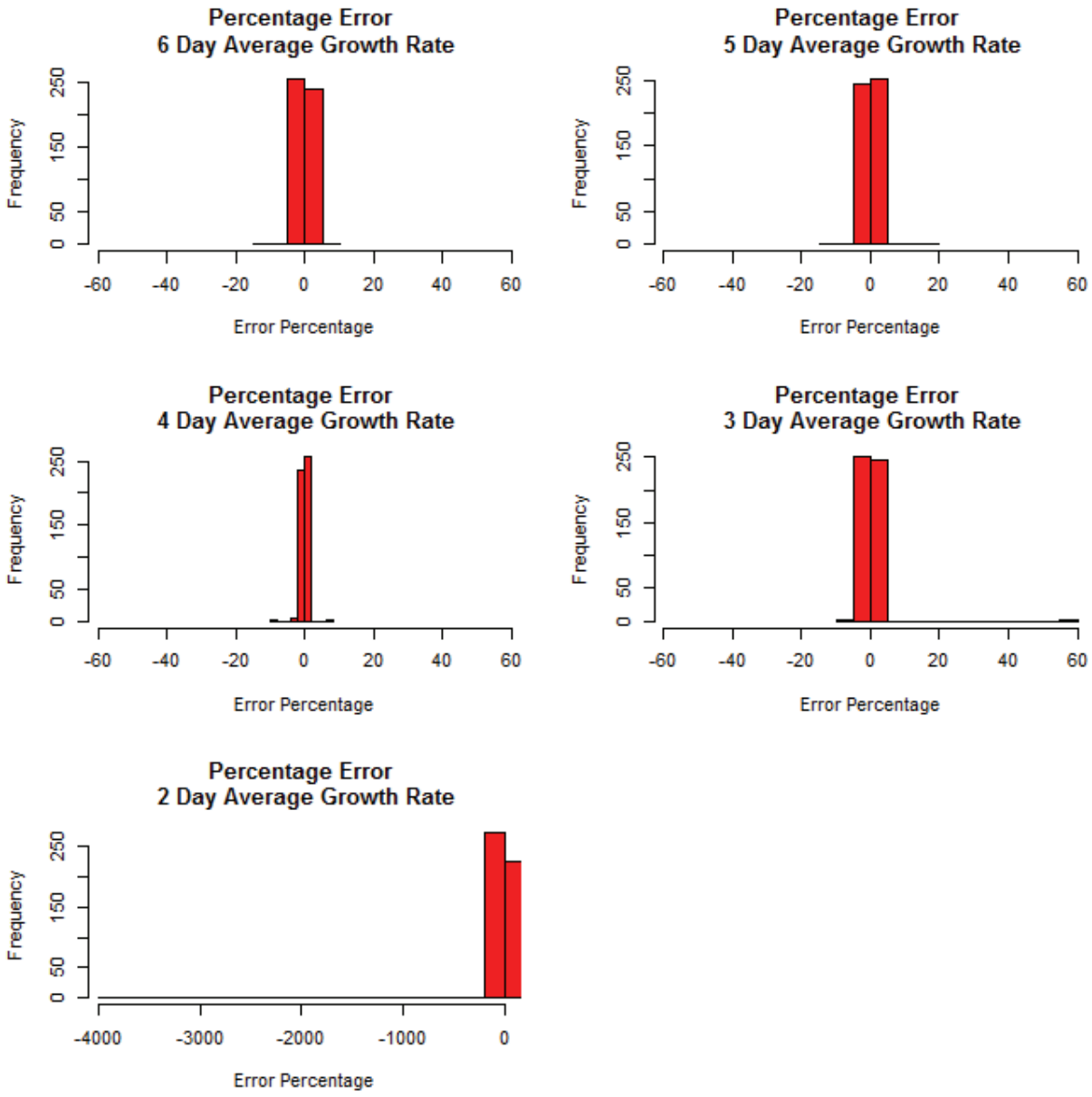
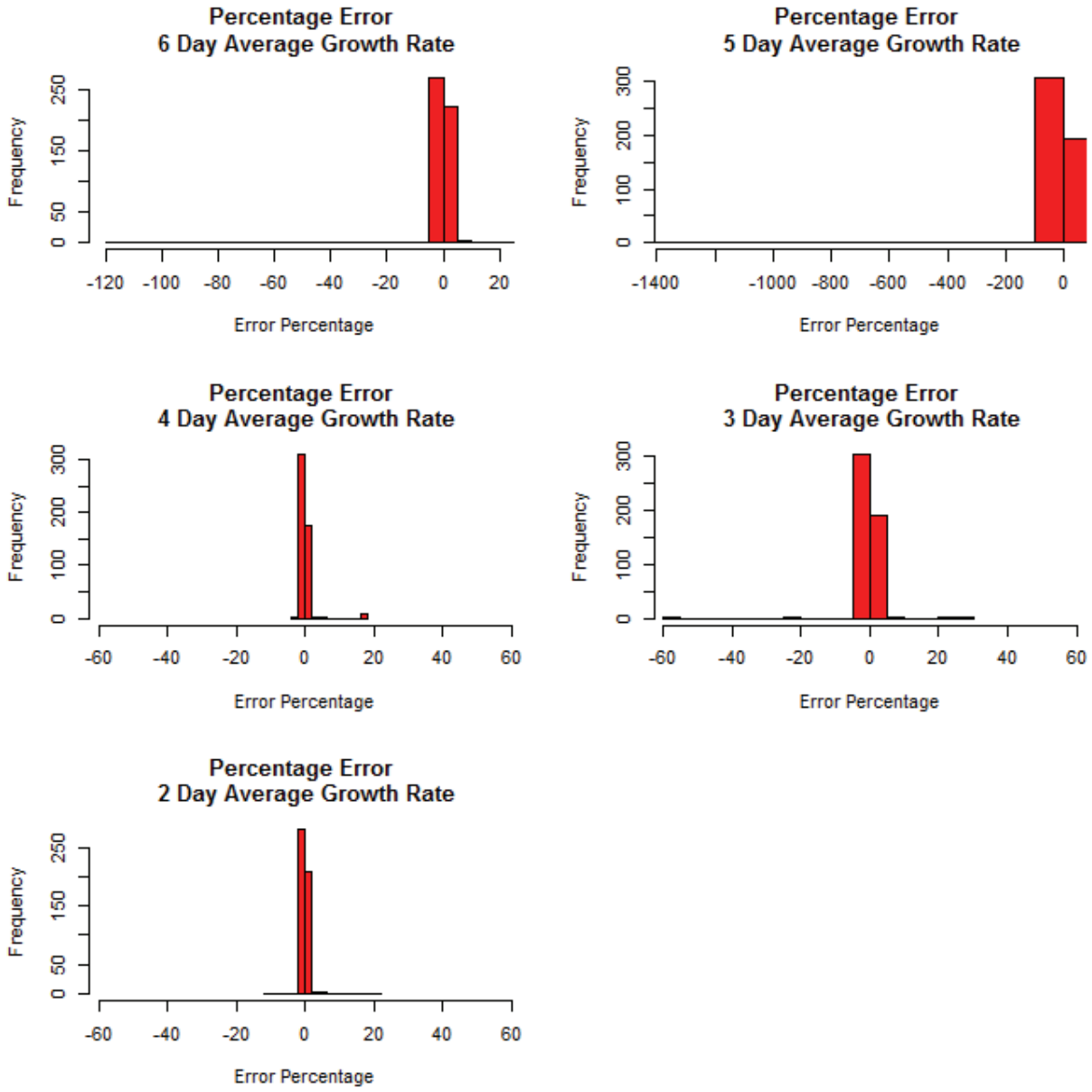


