Musical Success in the United States and Brazil: Novel Datasets and Temporal Analyses

Gabriel P. Oliveira, Gabriel R. G. Barbosa, Bruna C. Melo, Juliana E. Botelho,
Mariana O. Silva, Danilo B. Seufitelli, Mirella M. Moro

Universidade Federal de Minas Gerais, Brazil
gabrielpoliveira@dcc.ufmg.br, grgb@ufmg.br,
{brunacamposmelo,juliana.botelho,mariana.santos,daniloboechat,mirella}@dcc.ufmg.br

Abstract. Music is not only a worldwide essential cultural industry but also one of the most dynamic. The increasing volume of complex music-related data defines new challenges and opportunities for extracting knowledge, benefiting not only different music segments but also the Music Information Retrieval research field. In this article, we assess musical success in the United States and Brazil, two of the biggest music markets in the world. We first introduce MUHSIC and MUHSIC-BR, two novel datasets with enhanced success information that combine chart-related data with acoustic metadata to describe the temporal evolution of musical careers. Then, we use such enriched and curated data to cluster artists according to their success level by considering their high-impact periods (hot streaks). Our results reveal three groups with distinct success behavior over time. Furthermore, Brazil and the US present specific music success patterns regarding artists and genres, reflecting the importance of analyzing regional markets individually.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: Time series analysis; H.5.5 [**Sound and Music Computing**]: Modeling; H.3.3 [**Information Search and Retrieval**]: Retrieval Models

Keywords: hot streaks, hit song science, music information retrieval, open data, time series analysis

1. INTRODUCTION

Music is not only a worldwide essential cultural industry but also one of the most dynamic. Such a powerful domain has a long relationship with Computer Science, as it has helped the music industry over a myriad of different problems. Specifically, Music Information Retrieval (MIR) is an emergent research field that combines musicology and computing techniques to extract knowledge from music-related content. In fact, much has already been done in MIR since the seminal position paper by Byrd and Crawford [2002], from automatically classifying music genres to tackling user gender bias in recommendation algorithms [Melchiorre et al. 2021]. Usually, the first step for all such tasks is gathering data regarding song characteristics and relevant metadata on the music ecosystem (i.e., success charts, artist and genre metadata). However, the lack of music data sources with unified information requires the execution of data collection and data integration steps, which is still challenging.

With more information available over the Web, solving such challenges inevitably faces huge data volumes, which brings additional intrinsic processing issues. Nonetheless, MIR tasks have much to benefit from online data sources such as Billboard and Spotify, as they provide distinct information related to the music ecosystem. For instance, while Billboard offers chart-based success features, Spotify API provides acoustic fingerprints on the songs and artist and genre metadata. This meaningful set of characteristics is paramount to Hit Song Science (HSS), an emerging field within MIR that aims at predicting a song's popularity from its features [Çimen and Kayis 2021; Pachet 2011].

The authors would like to thank Prof. Anisio Lacerda for all relevant insights to model and build our datasets. This work was supported by CAPES, CNPq, and FAPEMIG, Brazil.

Table I: Comparison of datasets with popularity data.

Year	Dataset	Size	Songs	Charts	Artists	Genres	Time Series
2011	MSD [Bertin-Mahieux et al. 2011]	1,000,000	✓	×	✓	×	×
2016	TPD [Karydis et al. 2016]	23,385	✓	✓	✓	×	×
2018	MuMu [Oramas et al. 2018]	147,295	✓	×	×	✓	×
2019	HSPD [Zangerle et al. 2019]	1,000,000	✓	✓	×	✓	×
2019	SPD [Cosimato et al. 2019]	101,939	✓	×	✓	✓	×
2019	ABGD [Bogdanov et al. 2019]	1,935,991	✓	×	×	✓	×
2019	MusicOSet [Silva et al. 2019]	20,405	✓	✓	✓	✓	×
2020	MGD [Oliveira et al. 2020]	13,880	✓	✓	✓	✓	×
2021	Unnamed [Bertoni and Lemos 2021]	881	✓	×	×	×	×
2021	LFM-BeyMS [Kowald et al. 2021]	1,084,922	✓	×	✓	✓	×
2021	**MUHSIC (This work)**	**22,635**	✓	✓	✓	✓	✓
2021	**MUHSIC-BR (This work)**	**2,595***	✓	✓	✓	✓	✓

* Digital Era only.

Besides predicting song success, music-related data are also relevant to AI-based talent identification, mainly through popularity peaks within artist careers. Such peaks are commonly grouped into hot streak periods, defined as the continuous periods of success or productivity above normal [Garimella and West 2019; Janosov et al. 2020; Liu et al. 2018; Liu et al. 2021b]. Indeed, identifying Hot Streaks is very useful for the music industry, as it is one way of investing in the right artist at the most relevant moment. Although the existing music-related datasets provide partial information to perform such a task [Cosimato et al. 2019; Silva et al. 2019], to the best of our knowledge, none of them contains temporal success information already processed for analysis of this type.

Here, we introduce MUHSIC (***Mu**sic-oriented **H**ot **S**treak **I**nformation **C**ollection*), an open dataset with temporal information on musical success in the United States, focusing on artist and genre careers. Specifically, we provide chart-based success time series from 1958 to 2020, as well as the hot streak periods detected in each time series. Besides, the dataset also contains metadata about the most relevant music elements, i.e., songs, artists and genres. This novel set of features and its ease usability and reproducibility make MUHSIC a valuable resource for different MIR applications, such as Hit Song Science and Music Genre Classification. Moreover, the success time series may be used for analyzing temporal evolution and identifying success trends in the music industry.

This article extends a paper from the Dataset Showcase Workshop of the 36^{th} Brazilian Symposium on Databases [Oliveira et al. 2021b]. Its main contribution is an extension of MUHSIC with success data from Brazil, called MUHSIC-BR. We use such novel data to build artists' success time series and to detect Hot Streak periods in the Physical (1990-2015) and Digital (2016–2020) Eras. We also present a cluster analysis for artists over both datasets as an example of real application, which allows to better visualize different success levels achieved by artists. Overall, our datasets and analyses represent a step further in understanding the patterns that govern success within the musical industry.

2. EXISTING MUSIC DATASETS

There are different datasets with information about music, including metadata, acoustic features, lyrics and popularity data. The subset of information varies according to the work purpose. For instance, recent music datasets focus on disentanglement learning [Pati et al. 2020], instrument classification [Castel-Branco et al. 2021] and playlist generation [Ferraro et al. 2021]. More data available enable more mining tasks over such data. Table I shows the top datasets that include popularity data (except MSD), information that is crucial to evaluate success.

Despite not having popularity data, the Million Song Dataset (MSD) [Bertin-Mahieux et al. 2011] is often used in MIR, as it contains audio features and metadata for one million music tracks, with over 280 GB of data. It is also criticized due to its lack of details on data extraction and integration. The Hit Song Prediction Dataset (HSPD) [Zangerle et al. 2019] was built upon MSD by including data about its tracks that were in the Billboard Hot 100 charts. Then, the Track Popularity Dataset (TPD) [Karydis et al. 2016] provides data on musical track popularity by considering different sources from 2004 to 2014. It contains features tailored for music information retrieval (e.g., identification spaces and contextual similarity) and considers both popular and non-popular audio tracks.

In a different perspective, MusicOSet [Silva et al. 2019] and SpotGenTrack Popularity Dataset (SPD) [Cosimato et al. 2019] focus on quality and provide metadata, lyrics, acoustic features and song popularity. MusicOSet content enables large-scale evaluations of song and music collaboration-based recommendations, whereas SPD is tailored for other MIR tasks, such as genre classification and auto-tagging. Then, the Music Genre Database (MGD) [Oliveira et al. 2020] expands on such possibilities by providing information on music genres: genre collaboration networks and genre mapping.

Indeed, genre-related tasks are also relevant within MIR, with many datasets to assess them. One task is *genre classification*, which assigns one (or more) musical genre to a song. For such a task, MuMu [Oramas et al. 2018] is a multimodal music dataset that combines audio, images, text and multi-label genre annotations. It is also based on MSD and includes information from Amazon reviews. Likewise, Bogdanov et al. [2019] propose the AcousticBrainz Genre Dataset (ABGD), which combines different music data sources (including MSD) to provide hierarchical multi-label genre annotations. Then, Kowald et al [2021] use genres in their LFM-BeyMS dataset to analyze user behavior in LastFM.

Music-related datasets are the base for several real-world applications within the MIR field. For instance, Al-Beitawi et al. [2020] use K-Means in acoustic features from Spotify to uncover structural patterns related to the songs' popularity. Also, Roy et al. [2020] use cluster analysis to assess musical note structures in Indian classical music. Clustering people is also possible, as done by Georges and Nguyen [2019] to visualize similarities among European music composers.

