# A model based approach to Spotify data analysis: a Beta GLMM

**Mariangela Sciandra** · **Irene Carola Spera**

**Abstract** Digital music distribution is increasingly powered by automated mechanisms that continuously capture, sort and analyze large amounts of Web-based data. This paper deals with the management of songs audio features from a statistical point of view. In particular, it explores the data catching mechanisms enabled by Spotify Web API, and suggests statistical tools for the analysis of these data. Special attention is devoted to songs popularity and a Beta model including random effects is proposed in order to give a first answer to questions like: which are the determinants of popularity? The identification of a model able to describe this relationship, the determination within the set of characteristics of those considered most important in making a song popular is a very interesting topic for those who aim to predict the success of new products.

**Keywords** Spotify Web API · audio features · Popularity Index · Beta GLMM

**Riassunto** *La distribuzione di musica digitale è sempre più alimentata da meccanismi automatizzati che consentono l'acquisizione e l'analisi di grandi quantità di dati presenti sul Web. Questo articolo si occupa della gestione delle caratteristiche audio delle canzoni da un punto di vista statistico. In particolare, esplora i meccanismi di raccolta dati ottenuti dalla Spotify Web API, e suggerisce strumenti statistici per l'analisi di questi. Particolare attenzione è dedicata alla popolarità delle canzoni e viene proposto un modello Beta con effetti casuali per dare una prima risposta a domande quali: Cosa determina la popolarità? Perché una canzone è popolare? L'individuazione di un modello in grado di descrivere questa relazione, la determinazione delle caratteristiche Spotify considerate più*

M. Sciandra, I.C. Spera

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy
E-mail: mariangela.sciandra@unipa.it, irenecarola@libero.it

*importanti nel determinare la popolarità di una canzone è un argomento molto interessante, in particolare, per coloro che mirano a prevedere il successo di nuovi prodotti.*

**Parole chiave**
    *Spotify Web API - caratteristiche musicali - Indice di popolarità - Beta GLMM*

## 1 Introduction

Music plays an important role in everyday life of people, and with digitalization, large collections of musical data are formed, which tend to be further cumulated by music lovers (Sloboda, 2011). This has led to music collections, not only on the private shelf as audio or video discs and domain discs, but also on the hard disk and online, to grow beyond what was previously impossible. With the advent of new technologies, it has become impossible for a single individual to keep track of the music and the relationships between different songs. The techniques of data mining and automatic learning can help the navigation in the world of music (Lerch, 2018).

Data mining strategies are often based on two main problems: the type of available data and the use you want to make of them. What kind of data is the music? A collection of music tracks consists of various types of data; for example, data could consist of music audio files or metadata such as track title and artist name (Pachet, 2011). What kind of analysis can be carried out? The musical data mining provides specific methods to answer to the most varied questions: e.g. gender classification, identification of artists/singers, mood/emotion detection, instrument recognition, similarity search music, musical synthesis and so on.
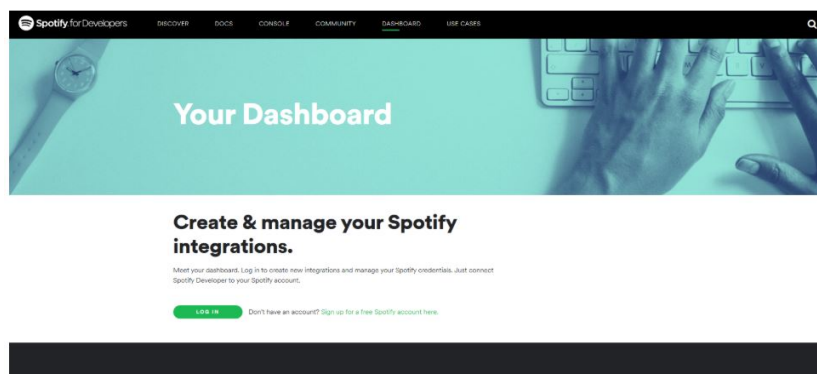
This research investigates the relationship between song data audio features obtained from the Spotify database (e.g. key and tempo) and song popularity, measured by the number of streams that a song has on Spotify. Previous researches on the topic of new product success prediction have identified multiple approaches to answer to this question. Moreover, the existing body of research which defines many popularity prediction models stresses the complexity of the mechanisms of song popularity. Research from Lee and Lee (2018) shows that it is feasible to predict the popularity metrics of a song significantly better than random chance based on its audio signal. Additionally, Ni et al. (2015) also show that certain audio features such as Loudness, duration and harmonic simplicity correlate with the evolution of musical trends. Dhanaraj and Logan (2005) propose features from both songs' lyrics and audio content for prediction of hits and also study a hit detection model based solely on lyrics' features. In an attempt to predict the popularity of a song from Spotify's song data, the research of Will Berger (2017) uses (Echo-Nest) audio features similar to this research and uses Spotify's own calculated metric "popularity" to measure popularity. Other attemps through classical linear regression or quadratic models can be found on the net but they are not exaustive works and do not take into accont the particular data structure and other aspects that could led to biased predictions. This is the reason why this paper can be considered as an innovative way to look at popularity predictions and represents an innovative approach inside literature. Paper is organized in the following way: in Sec. 2 the way to connect Spotify Web API to the R software is explained; available songs' Spotify

