Understanding Cultural Differences Using Spotify Data

Sam Laing
6283670
sam.laing@student.uni-tuebingen.de

Albert Catalan Tatjer
6443478
albert.catalan-tatjer@student.uni-tuebingen.de

Abstract

The main objective of this project is to examine the most streamed songs from a collection of different countries and investigate whether cultural similarities could be observed through analysis of song features obtained from the Spotify API. To achieve this, we apply hypothesis testing, random trees embeddings and a number of data visualization techniques.

1 Introduction

Music has been a fundamental aspect of human culture since the earliest days of our species [2]. It plays a crucial role in shaping cultural identity and has been the subject of ongoing research in fields such as anthropology and psychology. Similarly, the cultural evolution of wild bird songs has also been studied as a means of understanding the mechanisms driving such differences [4]. In today's globalized world, regional cultural differences in music continue to be an important topic of study, shedding light on both human history and current geopolitical relationships. The aim of this study is to investigate differences in international music using data from Spotify's weekly national ranking, and Spotify's API song feature data.

2 Data Discussion

2.1 Source

The data collected comes from two different, Spotify provided, sources. Spotify provides the 200 most listened to songs every week from 61 different countries, along with the number of streams of each song in that week. For pragmatic reasons, it was decided that just the top 50 songs every week for each country in the year 2022 would be collected. From here, we then scraped data from the Spotify API on each of the song urls we collected. For the purposes of our analysis, the following features were extracted:

- Length
- Danceability
- Acousticness
- Energy
- Liveness
- Loudness
- Speechiness
- Tempo

The following link provides a description of each of the collected features as defined by the Spotify for Developers teams.

2.2 Data Exploration

Before diving into any sort of modelling, it is prudent to consider some elementary visualizations of the data. Firstly, we considered a correlation matrix for the set of features on the unique set of

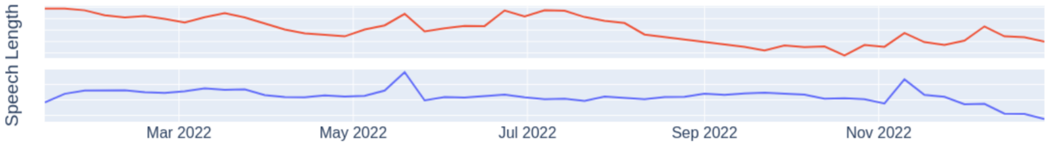

Figure 1: Weekly Mean of Two Features

songs. The largest correlation was -0.53. All other pairs have low correlation, indicating that there is minimal redundancy in these features.

By plotting the weekly means 1, It became clear that a number of these features were dependent on the chosen week, and some even appeared to exhibit seasonal behaviour.

3 Method

We considered two different approaches. Firstly, for each country, we investigate which features are significantly different from the global ranking. Secondly, we perform a Random Trees embedding of the weekly, per-country top 50 features into a two-dimensional space, where we can easily visualize and interpret the results.

3.1 Distance to Global

To investigate which countries deviate the most from the "average music taste", we compare the music of each country to the global average. For reasons discussed above, we conduct separate tests for each month. For each country, we have 200 data points per month. We can therefore safely assume that we have an accurate estimate of the variance of each feature for each country. For this reason, we perform a pairwise z-test for each feature at 0.05 significance level. Features such as "time signature" and "instrumentalness" were clearly not normally distributed, so they were omitted from the hypothesis tests, as z-tests assume a normal distribution [3].

3.2 International Distance

We are concerned with comparing the weekly ranking of the different countries. For each country, the weekly information consists of 50 songs per week and 7 features per song. We need to extract interpretable information from this high dimensional space. We consider every song in the weekly top 50 to be equally relevant for the country's cultural expression. Therefore, we compute the unweighted weekly mean of each feature, resulting in 7 features for each week and each country. Under these conditions, we consider a Random Trees [1] embedding to a two-dimensional space.

4 Results

4.1 Distance to Global

In the above heatmap 2, we illustrate the degree of deviation of each country from the global data, parametrized by the number of statistically significant differences.

Out of the $7 \times 12 = 84$ hypothesis tests performed for each country, Indonesia exhibited the largest number of statistically significant deviations, at 72. New Zealand and United Arab Emirates tied for the lowest number of statistically significant differences at 13. On average, countries in North America, Oceania and Europe (except for France and Spain) deviated less from global whereas countries in Asia, South America and Africa (except for South Korea and United Arab Emirates) exhibited more on average.