Regardless of purpose, most music datasets with popularity data contain only (or primarily) information from the United States. Previous studies indicate each music market has its own behavior [Oliveira et al. 2020] and, therefore, success analyses for such markets need specific regional data. For the Brazilian market, which we analyze in this article, Bertoni and Lemos [2021] make the first step by providing three datasets with acoustic features for hit and non-hit songs. However, they still lack information on artists and music genres to allow further analyses.

Overall, our datasets MUHSIC and MUHSIC-BR naturally share content with their predecessors; but they also tackle temporal success through *time series modeling* and *hot streaks detection*, which are their most complex and important novelties. Moreover, aiming to embrace several music data mining tasks, they include information on temporal success of both artists and their genres.

3. MUHSIC DATASET

In this section, we present MUHSIC itself, an open dataset focused on temporal success information. We first describe its building approach, from data collection to processing (Section 3.1). Then, we overview its main data (Section 3.2) and perform a short exploratory analysis over it (Section 3.3).

3.1 Building Methodology

MUHSIC building process has five steps as shown in Figure 1 and detailed next: collecting music charts for temporal success data; collecting song and artist metadata from Spotify; integrating all data; modeling musical careers in success time series; and generating hot streak information.

Success Chart Collection. Billboard is an American-based music magazine (with operations in Canada, Brazil, Greece, Japan, South Korea and Russia) widely known for its exclusive charts on trends across all musical genres. It publishes the *Hot 100* Chart,[1] which is also the main all-genre

[1] Billboard Hot 100 Chart: `https://www.billboard.com/charts/hot-100`

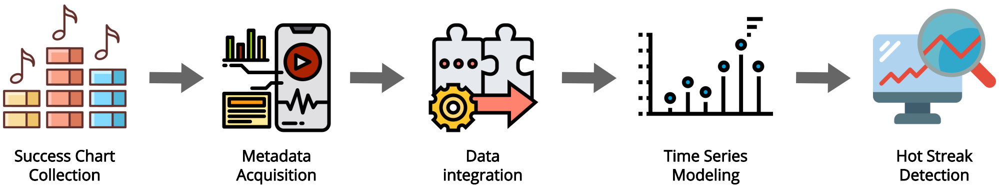


Fig. 1: MUHSIC building methodology.

song ranking in the country and possibly the world, as according to IFPI,[2] the United States is the biggest music market worldwide. Billboard Hot 100 has been weekly published since 1958, and is currently built by summing up songs' sales, airplay and streaming consumption. To model artist success over time, we collect all Hot 100 charts from August 11, 1958 to August 22, 2020 (data collection time) by using the Python package *billboard.py*.[3] Each chart is composed of its 100 entries, ranked from the most popular song to the least popular on that week.

Metadata Acquisition. Billboard chart entries are composed only of song name and its artists, which is not enough for complex research questions. Hence, we improve the dataset from the previous step by collecting data from Spotify, the world's most popular audio streaming service with more than 365 million users in 178 markets (as of August 2021). Its API[4] provides information on artists and songs, including artist genres and debut date, acoustic features for each song (e.g., key, mode, energy).

Data Integration. Building a dataset from multiple sources needs data integration; i.e., MUHSIC requires to link all songs from the Hot 100 entries to their correspondent Spotify song records. We then use the *SequenceMatcher* class from the Python *difflib*[5] package, *Jaro-Winkler* from the *python-string-similarity*[6] package and four similarity functions (**ratio, partial_ratio, token_sort_ratio, WRatio**) from the FuzzyWuzzy[7] package. A match happens when the similarity between records is at least 0.9; i.e., some Hot 100 entries have no match on Spotify, due to distinct song/artist name spelling or the unavailability of songs in Spotify. Overall, around 85% of all collected songs are successfully matched.

Time Series Modeling. Music success evolves by following the audience tastes, worldwide trends and other factors such as media platforms dynamics, new music styles and new song releases. Here, we model success over time based on the Hot 100 charts and Spotify data for both artist and genres, by including aggregated acoustic features of the songs that appear in a given week (e.g., the number of explicit songs and the median acousticness) and success information as follows.

For each **artist**, we build a time series from the debut date (i.e., first release date on Spotify) to the last chart collected. Thus, time series points represents the success of its artist in a given week, according to the Hot 100 chart. An artist's success is given by the *rank scores* for all of their songs that appear on the week chart. The *rank_ score* of a song i is $rank\_score(i) = max\_rank - rank(i) + 1$, where *max_ rank* is the lowest possible rank (i.e., 100), and *rank(i)* is the song position on the chart. Then, the rank scores of an artist must be aggregated somehow. The easiest idea is to sum up all rank scores of the songs. However, that is not enough because an artist with the #1 song in one week is more successful than an artist with two songs in the middle of the chart (e.g., on positions #49 and #50). Hence, rank score aggregation uses the Discounted Cumulative Gain (DCG) [Aggarwal 2016], which emphasizes the most relevant records (i.e., the highest ranked songs on the chart) and penalizes songs that appear in lower positions by a logarithmic factor, as defined by the following equation.

[2] IFPI Global Music Report: `https://gmr.ifpi.org/`
[3] billboard.py: `https://github.com/guoguo12/billboard-charts`
[4] Spotify Developer API: `https://developer.spotify.com/`
[5] *difflib*: `http://docs.python.org/3.6/library/difflib.html`
[6] *python-string-similarity*: `http://github.com/luozhouyang/python-string-similarity`
[7] *fuzzywuzzy*: `http://github.com/seatgeek/fuzzywuzzy`

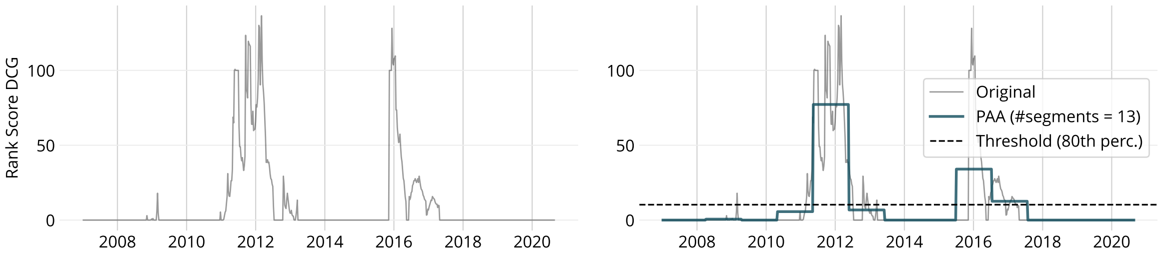


Fig. 2: Adele success time series (2007–2020). MUHSIC contains both the raw success time series (left) and the PAA fit with the hot streak threshold (right).

$$DCG = \sum_{i=1}^{n} \frac{rank\_score(i)}{log_2(i+1)}$$

We also build time series for analyzing the evolution of **genre** success. First, we assign artists' genres to their songs, as the songs themselves do not have such information. Then, for each week, we use DCG to aggregate songs from artists belonging to a given genre that appear on that week chart.

Hot Streak Detection. Professional careers tend to have phases of high productivity, reaching the career peak. Hot streaks (HS) is the term commonly used for continuous periods of success above normal. Previous work reveals that hot streaks can arise at any time in a professional career [Garimella and West 2019; Liu et al. 2018; Liu et al. 2021b]. In music, when an artist is at a hot streak, such an artist is also at the most profitable moment of a career. Therefore, we identify hot streak periods in both artist and genre time series (i.e., their careers) to allow further success analyses.

From raw time series, we apply Piecewise Aggregate Approximation (PAA) to reduce their dimensionality [Keogh and Pazzani 2000]. We use such a method to aggregate weeks into periods within careers, since highly impactful periods within artists' careers may contain weeks with low values for the success metric. In short, PAA reduces a given time series (with n points) into a new series with N segments, $1 \leq N \leq n$. Their values are the average of the points within such frames; i.e., the approximation of each time series point is made by assigning the PAA value of its corresponding segment. To compare careers from different artists/genres, we define a unique size of 52 weeks (i.e., one year) for each segment. Hence, the number of segments is calculated by dividing the time series length by this predefined size. Then, we define hot streaks as the periods when the success metric (approximated by PAA) is higher than a predefined threshold. The threshold value is specific for each time series and is calculated from the *activity rate* (AR), which is the ratio between the number of weeks in which the artist/genre appears on Hot 100 and the total number of weeks of the time series:

AR $\geq$ 20%: threshold is the *80th percentile* of the success metric;
15% $\leq$ AR $<$ 20%: threshold is the *85th percentile* of the success metric;
10% $\leq$ AR $<$ 15%: threshold is the *90th percentile* of the success metric; and
AR $<$ 10%: threshold is the *95th percentile* of the success metric.

