Staying on Top: *Survival of Music on Spotify's Top Charts*

Andrew Soldini

May 25, 2018

Abstract

Music has been revolutionized by instant streaming services such as Spotify. An implication of such a service is that the release of music can now become a global phenomenon incredibly quickly. Because music is traveling farther faster than ever, we are interested in seeing what music stays atop the charts for the longest. In this paper we will perform survival analysis on songs' times before falling out of the Spotify top rankings, either at all or specifically on the Global charts. We will show that there is significance in the correlation of survival of songs on the charts with their regions of origin and initial release successes.

1 Introduction

1.1 Background and Literature Review

Spotify, by turning music into a subscription, economically matches the marginal cost of creating an addition play of a song with its users cost of playing the song, at zero. This is a profound change in the flexibility people have in what they listen to. Individuals using Spotify have responded to this change by listening to more artists, representing a more diverse selection of music [EP17]. With people listening to more diverse music, we are then interested in what people are actually listening to around the world. Spotify occurring entirely online, there is incredible data to observe this.

The specific dataset that has allowed us to take on the task of answering these questions is a the top charts data from Spotify for the year of 2017. This dataset, contributes the songs that appeared on the top 200 of the charts for 53 countries for every day of 2017. We are interested in trying to analyze trends between country of origin, popularity of songs, and the survival, or continued appearance of a song, on these top charts.

1.2 Questions for Analysis

The primary question that we sought out to answer in this paper were in determining how long songs remain in the top charts. The questions we asked can be stated as:

1. Which regions' music are most likely to *survive* in the single chart for the top 200 songs globally?

2. How does the region of origin contribute to determining *survival* on any of the top charts of all the countries?

2 Methods and Research Design

2.1 Survival Analysis

As the title of the paper suggests, *survival* is a key concept to the questions that we are asking. By survival, we are referring to the field of statistical methods often called Survival Analysis or Time To Event Analysis (TTE). TTE was developed to analyze time until death in medical studies without assuming any specific functional form. In our case, we are interested in the number of days from a song first appearing on the charts that it falls off of the charts entirely. Typically, the data about each person or, in this case, song includes an initial starting point and a follow up point. This follow up is not random. In our case, the follow up will either be the point in time when a song falls off of the chart(s) or the observation period has ended with the song still on the chart(s). The non-observation of the event still provides useful information and is included by TTE methods. These items are classified as *right censored* because they happen in time after, or to the right, of our study.

The Kaplan Meier Plot

Developed in 1958 by Kaplan and Meier, the Kaplan Meier Plot is the method of visualization for TTE analysis [KM58]. This plot provides an empirical method showing the probability of survival, or the *survival function*($S(t)$). The survival function is the basis of TTE analysis, represented by:

$$S(t) = P(T > t)$$

Where T is the random variable representing the time of the event occurs and t represents a given time. The function $S(t)$ thus represents the probability of survival at a given point of time. The Kaplan Meier Plot approximates this by comparing the number of occurrences of the event, d_i, to the number of individuals at risk, n_i, within given time intervals. Individuals at risk is the number of individuals that are still alive at a given point that have not yet reached their follow-up or the event occurs for in the interval. If the individual is censored in the time interval, they are not included in the at risk in that interval or subsequent ones. Mathematically the estimated survival function is then:

$$\hat{S}(t) = \prod_{t_i \leq t} \left[1 - \frac{d_i}{n_i}\right]$$

The survival is approximated by a product of these intervals because each successive interval represents the probability of having survived to that point and then also surviving that time period. All plots in this paper will be Kaplan Meier Plots.

The Cox Proportional Hazards Model

With the survival function, we then become interested in not only the probability of survival, but also the instantaneous rate of occurrence. This rate is represented

by what is called, the *hazard function*. Where $f(t)$ represents the density function of the occurrence of events over time the hazard function can be demonstrated by:

$$h(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)}$$

The intuition behind this is that the limit describes the probability that the event occurs in a infinitesimal window of time given that the individual has survived up until this point. Therefore the density represents the probability of occurrence and the survival function represents the probability of surviving up until and at that point.

The *Cox Proportional Hazards* model is the way in which we model covariates to the occurrence of the event. The Cox model is:

$$\log(h(t)) = \log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Where β_j and x_j are the coefficient and associated occurrences of covariate j. $h_0(t)$ represents the hazard in cases where the covariates are not present. Building from this, the single covariate model can then be represented by:

$$h_1(t) = h_0(t) \times e^{\beta_1 x_1}$$

We then can easily see the transformation of this to the *Hazard Ratio*:

$$HR(t) = \frac{h_1(t)}{h_0(t)} = e^{\beta_1}$$

The hazard ratio is a proportional effect that the covariate has on the hazard function overall, hence the title "proportional hazards model." The hazard ratio is the statistic by which we will interpret the correlations of covariates:

- A Hazard Ratio greater than one ($HR > 1$) imply hazard is increased in the presence of the covariate, implying a positive correlation
- A Hazard Ratio less than one ($HR < 1$) imply hazard is decreased in the presence of the covariate, implying a negative correlation

2.2 Setting Up the Model

Before we began to address the specific question at hand, we needed to first work the dataset to have have the needed properties. These components needed were the region of origin, maximum number of streams in an interval, time of follow up, and status at follow up. We will describe where this information was derived here so that you may understand the biases that we introduced in our results.