audio features are described in Sec. 3; Sec.4 is devoted to the proposal of a new class of models to predict songs'popularity as a function of the Spotify audio features. An application to a real datataset is carried out in Sec. 5, future work and conclusions follow.

## 2 Spotify and R

Spotify (SpotifyWebAPI, 2019) is one of the most famous music Apps in the world, a program of last generation grown up exponentially over the last few years. This platform allows you to listen to music streaming to computers, smartphones and tablets, choosing from more than 30 million tracks, old and new, of the main international record companies, without having to purchase individual songs or albums legally.

In this section a brief description of the steps necessary to connect Spotify Web API to the R software are reported. First you have to create an app on the Spotify's developer platform, accessing this dashboard:



After accessing the page, you have to login and create an app:

1. give a name to the application and provide a brief description;

2. then you have to specify the purpose for which it is created (example for commercial purposes, teaching, etc. ... );

3. finally, accept the conditions to complete the account configuration.

After this procedure, user identification code and password are generated, both are alphanumeric codes that are essential to release the access token to connect Spotify with R.

The packages used to make the connection are:

– *rvest*: it allows you to extract data from a web page (web scraping);

– *tidyverse*: it is used for data transformation and cleaning;

– *spotifyr*: it extracts the song's audio features from Spotify's Web API.

After loading the R libraries listed above, the following functions are used to connect and extract data from Spotify (Charlie, Charlie):

```
Sys.setenv(SPOTIFY_CLIENT_ID = 'xxxxxxxxxxxxxxxxxxxx')
Sys.setenv(SPOTIFY_CLIENT_SECRET ='xxxxxxxxxxxxxxxxxxxx')
access_token ← get_spotify_access_token()
Liga← get_artist_audio_features('Ligabue')
```

They are especially useful for:

– Sys.setenv: storing in R the identification code and password to access Spotify;

– get_spotify_access_token(): generating the token to access all data using the Spotifyr library;

- **get_artist_audio_features**: extracting all the audio characteristics of the songs that make up the discography, by specifying the name of the singer.

## 3 Spotify Audio Features

Spotify Web API makes users able to extract several audio features of songs. The available features that also have been used in this paper are listed in Table 1:

**Table 1**: Spotify audio features.

| acousticness | float | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
|---|---|---|
| analysis_url | string | An HTTP URL to access the full audio analysis of this track. An access token is required to access this data. |
| danceability | float | Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. |
| duration_ms | int | The duration of the track in milliseconds. |
| energy | float | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. |
| id | string | The Spotify ID for the track. |
| instrumentalness | float | Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| key | int | The key the track is in. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C#/Db, 2 = D, and so on. |
| liveness | float | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. |
| loudness | float | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. |
| mode | int | Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. |
| speechiness | float | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |
| tempo | float | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. |
| time_signature | int | An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). |
| track_href | string | A link to the Web API endpoint providing full details of the track. |
| type | string | The object type: "audio_features" |
| uri | string | The Spotify URI for the track. |
| valence | float | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). |

The aim of this paper is to investigate if audio features from Spotify can be considered as determinants of the stream popularity of songs (Nijkamp, 2018).

3.1 Track API popularity

Among all the features returned by spotify, song popularity plays an important role. The popularity of a track is a value between 0 and 100, with 100 the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of track plays and taking into account how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity. Note that the popularity value may lag actual popularity by a few days: the value is not updated in real time.

Songs popularity is an important issue for music industry. In 2017 music industry generated $8.72 billion in the United States alone. Thanks to growing streaming services (Spotify, Apple Music, etc) the industry continues to flourish. The top 10 artists in 2016 generated a combined $362.5 million in revenue. The question of what makes a song popular has been studied before with varying degrees of success (Giles, 2006). Every song has key characteristics including lyrics, duration, artist information, temp, beat, Loudness, chord, etc. Previous studies that considered lyrics to predict a song's popularity had limited success.