Figure 2 illustrates Adele's career, as an example of the time series available in MUHSIC. From her success timeline measured by DCG (left), we detect three hot streak periods (right). Therefore, our dataset provides all such data for further success analyses.

3.2 Data Content

We organize MUHSIC in a relational schema. The dataset consists of 11 tables that contains the information collected, curated and enriched, as designed in Figure 3. Such tables may be classified into five categories: charts, songs, artists, genres and associative tables. The first four categories represent the main elements of the musical ecosystem, and the associative ones connect them.

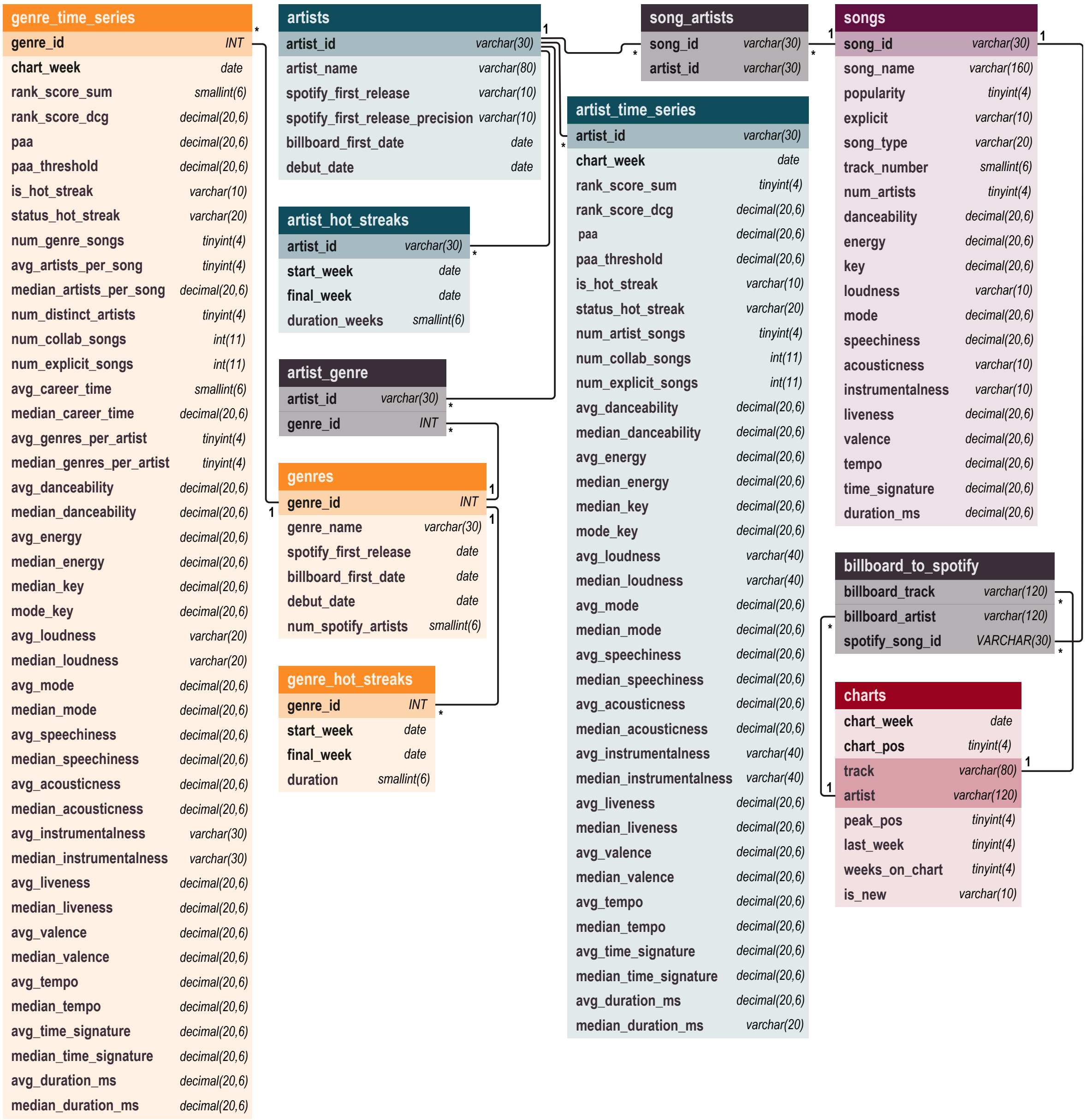

genre_time_series
genre_id INT
chart_week date
rank_score_sum smallint(6)
rank_score_dcg decimal(20,6)
paa decimal(20,6)
paa_threshold decimal(20,6)
is_hot_streak varchar(10)
status_hot_streak varchar(20)
num_genre_songs tinyint(4)
avg_artists_per_song tinyint(4)
median_artists_per_song decimal(20,6)
num_distinct_artists tinyint(4)
num_collab_songs int(11)
num_explicit_songs int(11)
avg_career_time smallint(6)
median_career_time decimal(20,6)
avg_genres_per_artist tinyint(4)
median_genres_per_artist tinyint(4)
avg_danceability decimal(20,6)
median_danceability decimal(20,6)
avg_energy decimal(20,6)
median_energy decimal(20,6)
median_key decimal(20,6)
mode_key decimal(20,6)
avg_loudness varchar(20)
median_loudness varchar(20)
avg_mode decimal(20,6)
median_mode decimal(20,6)
avg_speechiness decimal(20,6)
median_speechiness decimal(20,6)
avg_acousticness decimal(20,6)
median_acousticness decimal(20,6)
avg_instrumentalness varchar(30)
median_instrumentalness varchar(30)
avg_liveness decimal(20,6)
median_liveness decimal(20,6)
avg_valence decimal(20,6)
median_valence decimal(20,6)
avg_tempo decimal(20,6)
median_tempo decimal(20,6)
avg_time_signature decimal(20,6)
median_time_signature decimal(20,6)
avg_duration_ms decimal(20,6)
median_duration_ms decimal(20,6)
artists
artist_id varchar(30)
artist_name varchar(80)
spotify_first_release varchar(10)
spotify_first_release_precision varchar(10)
billboard_first_date date
debut_date date
artist_hot_streaks
artist_id varchar(30)
start_week date
final_week date
duration_weeks smallint(6)
artist_genre
artist_id varchar(30)
genre_id INT
genres
genre_id INT
genre_name varchar(30)
spotify_first_release date
billboard_first_date date
debut_date date
num_spotify_artists smallint(6)
genre_hot_streaks
genre_id INT
start_week date
final_week date
duration smallint(6)
song_artists
song_id varchar(30)
artist_id varchar(30)
artist_time_series
artist_id varchar(30)
chart_week date
rank_score_sum tinyint(4)
rank_score_dcg decimal(20,6)
paa decimal(20,6)
paa_threshold decimal(20,6)
is_hot_streak varchar(10)
status_hot_streak varchar(20)
num_artist_songs tinyint(4)
num_collab_songs int(11)
num_explicit_songs int(11)
avg_danceability decimal(20,6)
median_danceability decimal(20,6)
avg_energy decimal(20,6)
median_energy decimal(20,6)
median_key decimal(20,6)
mode_key decimal(20,6)
avg_loudness varchar(40)
median_loudness varchar(40)
avg_mode decimal(20,6)
median_mode decimal(20,6)
avg_speechiness decimal(20,6)
median_speechiness decimal(20,6)
avg_acousticness decimal(20,6)
median_acousticness decimal(20,6)
avg_instrumentalness varchar(40)
median_instrumentalness varchar(40)
avg_liveness decimal(20,6)
median_liveness decimal(20,6)
avg_valence decimal(20,6)
median_valence decimal(20,6)
avg_tempo decimal(20,6)
median_tempo decimal(20,6)
avg_time_signature decimal(20,6)
median_time_signature decimal(20,6)
avg_duration_ms decimal(20,6)
median_duration_ms varchar(20)
songs
song_id varchar(30)
song_name varchar(160)
popularity tinyint(4)
explicit varchar(10)
song_type varchar(20)
track_number smallint(6)
num_artists tinyint(4)
danceability decimal(20,6)
energy decimal(20,6)
key decimal(20,6)
loudness varchar(10)
mode decimal(20,6)
speechiness decimal(20,6)
acousticness varchar(10)
instrumentalness varchar(10)
liveness decimal(20,6)
valence decimal(20,6)
tempo decimal(20,6)
time_signature decimal(20,6)
duration_ms decimal(20,6)
billboard_to_spotify
billboard_track varchar(120)
billboard_artist varchar(120)
spotify_song_id VARCHAR(30)
charts
chart_week date
chart_pos tinyint(4)
track varchar(80)
artist varchar(120)
peak_pos tinyint(4)
last_week tinyint(4)
weeks_on_chart tinyint(4)
is_new varchar(10)

Fig. 3: MUHSIC relational schema, in which primary keys have a bold color background. The original PDF file can be downloaded in Zenodo (see Section 5).