Region of origin was not originally in the dataset so we manually added it for roughly 800 songs. We only looked at songs that first appeared after the first day of observation to avoid left-censorship, this introduces some bias. Of those 800 songs, roughly half were picked randomly from the songs that appear in the data. The other half were picked specifically for being popular (highly ranked and streamed). The non-random selection was essential for answering the first question, but the random expanded the dataset for the second question.

When we say "Region," we are referring to a generalization of the area in the world that the song came from. There were six regions in all, representing varying numbers of countries and numbers of listeners and production of music, impacting their significance in the results. The regions and their number of included countries are as follows:

- North America: 2 (US, CA)
- Latin America: 15 (Including Mexico and South America)
- Europe: 26
- Asia: 9 (Notably not including China, India, Korea)
- Middle East: 1 (Turkey)
- Oceania: 2 (AU, NZ)

Because of a lack of a significant amount of songs in Oceania and the Middle East, we do not include them in the data analysis. Also, Africa is notably missing entirely.

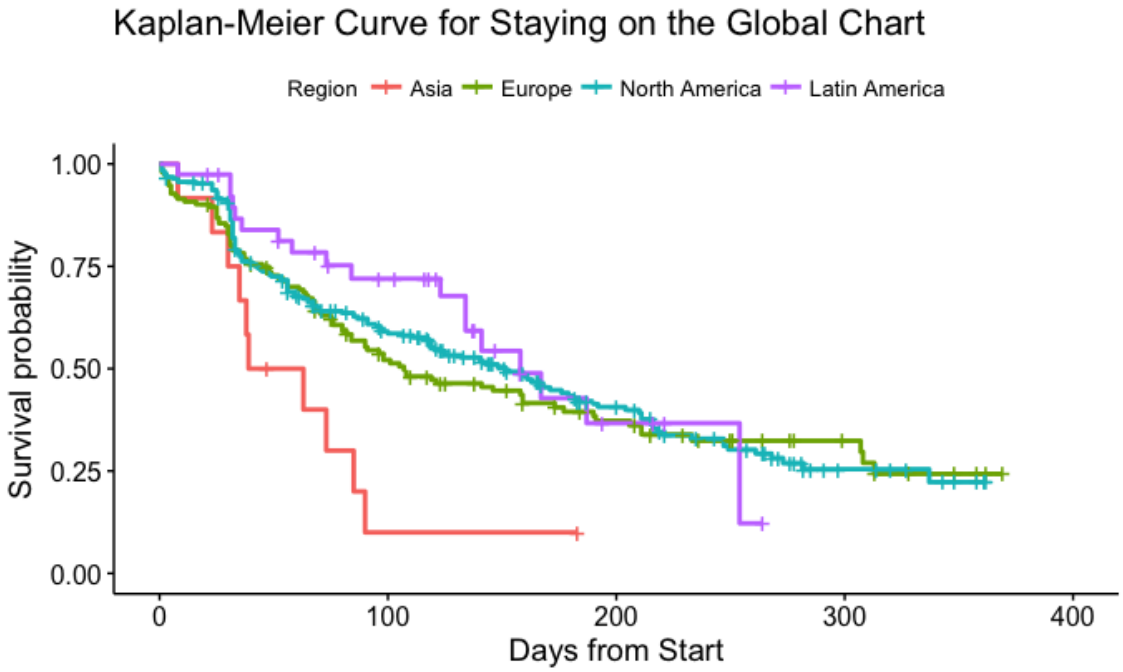
The number of streams on a song's first day was a covariate used in the analyses to try to mitigate the effects of a song's original position in the ranks effect to more isolate the country of origin's effect. To create this, we aggregated the number of streams of a song on its first day. We used the logarithm of this value in order to deal with its heavy right-skew.

The follow up time is the number of days after a song first appears on the chart after which it either is: no longer on the chart or censored. Because the time origin is relative, censorship takes place at different times. Also there is a fairly robust assumption we made here that songs do not fall off of the chart and go back on at a later point.

3 Analysis of Results

In this section we will present the Kaplan-Meier plots associated with each analysis question followed by a table of the Cox model hazard ratios. We will then give a very brief analysis of what this result is saying.