Figure 20: SoundCloud Estimated Imputation Error Method 2

Total Range of Percentage Errors: -1,353.1% to 25.1%



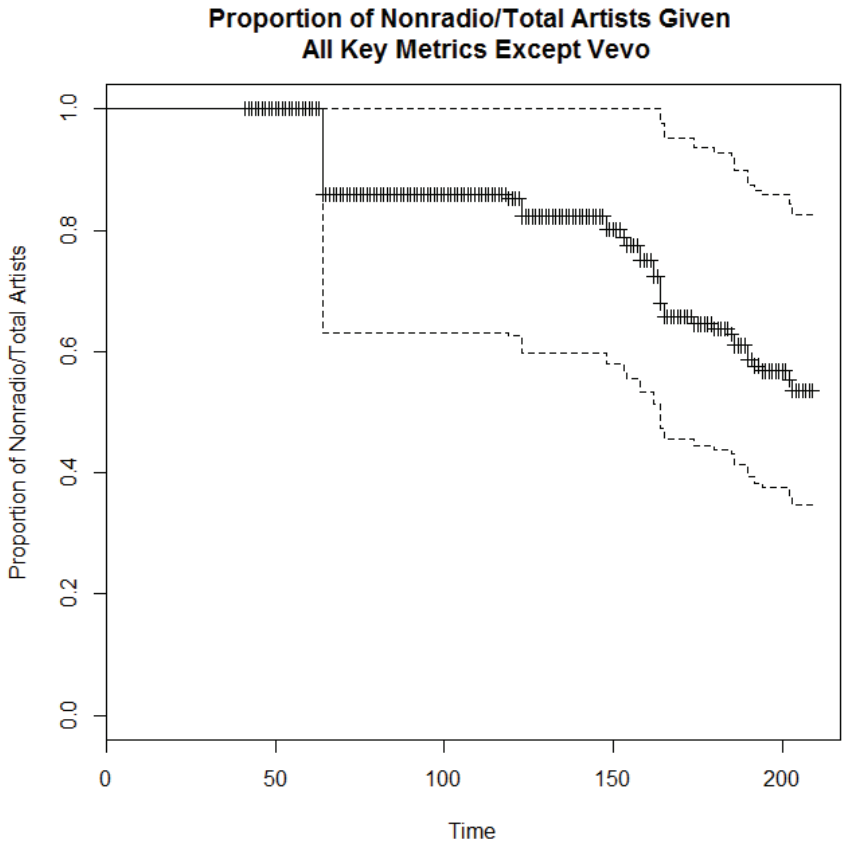
Appendix C: Cox Proportional Hazard Model of All Key Metrics Excluding Vevo Plays