Charts. Music charts are the core of MUHSIC, as success metric and artists are all derived from them. All Hot 100 charts are in the *charts* table that contains ranking position, track and artist names, which are used in the Spotify matching process. Other chart-related information extracted from Billboard includes peak position, number of weeks on the chart and previous week's position.

Songs. The table *songs* contains information of hit songs from Billboard Hot 100. All songs in the table were obtained from the Spotify matching explained in the previous section. Each record has data such as the *explicit* flag, which indicates whether the song lyrics contain language unsuitable for children. In addition, the table has acoustic features that describe aspects of the audio content, such as key, tempo and *acousticness* (i.e., the probability of a song being acoustic).

Artists. This category contains three core tables with data about the artists themselves, their success

Table II: MUHSIC basic statistics: number of records and file size.

Table	Records	Size	Table	Records	Size
artists	6,066	1.5 MB	genres	998	80 KB
artist_genre	21,123	4.0 MB	genre_hot_streaks	2,248	128 KB
artist_hot_streaks	7,557	1.5 MB	genre_time_series	2,067,770	1.0 GB
artist_time_series	6,251,441	3.9 GB	songs	22,635	8.5 MB
billboard_to_spotify	26,919	5.5 MB	song_artists	26,790	5.0 MB
charts	322,380	29.6 MB			

time series and their hot streak data. Specifically, table *artists* comprises the metadata obtained from Spotify and Billboard, such as their first release date (*spotify_first_ release*) and their first entry date at Billboard (*billboard_first_ date*). We consider the artist *debut_ date* as the minimum between those two dates. Next, table *artist_ time_ series* describes the careers of all artists. For each artist and weekly chart, we calculate the aggregated rank scores in two ways (sum and DCG), the PAA approximation and the threshold used to define hot streak periods. The field *is_ hot_ streak* informs whether the week belongs to a hot streak or not. Besides, we provide the aggregated acoustic features for all songs that appear in that weekly chart. Finally, the table *artist_ hot_ streaks* summarizes the hot streak periods, with the period of each one, as well as its duration in weeks.

Genres. The genre category is similar to the artist one with three tables, as we also build genre time series and detect their hot streaks. Table *genres* contains the metadata from Spotify, including the number of artists belonging to a specific genre (*num_ spotify_ artists*). Table *genre_ time_ series* comprises the genre success time series, and most of the columns are similar to the corresponding artist table. However, we added new features that are possible due to this genre perspective. For instance, we calculate the number of distinct artists of the genre for each week (*num_ distinct_ artists*), as well as the average and median career time of such artists (*avg_ career_ time and median_ career_ time*), in which career time is based on the number of days between debut and current week. Then, table *genre_ hot_ streaks* summarizes the hot streak periods for the genre careers.

Associative Tables. The associative tables represent many-to-many relationships and link tables from the other categories. For example, table *billboard_ to_ spotify* relates the Hot 100 entries to their Spotify correspondent instances. This relationship is made using columns *track* and *artist* from table *charts* and the *song_ id* from table Spotify *songs*. Also, tables *song_ artists* and *artist_ genre* represent the list of artists who sing a hit song and the music genres from a given artist, respectively.

3.3 Exploratory Data Analysis

We perform an exploratory data analysis over MUHSIC and present basic statistics to understand the data. Its final version is composed of: 3,238 weekly charts; 22,635 distinct songs; and 6,066 artists who belong to 998 music genres. Such enriched, curated data enable to build success-based time series to explore artists' careers and genres' evolution. Table II summarizes basics statistics for each table.

One key feature of MUHSIC is the artists' music genre. Figure 4 presents its 20 most frequent genres, according to the number of artists – in Spotify, the genre is linked directly to the artist, not to each song. We highlight the presence of widely popular super-genres such as *rock*, *pop*, *rap*, *r&b*, *soul* and *country*. There is no standardization in the Spotify genres; hence, there are several derived genres that are also frequent, including *dance pop, brill building pop* and *southern hip hop*. We use all such information to generate the genre success time series and to detect their hot streaks.

Next, we analyze the hot streak periods for artists through a simple characterization analysis of table *artist_ hot_ streaks*. Figure 5a shows the number of hot streaks (HS) per artists. In general, most artists (about 90%) have between one and two hot streak periods in their careers. Only a few manage to achieve more than five periods. Examples of such artists include Bruce Springsteen (8 HS), Michael Jackson (8 HS) and Mariah Carey (7 HS). Besides, most hot streaks last for around one year

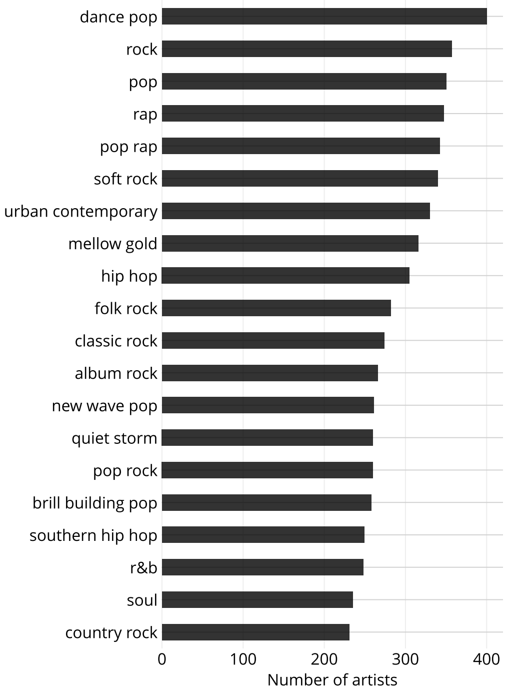
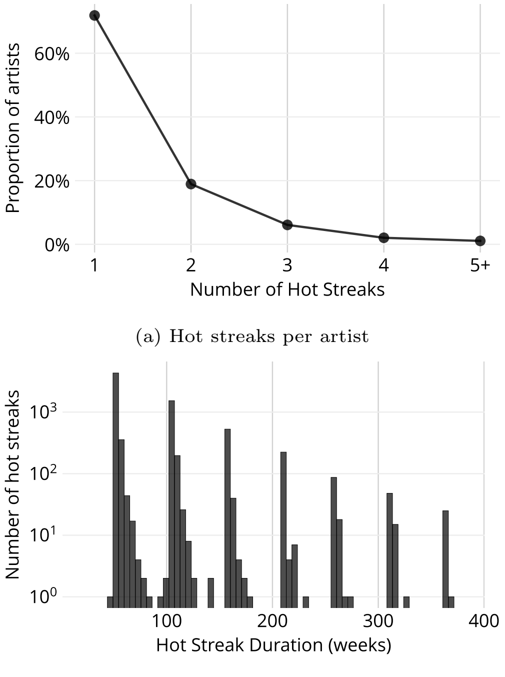


Fig. 4: Top 20 music genres in MUHSIC.



(a) Hot streaks per artist

(b) Duration of hot streaks

Fig. 5: Hot streak statistics in MUHSIC.

(52 weeks), as shown by Figure 5b. There are also many HS periods during two, three and four years, suggesting a yearly pattern in this phenomenon. Note the log scale in the y-axis.

4. MUHSIC-BR DATASET

MUHSIC is a powerful tool to assess music success throughout the years. Still, a limitation is to consider only data from the United States (i.e., Billboard Hot 100). Now, we make a first step for solving such an issue by presenting MUHSIC-BR, an extension of the original dataset that considers success data from Brazil. We also assess the evolution of musical careers by adding data from physical (1990–2015) and digital (2016–2020) eras. Following the organization of Section 3, we present methodology (Section 4.1), dataset content (Section 4.2) and exploratory data analysis (Section 4.3).

4.1 Building Methodology

To build MUHSIC-BR, we follow three steps for both eras: collecting temporal success data; modeling musical careers in success time series; and generating hot streak information from such careers.