4.2 International Distance

The embedding mentioned in 3.2 is shown in figure 3. From here, we will refer to the representative point of a country, when addressing a country's position in the embedding space. Data points that

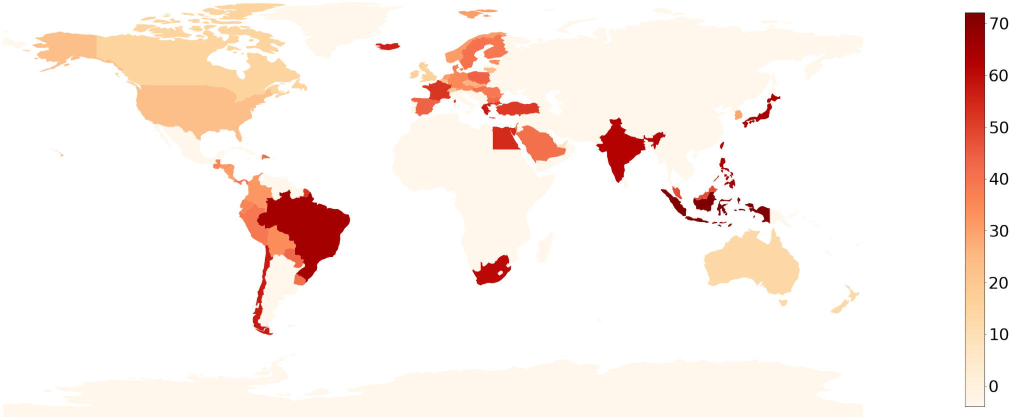

Figure 2: The heatmap representing the number of paired z-tests with a statistically significant difference (at $\alpha = 0.05$ significance level) from global. Darker shades of red correspond to numerous significant differences, and lighter shades correspond to fewer differences. Countries with no available data have been filled in with another shade for easier visualisation.

appear near to one another in the embedding space correspond to data points whose features, in the source space, are similar.

We start with the country groups that form proximity clusters. In the top-left corner, and far away from most of the countries, there are Indonesia and Iceland. In the top-centre, the south-east Asian countries Philippines, Hong-Kong, Taiwan and Malaysia are grouped together. Further down, there are middle eastern countries close together, Turkey, Saudi Arabia and Israel, interestingly India and Singapore appear in the same cluster. Right of centre, there is a dense area where the United States of America, Australia, New Zealand, Korea, United Arab Emirates, most of the European countries lie and Nicaragua. We interpret this cluster to be western-influenced countries. In the bottom right corner, there is the Spanish-speaking cluster, Spain, Ecuador, Bolivia, Costa Rica, Peru, Colombia, etc. are all close together. Additionally, Brazil, South Africa, and Japan appear in the left-centre of the embedding space, far away from most of the other countries. Finally, in the centre of the plot, we see that Egypt is the most isolated point. One possible explanation is a lack of data from African countries.

Interestingly, the distance of the countries to the global data point, visualized in 2, resembles that of the embedding space.

5 Conclusion

The fact that North American, Oceanic and European countries deviated less from global corroborates the idea that Spotify's global music rankings are strongly western-influenced. Although the Random Trees embedding proved to be quite effective in finding clusters of related countries, there are some rather unexpected results. Nicaragua belongs to the "western cluster" instead of the Spanish-speaking cluster, Japan, Brazil and South Africa are close together, and Iceland and Indonesia are also close together. This could be either a result of the information lost during the embedding, the fact that the mean of the top 50 songs could obscure some complexity of the data, or an unexpected actual cultural similarity between the countries that we are unaware of. It should be noted that PCA embedding yielded similar but less distinguished clusters.