Multivariate Cox Proportional Hazards Model Excluding Vevo – All Weeks

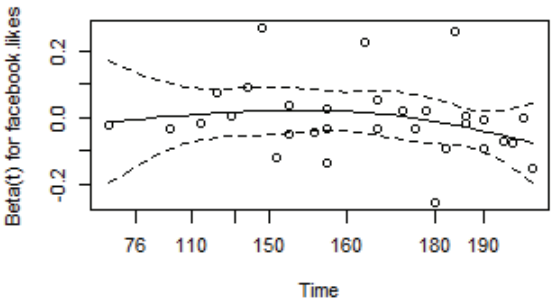
31 Radio Artist, 111 Total Artists

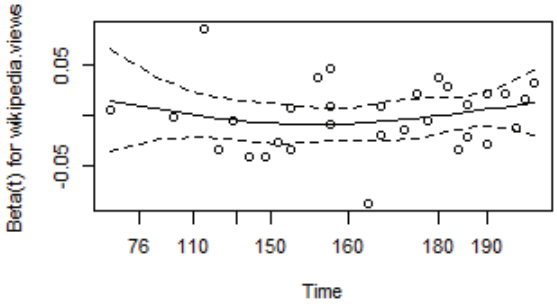
Global Proportion Hazard Assumption = 0.77

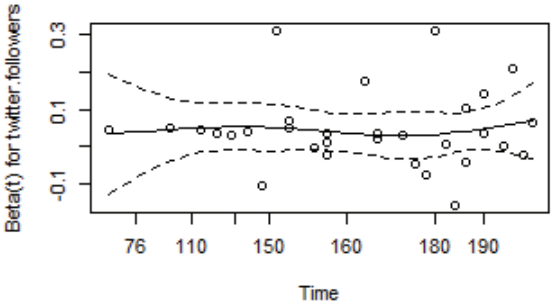
Variable	Coefficient	Exponential of Coefficient	P-Value	Proportional Hazard Assumption P-Value
Facebook Page Likes	-0.0061	0.994	0.76	0.27
Wikipedia Page Views	-0.0013	0.998	0.82	0.61
Twitter Followers	0.045	1.046	0.0087	0.96
Youtube Plays	0.00045	1.00045	0.0055	0.60
SoundCloud Plays	0.026	1.026	0.000036	0.72

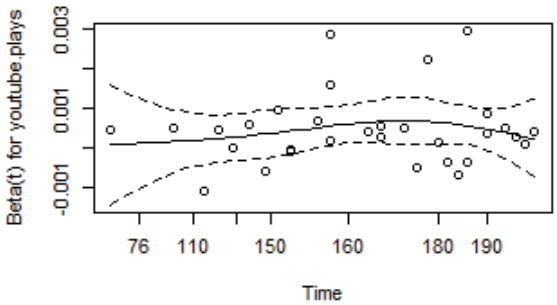


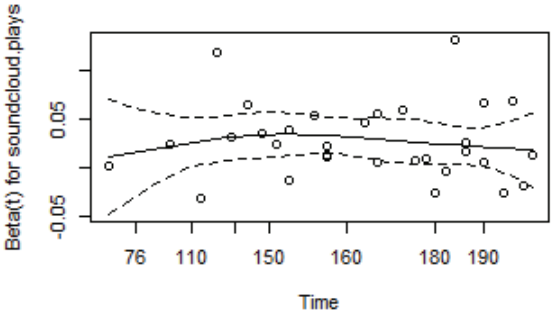
Diagnostic Plots of Proportional Hazard Assumption











Appendix D: All Next Big Sound Metrics

Next Big Sound Online Sources and Variables	
Facebook	Page Likes
	Page Views
	Unique Visitors
	Engaged Users
	Talked About This Today
Wikipedia	Views
Twitter	Followers
	Mentions
	Tweets
	Retweets
Youtube	Video Views
	Subscribers
	Comments
	Likes
	Unique Views
	Video Favorites
	Shares
	Minutes Watched
	Average View Duration
	Average View Percentage
Vevo	Views
SoundCloud	Plays
	Followers
	Comments
	Downloads
Last.fm	Plays
	Listeners
	Shouts
Instagram	Followers
	Comments
	Likes
	Photos
Tumblr	Original Posts
	Posts
	Notes on Original Post