Success Data Collection. For the Physical Era, we gather data from Pró-Música Brasil (PMB), the official representative of Brazilian record labels that represents artists (legally and financially) and issues certification awards. Such certification awards are given according to sale numbers in the form of "special discs", i.e., Gold, Platinum and Diamond discs. We collected the awards data available in the PMB website[8] in February 5th, 2021. We also gather the information if an artist is Brazilian or not, once the threshold sales number for each certificate depends on such information. We obtain it from Wikipedia using a Python lib.[9] In summary, we collect 3,243 musical works from 574 artists

[8] PMB Certificates: `https://bit.ly/CertificatesPMB`

[9] Wikipedia Python Library: `https://github.com/goldsmith/Wikipedia`

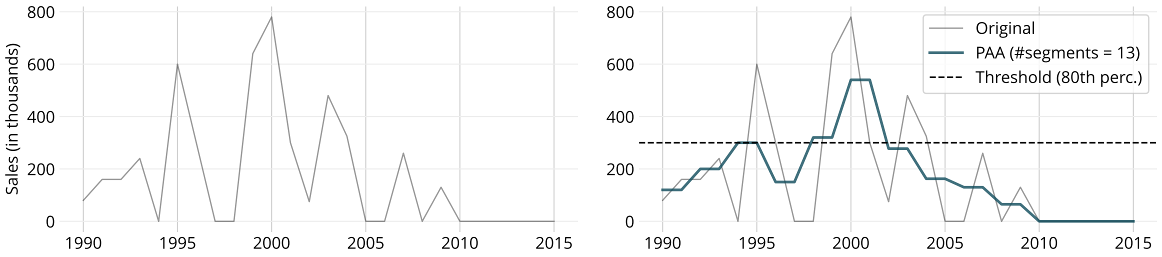


Fig. 6: Roberto Carlos' success time series in the Physical Era (1990–2015). MUHSIC-BR contains both the raw success time series (left) and the PAA fit with the hot streak threshold (right).

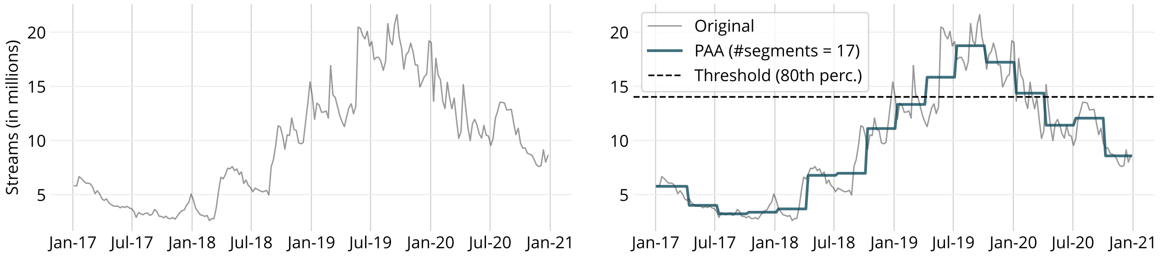


Fig. 7: Marília Mendonça's success time series in the Digital Era (2017–2020). MUHSIC-BR contains both the raw success time series (left) and the PAA fit with hot streaks (right).

considering only the period between 1990 and 2015 (i.e., Physical Era). Finally, the artists' genre metadata were collected from Spotify and integrated into MUHSIC-BR as the process in Section 3.1.

A key update occurred between 2016 and 2017 in PMB's metric: it changed from physical media (CD and DVD) towards digital media (Single and Albums). Meanwhile, streaming was already the primary revenue source for digital media in the global market, with around 58.5% of the total revenue.[10] Hence, we select streaming services as a data source for the Digital Era. We collect the weekly Spotify Top 200 with the most streamed songs in Brazil from January 2017 to December 2020. Each chart entry contains song name and artist(s), number of streams, Spotify URL, and its chart position. We also collect artist data such as name, number of followers and their list of genres through the Spotify API. Our dataset for Digital Era comprises 2,595 songs from 1,018 artists obtained from 108 weekly charts.

Time Series Modeling. For Physical Era, we model artists' time series with the total sales per year, which is equivalent to the aggregate of received certificates from each artist. As shown in Figure 6 (left), the artists' time series covers the whole physical Era (1990-2015). For Digital Era, the time series reflects the total streams per week for the whole Era (2016-2020) as shown in Figure 7 (left). Unlike MUHSIC, we do not use the rank score obtained from Spotify charts (Digital Era) because we believe that the total of streams is a more accurate success indicator for artists.

Hot Streak Detection. This part is similar to Section 3.1. We run empirical experiments to define the PAA segment size for both Physical and Digital Eras. The minimum segment size for Physical Era is two years, because it is a yearly timeseries. The Digital Era segment size is 12 weeks, which is reasonable to analyze the continuous periods of great success in streaming platforms such as Spotify. Figure 6 shows the time series and respective PAA of Roberto Carlos as an example of the Physical Era; then, Figure 7 shows the same information for Marilia Mendonça[11] as an example of the Digital

[10] Pró-Musica Brasil 2016: `https://bit.ly/ProMusica2016`

[11] Marília Mendonça was a young singer-composer and strong representative of the *feminejo* – the women movement within the male dominated "sertanejo" genre. She passed away on November 5h, 2021, while we were writing this article and had already used her as good example of success in Brazilian music.

Table III: MUHSIC-BR basic statistics: number of records, size and file description.

File name.csv	Records	Size	Description
physical_artist_hot_streak_summary	772	58.0 KB	Summary of each hot streak period achieved by the artists in the Physical Era (i.e., duration, start and end year)
physical_artist_timeseries	20,280	817.4 KB	Detailed time series for each artist in the Physical Era
physical_certificates	3,243	118.0 KB	Success information obtained from Pró-Música Brasil (i.e., certificate type, number of sales)
digital_artist_hot_streak_summary	599	63.3 KB	Summary of each hot streak period achieved by the artists in the Digital Era (i.e., duration, start and end week)
digital_artist_timeseries	211,744	15.6 MB	Detailed time series for each artist in the Digital Era
digital_spotify_charts	41,600	3.3 MB	Spotify Top 200 (weekly chart) for the Brazilian market
digital_spotify_song_info	2,595	565.4 KB	Metadata and acoustic features for hit songs

Era. The gray line corresponds to the original artist time series, while the green line represents the new time series after applying the PAA. The dotted line shows the threshold for the artists' careers. Consequently, all periods above this dotted line represent the Hot Streak period for their careers.

4.2 Data Content

We organize MUHSIC-BR in a set of CSV files, with artist time series and summarized hot streak information. It also includes success information for each Era, i.e., certificates awarded for the Physical Era and streaming charts for Digital Era. It provides information about hit songs from Spotify charts in Digital Era. Table III presents basic statistics and a brief description for each file in MUHSIC-BR.

Success Information. The core of MUHSIC-BR Physical Era dataset is certificates and song sales. Therefore, *physical_ certificates* has those certificates issued by PMB from 1990 to 2015. Besides media type (CD or DVD), such a table also contains information on sales number, awarded artist's name and nationality. In the Digital Era, *digital_ spotify_ charts* has Spotify's weekly Top 200 charts from 2017 to 2020, with artists' names and their songs, position, id and streaming amount.

Time Series. Both *artist_ timeseries* table contain data about the artists' success time series. We calculate aggregated success score, PAA approximation, and threshold used to define hot streak periods for each artist and weekly chart in the Digital Era dataset. The column *is_ hot_ streak* informs whether the week belongs to a hot streak or not, and the column *status_ hot_ streak* stores if the week comes right after of right before a hot streak. The same information is available for Physical Era by using yearly success regarding sales.

Hot Streaks. Both *hot_ streak_ summary* files are similar for both eras. They contain information on every detected hot streak (i.e., artist, duration, start/end date and artists' genres). The only differences are the artist id, available only in the Digital Era table, and the adopted time measure: years and weeks for physical and digital media, respectively.

Songs. The *digitial_ spotify_ song_ info* file is only available for the Digital Era because the Physical Era contains other types of content (as albums and live performances). Therefore, each song entry has relevant data, such as the explicit flag, which indicates inappropriate content for children. In addition, the table provides acoustic features of songs (e.g., danceability, key and tempo). Finally, it also contains data about artists given credit for the music, including collaboration artists.