3.1 Global Charts

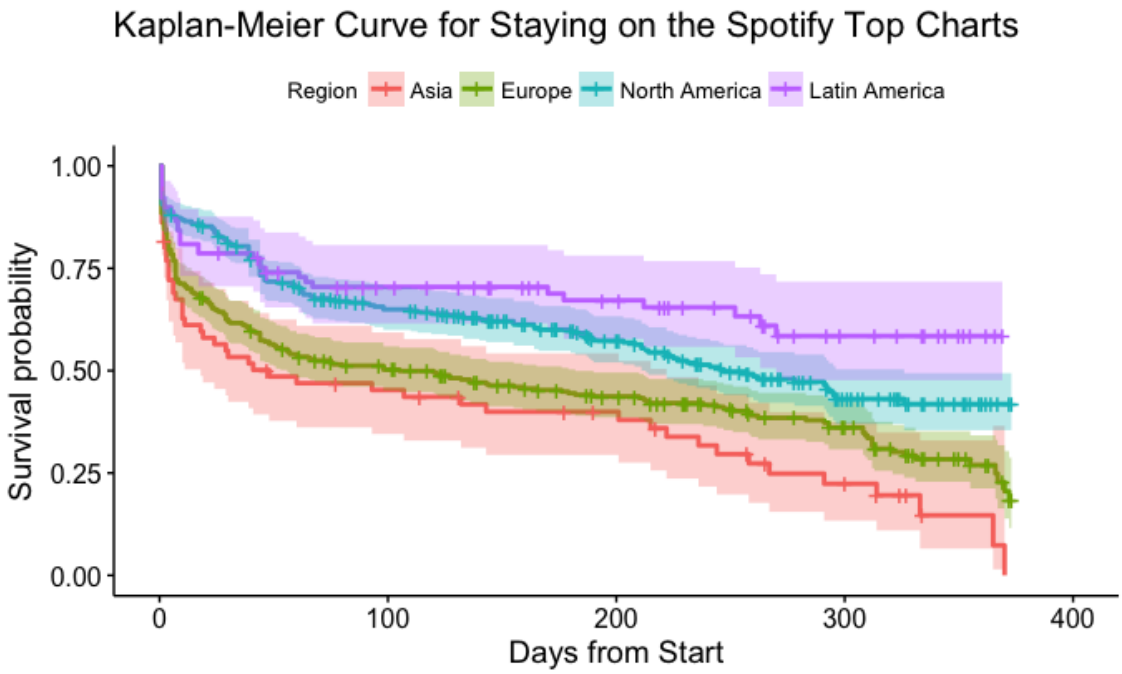


Predictor	Hazard Ratio	p-Value
Asia	default	default
Europe	0.5036	0.0418
North America	0.4842	0.0286
Latin & South Americas	0.3697	0.0112
Streams on First Day	0.8842	2.9e-06

Table 1: Survival in Global Charts by Region

In the graph above, we see that when Asian songs are in the global top 200 songs, they fall out much more quickly than the songs from the other regions. Looking at the hazard ratios from the cox model, this is accurately reflected by all regions having a lower hazard than Asia. Notably, Latin America has the best survival of all regions on the global rankings. North America, then Europe, and then Asia complete this ranking.

3.2 All Charts



Predictor	Hazard Ratio	p-Value
Asia	default	default
Europe	1.1667	0.34023
North America	0.8335	0.27596
Latin & South Americas	0.5053	0.00285
Streams on First Day	0.7769	$<$ 2e-16

Table 2: Regional Successes and Chart Survival

In the graph above, we see a clear stratification of the regions. Looking at the hazard ratios from the cox model, this is accurately reflected by all regions having a lower hazard than Asia. Again, Latin America has the best survival of all regions on the global rankings. North America, then Asia, and then Europe follow and complete this ranking. In this analysis, however, Latin America is the only region that is a statistically significant predictor.

4 Discussion

The main takeaway that we found to be in this analysis is that Latin America performs surprisingly well on the charts beyond the initial popularity of songs. In this analysis we can not really being to explain why this is but by the differences in the setup we can more clearly frame our intentions.

In the global charts, success has more to do with popularity around the world in order to stay in the rankings. Therefore the songs must appeal to the broadest audience possible. The success of Latin America here thus seems to point to broad global success of music from the region.

When looking at all of the charts, songs can be popular in just their home regions and remain in the surviving pool. Therefore longer success of songs in this analysis

could imply that some markets have slower turnover of music in general.

To understand more of what is going on here, there could be many other interesting variables to consider. Some specific ones that we would like to consider in further work would be genre of the song, language of the song, and perhaps components within the song. All of these might paint a more detailed picture of what makes songs popular, or at very least resilient on the top charts.

We are also limited by our small sample size that makes significance harder to determine and a further study would want to include more songs in the analysis. A further study on this dataset may also wish to try other methods such as machine learning or cluster analysis in order to make better use of the lines of data available.

Throughout our analysis we also made assumptions that Spotify itself does not influence the ways in which people listen to music on its platform significantly. We surmise that Spotify has a strong track on its effects on listening behavior and that there could be a small effect that they can look out for. Perhaps it is even possible for them to run causal experiments.

Overall, while we now have more questions than when we started, we have discovered interesting trends in the data that can be more deeply explored in further analyses.

References

[EP17] David Erlandsson and Jomar Perez. Listening diversity increases nearly 40 percent on spotify, 2017.

[KM58] Kaplan and Meier. Nonparametric estimation from incomplete observations. 1958.