When considering these results, it is important to be mindful of the following points:

The data is gathered from Spotify. This is substantial bias. It is a Swedish company with paid membership. The user base is, therefore, the subpopulation of the world willing, and able, to pay for it. Another issue is that Spotify does not operate in all the countries of the world. With consequences that are two-fold. Firstly, we lack data of a lot of countries that would have potentially changed the embedding function meaningfully. Secondly, there are artists from certain countries that cannot

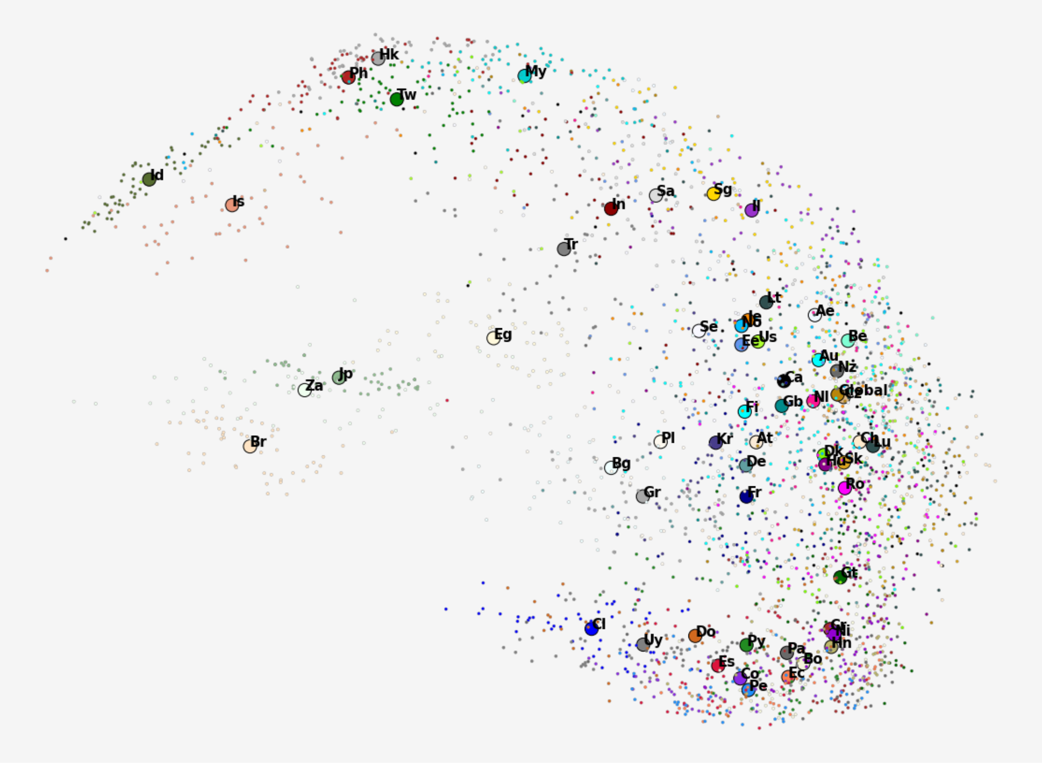

Figure 3: The Random Trees embedding space representation of the data. The source space is $\mathbb{R}^F$ where there are $C \times W$ data samples. $C = 62$ is the number of sampled countries, $W = 52$ is the number of weeks, and $F = 7$ is the mean of each song feature of a country for over a week. Note that there is a unique colour for every data point that belongs to the same country, and there are 52 points per country, one for each week of the year. For clarity, we plotted the mean of the 52 embedding points of each country to serve as a representative point.

post their music in the platform, thus biasing the song space users can choose from. Additionally, Spotify's recommendation algorithm has a huge impact on the rankings, and the data we acquired is conditional on its bias.

On the other hand, all the analysis done relies on the feature metrics devised by Spotify. These metrics can not possibly encapsulate all the qualitative nuances of music. In particular, music from Asia and Africa is often microtonal and does not lie within the twelve-tone equally tempered music which is almost exclusively seen in popular western music. Moreover, a lot of African and South American music feature intricate and subtle rhythmic distinctions which cannot be adequately described by one or two features such as "Danceability" or "Liveness". In short, the metrics used by Spotify are sure to be in some way culturally biased and too sparse to sufficiently parametrize such a diverse art form.

References

[1] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[2] A.P. Merriam and C. Bithell. *The Anthropology of Music*. Number S. 3-16. Northwestern University Press, 1964. URL https://books.google.de/books?id=30ghzQEACAAJ.

[3] Ajay Singh and Micah Masuku. Assumption and testing of normality for statistical analysis. 3: 169–175, 06 2021.

[4] Heather Williams and Robert F. Lachlan. Evidence for cumulative cultural evolution in bird song.