4.3 Exploratory Data Analysis

The exploratory data analysis on both eras of Brazilian market are similar to Section 3.3. The final version of MUHSIC-BR is split between physical and digital media. On physical media, the final version is composed of: 26 annual charts (collected from 1990 to 2015); 3243 certificates; and 574 artists belonging to 365 musical genres. The final version of the digital media comprises: 208 weekly charts; 2595 hit songs; and 1018 artists belonging to 254 musical genres. Such innovative data provided

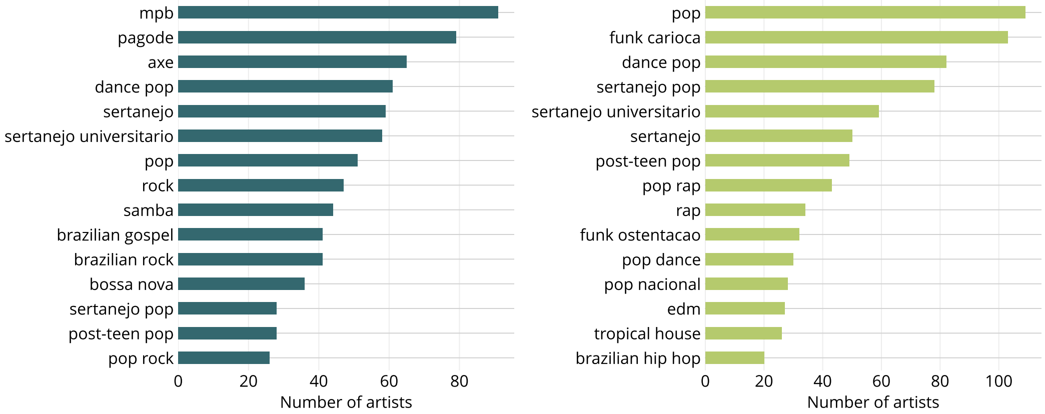


Fig. 8: Top 15 music genres in MUHSIC-BR for Physical (left) and Digital (right) Eras sorted by the number of artists.

by MUHSIC-BR allow analysis about the evolution of artists' careers and genres. Table III summarizes the basic statistics for each table on the dataset.

The genre of artists is also a relevant feature of MUHSIC-BR. Figure 8 shows the 15 most frequent genres in the Brazilian music market. As described in Section 4.1, the data integration for physical media was made between Pró-Música Brasil and Spotify data. For both eras, the genre is related to the artist, not to each song. In the context of physical media, the super-genres *MPB*, *pagode*, *sertanejo*, *pop*, *rock* and *samba* appear. There are several derivative genres, such as *sertanejo universitário*, *brazilian rock* and *post-teen pop*. In the digital age, *pop*, *funk*, *sertanejo* and *rap* stand out. New genres appear in the current era, such as *funk carioca*, *pop rap* and *tropical house*. The three most popular genres in the physical period do not even appear among the 15 in the digital age, which indicates a significant change in musical consumption by Brazilians. Nonetheless, there is a marked consumption of national music in both periods. We can analyze these and other perceptions from the time series generation and detect their hot streaks.

We analyze hot streak periods for both eras. First, Figure 9 (left) shows the number of hot streaks (HS) per artist in the Physical Era. Most artists (72%) had only one hot streak period in their careers, while 22% of artists had two. Artists including Michael Jackson, Tim Maia, Nelson Gonçalves and Led Zeppelin make up a select group of artists with four hot streaks in their careers. Figure 9 (right) shows the duration of hot streaks, most lasting two years. Few artists such as Black Eyed Peas, Seu Jorge and Raça Negra obtained eight-year hot streaks. Jota Quest was the only band that endured to maintain hot streak success for ten years.

For Digital Era, Figure 10 (left) indicates that 80% of artists had only one hot streak in the Digital Era so far, while only 2% (12 artists) had three hot streaks in their careers. We highlight international artists between them, such as Beyoncé, Calvin Harris and Taylor Swift. However, Brazilian artists also have a strong presence, such as Simone & Simaria, Anitta, Luan Santana and Pabllo Vittar, which hold hot streaks for 12 or 24 weeks. In addition, singers from different musical genres achieved hot streaks lasting 48 weeks, although only Billie Eilish sustained a 60-week hot streak.

5. FORMAT AND USAGE

Both MUHSIC and MUHSIC-BR are publicly available at Zenodo [Oliveira et al. 2021a], an open dissemination research data repository that shares, curates and publicizes data and software for everybody. The datasets are available in two formats:

MySQL dump (.sql file), which creates the 11 tables of the relational schema of MUHSIC (Section

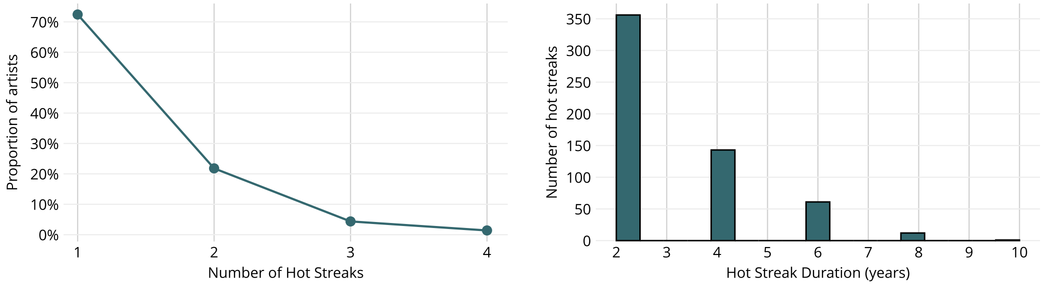


Fig. 9: Hot streak statistics in MUHSIC-BR for Physical Era: hot streaks per artist (left) and duration (right).

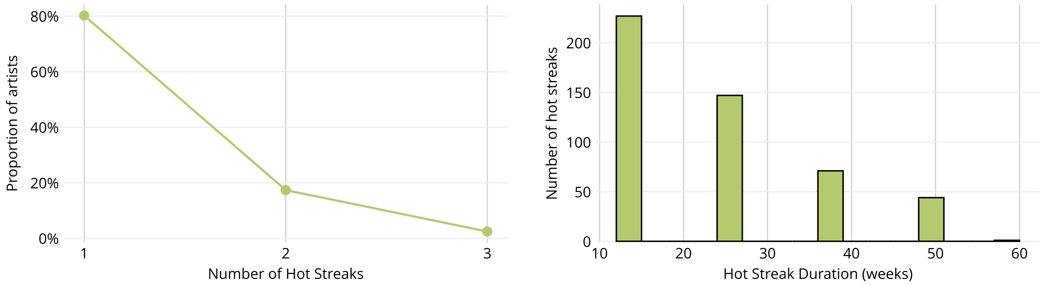


Fig. 10: Hot streak statistics in MUHSIC-BR for Digital Era: hot streaks per artist (left) and duration (right).

3.2). This format is recommended for simple and complex queries, as MySQL is better at dealing with large amount of data; and

CSV files, useful for processing data in Python or R to perform complex analyses and visualizations.

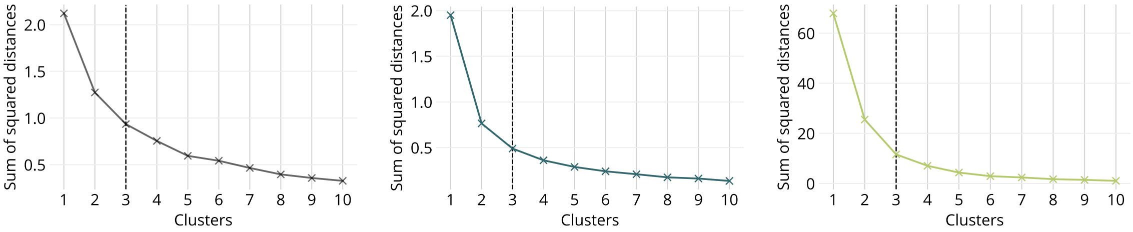
6. CLUSTERING ARTISTS BY SUCCESS LEVEL

Although appearing in the charts or receiving a sales certificate is already proof of an artist's success, they present different success levels when compared with each other. For instance, the song *Blinding Lights* by The Weeknd is the longest charting hit in Hot 100 history (88 weeks) and the song with the longest time in the chart's Top 5 (43 weeks).[12] In contrast, the Brazilian singer Anitta made history as she debuted in Hot 100 in position 91 with her single *Me Gusta* featuring Cardi B and Myke Towers. To better understand the characteristics of the different success levels achieved by artists, we perform a cluster analysis as an example of application using success data from MUHSIC and MUHSIC-BR. The preliminary cluster results for Brazil are published in Barbosa et al. [2021]. Here, we enhance such analyses by applying the methodology for the United States data and comparing both markets.

We use K-Means for clustering, as it is the most used method for dividing a dataset into k groups. We perform three distinct analyses: United States (MUHSIC), Brazil in Physical Era (MUHSIC-BR), and Brazil in Digital Era (MUHSIC-BR). For each artist, we consider the following features as input for K-Means: the number of hot streaks, the aggregated success metric and the time series threshold (see Sections 3.1 and 4.1). Regarding the success metric, we consider the sum of Rank Score DCG for MUHSIC and the total sales and streams for MUHSIC-BR (Physical and Digital, respectively).

To find the number k of clusters (input of K-Means), we use the Elbow Method [Bholowalia and Kumar 2014]. It plots the explained variation according to the number of clusters and chooses the curve *elbow* as the optimal k. For all three experiments, its outcome suggests $k = 3$, as shown in Figure

[12] Billboard: `https://bit.ly/BlindingLightsHot100`



(a) United States

(b) Brazil (Physical Era)

(c) Brazil (Digital Era)

Fig. 11: Elbow method. For all experiments, we choose $k = 3$ (dashed lines), as it is the elbow of the curves.

Table IV: Main statistics on the artist clusters.

United States					Brazil (Physical Era)					Brazil (Digital Era)				
	All	SHA	BHA	THA		All	SHA	BHA	THA		All	SHA	BHA	THA
NA	6,066	5,428	587	51	NA	574	527	38	9	NA	1,018	940	70	8
AHS	1.2	1.1	1.7	2.2	AHS	1.3	1.3	1.8	1.9	AHS	0.6	0.5	1.6	1.6
MRS	187.6	148.8	2,919.1	13,072.5	MSA	8.5	8	152	507	MST	7.6	3.9	2,190	10,857
MT	0	0	8.1	29.3	MT	0	0	72,500	300,000	MT	0	0	2,052,131	7,145,116

NA: Number of Artists **AHS**: Average Number of HS **MRS**: Median Rank Score DCG **MSA**: Median Sales (10^4)
MST: Median Streams (10^5) **MT**: Median Threshold

11. We evaluate clustering quality using Silhouette Coefficient, which measures the distinctness of the clusters.[13] Results indicate a good quality, as the Silhouette values are all close to 1 (US – 0.702; Brazil, Physical – 0.742; Brazil, Digital – 0.850). We then name the resulting clusters according to the success metric: *Spike Hit Artists* (SHA), *Big Hit Artists* (BHA) and *Top Hot Artists* (THA). Table IV presents the main statistics on the clusters for the United States (MUHSIC) and Brazil (MUHSIC-BR).

Spike Hit Artists (SHA). This cluster contains most artists (around 90%) in all three experiments. The median threshold for such artists is zero, indicating that their time series is mainly composed of zero values (i.e., no presence in the Hot 100, no PMB certificate or presence in Spotify Brazilian chart). Consequently, their success metric (MRS, MSA and MST) is lower than other clusters. Finally, SHA present fewer hot streak periods than other artists, on average. Such a result may indicate that success happens sparsely in their time series as there are not many hot streaks even with a low threshold.

Big Hit Artists (BHA). This cluster represents an intermediate level of success when compared to the other clusters and contains approximately 5% to 10% of total artists. There is a significant increase in the success metric and the average number of hot streaks in all three experiments. In addition, the median threshold for Big Hit Artists is not zero, indicating a higher presence in the charts that reflects on the success time series.

Top Hit Artists (THA). This is the cluster with the most successful artists (approximately 0.5% to 2% of the total). Such artists present the highest median success metric, indicating a solid and consistent presence in the charts in the US and the highest sales and stream count in Brazil. In addition, their threshold value is much higher when compared to SHA and BHA, which means that besides having a higher presence in the charts, they also present higher success numbers.

Table V presents the Top 5 artists for each cluster in all three experiments (US and Brazil – Physical and Digital). All clusters have specific patterns that reveal relevant information about success in the US and Brazil. For example, the most successful artists (i.e., THA) in the US are primarily English-speakers, while in Brazil, they are all native Brazilians. In addition, the success metric is a relevant

[13] The Silhouette Coefficient varies from -1 to 1. Therefore, values close to 1 indicate that clusters are clearly distinguished, while -1 means clusters are assigned incorrectly [Rousseeuw 1987].

Table V: Top 5 artists of each cluster, sorted by their success metric.

	United States			Brazil (Physical Era)			Brazil (Digital Era)		
	Artist	**# HS**	**Rank DCG**	**Artist**	**# HS**	**Sales**	**Artist**	**# HS**	**Streams**
THA	Drake	1	36,197.7	Ivete Sangalo	1	9.525M	Marília Mendonça	1	2.028B
	Rihanna	2	28,970.7	Padre Marcelo Rossi	2	6.09M	Zé Neto & Cristiano	1	1.345B
	Mariah Carey	4	23,589.4	Zeze C. & Luciano	2	5.58M	Matheus & Kauan	2	1.129B
	Lil Wayne	1	21,856.7	Aline Barros	2	5.575M	Jorge & Mateus	2	1.086B
	Taylor Swift	2	19,177.4	Roberto Carlos	1	4.530M	Henrique & Juliano	1	1.085B
BHA	Stevie Wonder	4	12,988.5	Bruno & Marrone	2	3.96M	MC Kevinho	2	644M
	T.I.	2	11,569.7	É o Tchan	1	3.27M	Alok	1	641M
	Boyz II Men	2	11,254.5	Só Pra Contrariar	1	3.215M	Felipe Araújo	2	590M
	The Rolling Stones	5	11,238.4	Paula Fernandes	1	3.095M	MC Kevin o Chris	1	527M
	Bee Gees	2	10,892.8	Jota Quest	1	3.09M	Gustavo Mioto	2	503M
SHA	21 Savage	1	3,847.7	Adele	1	1.64M	MEDUZA	1	137M
	Iggy Azalea	1	3,541.1	Victor & Leo	2	1.42M	The Chainsmokers	1	137M
	Creedence	1	3,466.1	João Paulo & Daniel	1	1.36M	George H. & Rodrigo	2	130M
	B.o.B	1	3,465.6	Frank Aguiar	2	1.36M	Orochi	1	127M
	ABBA	1	3,446.7	Terra Samba	1	1.34M	Humberto & Ronaldo	2	126M

factor in distinguishing between clusters. For instance, the Top 1 artists in THA have a metric value around three times higher than the Top 1 artists in BHA. A similar relation happens between BHA and SHA. Overall, all such observations reinforce the importance of our clustering approach since it allows us to distinguish the levels of successful artists.

7. OTHER POTENTIAL APPLICATIONS

Besides clustering artists by success level, in this section, we point out other scenarios and applications that illustrate the potential impact and usability of MUHSIC and MUHSIC-BR for Music Information Retrieval. Such a multidisciplinary area covers a wide range of research topics, including music classification, recommendation, genre identification, and others. In particular, the metadata and success-related content available in both datasets act as valuable resources to be used in different MIR tasks. Next, we highlight two examples where the dataset can be directly applied.

Hit Song Science (HSS). It aims to predict song success before being released [Pachet 2011]. The premise of HSS is hit songs comprise a specific feature set that makes them appealing to most people. Such attributes can then be exploited through machine learning methods to assess whether a song will become a chart-topping hit. In this context, MUHSIC and MUHSIC-BR are ready to be used as labeled datasets for training and testing such methods. Also, both datasets enable exploring of similarities, patterns, and differences through learning models as well.

Music Genre Classification (MGC). It aims to classify a song into one or more musical genres and is not task, as the boundaries between genres remain blurred [Scaringella et al. 2006; Oliveira et al. 2020]. Both MUHSIC and MUHSIC-BR allow an easy linkage of the song acoustic features to the genres of the artist who sings it. As recent work on MGC use deep learning to assign genres to a song [Castillo and Flores 2021; Liu et al. 2021a], the features provided by our datasets may enrich the experiments and thus enhance the classification results. Moreover, the variety of music genres available in the dataset may assist in specializing the classification task, as it helps distinguishing between distinct sub-genres (e.g., *dance pop* and *indie pop* as *pop* sub-genres) [Feng and Feng 2021].

There are other success-related applications that can benefit from our datasets, as they include several factors that shape what listeners consume. For example, Time Series Analysis is useful for understanding and predicting artists' success according to one or multiple attributes [Janosov et al. 2020]. Such applications are discussed in detail in the MUHSIC dataset paper [Oliveira et al. 2021b].

8. CONCLUSION AND FUTURE IMPROVEMENTS

This article assessed musical success in the United States and Brazil from a temporal perspective. Our first contribution was to present MUHSIC and MUHSIC-BR, two enhanced open datasets including metadata and success information on the main elements of the music ecosystem within each country. MUHSIC combined information extracted from Billboard Hot 100 and Spotify, whereas MUHSIC-BR was built from Pró-Música Brasil certificates and Spotify charts to represent success in Physical and Digital Eras respectively. Although they share much content with other music datasets, their novelty relies on temporal success by providing success time series as representatives of musical careers.

As a first application of our datasets, we grouped artists into different success levels based on features extracted from their time series and hot streak periods. For both the United States and Brazil, we identified three distinct clusters: Spike Hit Artists (SHA), whose success happens sparsely in their time series; Big Hit Artists (BHA), with an intermediate level of success; and Top Hit Artists (THA), who have a solid and consistent success over time. All clusters present distinct success patterns in such markets, revealing relevant insights on music consumption. For instance, our results showed that success may be directly related to local culture, as the most successful artists are mainly from that country and sing in its language (e.g., English for the US and Portuguese for Brazil).

Overall, our findings benefit not only the Music Information Retrieval (MIR) community but also the music industry as a whole. First, our datasets are accessible and ready-to-use for complex tasks, including hit song science, genre classification and hot streak prediction. Then, the artist clusters may assist in recommendation tasks (e.g., recommending artist partnerships or songs to listeners). Finally, the temporal aspect of musical success can enhance the understanding of drifts in musical preferences. In short, this work represents a meaningful step towards the science behind musical success and its temporal dynamics, allowing the music industry to connect people to content relevant to them.

Challenges and Limitations. MUHSIC and MUHSIC-BR are not free from limitations, which may be improved in future versions. The key challenges relate to the heterogeneity of the data sources used in the data collection and data integration phases. Another limitation is that the data sources consider only mainstream and popular music, generalizing the information. All such limitations are better discussed in the dataset paper [Oliveira et al. 2021b].

Future Work. We first plan to consider additional data sources to extend further our feature sets, such as lyrics, awards, and other relevant metadata. Moreover, the next step would be including non-hit songs and artists to increase the data diversity. Finally, we plan to expand the market coverage of our data, since local engagement shapes the global music environment.

REFERENCES

Aggarwal, C. C. (2016). *Recommender Systems - The Textbook.* Springer, Switzerland. doi:10.1007/978-3-319-29659-3.

Al-Beitawi, Z., Salehan, M., and Zhang, S. (2020). Cluster analysis of musical attributes for top trending songs. In *HICSS*, pages 1–7.

Barbosa, G. R. G., Melo, B. C., Oliveira, G. P., Silva, M. O., Seufitelli, D. B., and Moro, M. M. (2021). Hot Streaks in the Brazilian Music Market: A Comparison Between Physical and Digital Eras. In *Brazilian Symposium on Computer Music (SBCM)*, pages 155–162, Recife, Brazil.

Bertin-Mahieux, T. et al. (2011). The Million Song Dataset. In *Int'l Society for Music Information Retrieval Conf.*, pages 591–596, Miami, USA.

Bertoni, A. and Lemos, R. (2021). Três datasets criados a partir de um banco de canções populares brasileiras de sucesso e não-sucesso de 2014 a 2019. In *Brazilian Symposium on Databases: Dataset Showcase Workshop*, pages 11–20, Rio de Janeiro, Brazil. doi:10.5753/dsw.2021.17410.

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *Int'l J. of Computer Applications*, 105(9):17–24.

Bogdanov, D., Porter, A., Schreiber, H., Urbano, J., and Oramas, S. (2019). The acousticbrainz genre dataset: Multi-source, multi-level, multi-label, and large-scale. In *Int'l Society for Music Information Retrieval Conf.*, pages 360–367.

Byrd, D. and Crawford, T. (2002). Problems of music information retrieval in the real world. *Information Processing & Management*, 38(2):249–272. doi:10.1016/S0306-4573(01)00033-4.

Castel-Branco, G. et al. (2021). Puremic: A new audio dataset for the classification of musical instruments based on convolutional neural networks. *J. Signal Process. Syst.*, 93(9):977–987. doi:10.1007/s11265-021-01661-3.

Castillo, J. R. and Flores, M. J. (2021). Web-based music genre classification for timeline song visualization and analysis. *IEEE Access*, 9:18801–18816. doi:10.1109/ACCESS.2021.3053864.

Çimen, A. and Kayis, E. (2021). A longitudinal model for song popularity prediction. In *Int'l Conf. on Data Science, Technology and Applications (DATA)*, pages 96–104, Online Streaming. doi:10.5220/0010607700960104.

Cosimato, A. et al. (2019). The conundrum of success in music: Playing it or talking about it? *IEEE Access*, 7:123289–123298. doi:10.1109/ACCESS.2019.2937743.

Feng, M. and Feng, W. (2021). Evaluation of parallel and sequential deep learning models for music subgenre classification. *Math. Found. Comput.*, 4(2):131. doi:10.3934/mfc.2021008.

Ferraro, A. et al. (2021). Melon playlist dataset: A public dataset for audio-based playlist generation and music tagging. In *Int'l Conf. on Acoustics, Speech and Signal Processing*, pages 536–540, Toronto, Canada.

Garimella, K. and West, R. (2019). Hot streaks on social media. In *Int'l Conf. on Web and Social Media*, pages 170–180, Munich, Germany.

Georges, P. and Nguyen, N. (2019). Visualizing music similarity: clustering and mapping 500 classical music composers. *Scientometrics*, 120(3):975–1003.

Janosov, M., Battiston, F., and Sinatra, R. (2020). Success and luck in creative careers. *EPJ Data Sci.*, 9(1):9. doi:10.1140/epjds/s13688-020-00227-w.

Karydis, I., Gkiokas, A., and Katsouros, V. (2016). Musical track popularity mining dataset. In *IFIP Int'l Conf. on Artificial Intelligence Applications and Innovations*, pages 562–572, Greece. doi:10.1007/978-3-319-44944-9_50.

Keogh, E. J. and Pazzani, M. J. (2000). Scaling up dynamic time warping for datamining applications. In *ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 285–289, Boston, USA. doi:10.1145/347090.347153.

Kowald, D., Müllner, P., Zangerle, E., Bauer, C., Schedl, M., and Lex, E. (2021). Support the underground: characteristics of beyond-mainstream music listeners. *EPJ Data Sci.*, 10(1):14. doi:10.1140/epjds/s13688-021-00268-9.

Liu, C., Feng, L., Liu, G., Wang, H., and Liu, S. (2021a). Bottom-up broadcast neural network for music genre classification. *Multim. Tools Appl.*, 80(5):7313–7331.

Liu, L., Dehmamy, N., Chown, J., Giles, C. L., and Wang, D. (2021b). Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nature Communications*, 12(5392):249–272. doi:10.1038/s41467-021-25477-8.

Liu, L., Wang, Y., Sinatra, R., Giles, C. L., Song, C., and Wang, D. (2018). Hot streaks in artistic, cultural, and scientific careers. *Nature*, 559(7714):396–399. doi:10.1038/s41586-018-0315-8.

Melchiorre, A. B. et al. (2021). Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, 58(5):102666. doi:10.1016/j.ipm.2021.102666.

Oliveira, G. P., Barbosa, G. R. G., Melo, B. C., Silva, M. O., Seufitelli, D. B., Lacerda, A., and Moro, M. M. (2021a). MUHSIC: An Open Dataset with Temporal Musical Success Information. *Zenodo*. doi:10.5281/zenodo.4779002. https://doi.org/10.5281/zenodo.4779002.

Oliveira, G. P., Barbosa, G. R. G., Melo, B. C., Silva, M. O., Seufitelli, D. B., and Moro, M. M. (2021b). MUHSIC: An Open Dataset with Temporal Music Success Information. In *Brazilian Symposium on Databases: Dataset Showcase Workshop*, pages 65–76, Rio de Janeiro, Brazil. doi:10.5753/dsw.2021.17415.

Oliveira, G. P., Silva, M. O., Seufitelli, D. B., Lacerda, A., and Moro, M. M. (2020). Detecting collaboration profiles in success-based music genre networks. In *Int'l Society for Music Information Retrieval Conf.*, Montreal, Canada.

Oramas, S., Barbieri, F., Nieto, O., and Serra, X. (2018). Multimodal deep learning for music genre classification. *Trans. Int. Soc. Music. Inf. Retr.*, 1(1):4–21.

Pachet, F. (2011). Hit song science. In Tao Li, Mitsunori Ogihara, G. T., editor, *Music Data Mining*, chapter 10, pages 305–326. CRC Press, New York, USA.

Pati, A., Gururani, S. K., and Lerch, A. (2020). dmelodies: A music dataset for disentanglement learning. In *Int'l Society for Music Information Retrieval Conf.*, pages 125–133, Montreal, Canada.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Roy, S., Biswas, M., and De, D. (2020). imusic: a session-sensitive clustered classical music recommender system using contextual representation learning. *Multim. Tools Appl.*, 79(33-34):24119–24155.

Scaringella, N., Zoia, G., and Mlynek, D. (2006). Automatic genre classification of music content: a survey. *IEEE Signal Process. Mag.*, 23(2):133–141. doi:10.1109/MSP.2006.1598089.

Silva, M. O., Rocha, L. M., and Moro, M. M. (2019). MusicOSet: An Enhanced Open Dataset for Music Data Mining. In *Brazilian Symposium on Databases: Dataset Showcase Workshop*, pages 408–417, Fortaleza, Brazil.

Zangerle, E., Huber, R., and Yang, M. V. Y.-H. (2019). Hit Song Prediction: Leveraging Low- and High-Level Audio Features. In *Int'l Society for Music Information Retrieval Conf.*, pages 319–326, Delft, The Netherlands.