## 4 Using Spotify data to predict what songs will be hits

The aim of this section is the identification of the determinants of songs'popularity. In particular, we want to investigate the possible relationship between the audio characteristics of the songs in the Spotify database (for example, Energy, Loudness, etc. ...) and the popularity of the songs also available in the Spotify dataset. The identification of a model able to describe this relationship, the determination within the set of characteristics of those considered most important in making a song popular is a very interesting topic for those who aim to predict the success of new products. Then, the fundamental question is: What does determine popularity? Why is a song popular?

In cultural markets like music, forecasting is very complex. Studies in this field called Hit Song Science (HSS) are of interest to record companies but also to consumers themselves and to Spotify (Middlebrook and Sheik, 2019). Previous attempts in this direction have always referred to linear or quadratic model regression (Nasreldin, 2018). In this paper the application of a Beta regression with random effects is proposed. The choice of this class of models derives by the nature of the response variable (a continuous variable limited in $[0, 100]$) and by the correlation structure in the data: it is assumed, in fact, that songs belonging to the same album can be related to each other more than song from different albums; ignoring this level of hierarchy in the data could lead to biased or inefficient results.

4.1 Beta regression for correlated data

In this work we study the dependence of the popularity of songs on the musical characteristics by using an extension of the Beta regression model, including random effects. The resulting model will

be a generalized Beta model with mixed effects (Beta GLMM). Before defining the model from a theoretical point of view, following a brief remaind to the classical Beta regression and the Generalized Linear models with Mixed Effects (GLMMs); this brief summary will be useful to understand the reasons why we have chosen to focus on this particular model and for the theoretical definition of the model itself.

### 4.1.1 Beta regression

Beta distribution is a continuous probability distribution defined in the unitary range with a density function given by:

$$f(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},  \quad (1)$$

where $\Gamma(.)$ indicates the Gamma function. The parameter $\mu$ indicates the expected value of $Y$, i.e. $E(Y) = \mu$. The parameter $\phi$ meets the definition of a precision parameter because, for fixed $\mu$, the higher the value of $\phi$, the lower the variance of the dependent variable. More specifically,

$$Var(Y) = \frac{\mu(1-\mu)}{1+\phi}.  \quad (2)$$

In Beta regression models (Ferrari and Cribari-Neto, 2004), the parameter that indicates the average $\mu \in (0,1)$ of the Beta distribution is expressed as a function of the covariates, while the parameter of precision $\phi \in R^+$ is treated as a disturbance parameter. In order to ensure that the linear predictor takse on values in the space given by the dependent variable's support, the link logit represents the most commonly chosen link function:

$$g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^T \beta,  \quad (3)$$

where $x_{ij}^T$ denotes a vector of explanatory variables, and $\beta$ refers to the vector of regression coefficients, $i = 1, \ldots, N$. The Beta distribution is defined only on the open unit interval. If exact one and zero values are admitted, these values must be transformed in order to ensure the nature of the Beta distribution support (Bonat et al., 2014).The most frequently applied transformation is:

$$Y^* = [Y(N-1) + 0.5]/N  \quad (4)$$

where $Y^*$ is the transformed and Y is the untransformed dependent variable. Alternatively, it was suggested to add a small amount of $\varepsilon$, for example 0.005 or 0.01 to the lower limit, and subtract the same amount from the upper limit. Hunger et al. (2012) also observe that when the resulting values are too close to the boundary points, the accuracy of the estimates may decrease significantly.

### 4.1.2 Generalized Linear Mixed Models (GLMM)

Generalized Linear Mixed Models (or GLMM) are an extension of the Generalized Linear Model (GLM) in which the linear predictor contains random effects in addition to the usual fixed effects. For this model class, the assumption of homogeneity and independence of the sample units is lost. In addition, with regard to the distribution of the response variable, the GLMM inherit from the GLM the idea of extending mixed linear models to the non-normal data case (Lovison et al., 2011). GLMM provide a wide range of models for the analysis of data that have some form of grouping, since differences between groups can be modelled through the use of a random effect. The basic concept is the structure in *cluster*: the data with clustering structure has a univariate response variable *y* double indexed, *i* for the first level units and *j* for the second level units and a vector $x_{ij}$ of explanatory variables p for the j-*th* unit in the i-*th* cluster. It is important to remember that clusters can have different sizes and that this can influence the results of the analysis.

These models are useful in the analysis of many types of data, including longitudinal data. The general form of the model, in matrix notation, is:

$$y = X\beta + Zb + \varepsilon \tag{5}$$

where y is a column vector $N \times 1$; X is an array $N \times p$ of explanatory variable *p*; $\beta$ is a column vector $p \times 1$ of fixed effect regression coefficients; Z is the random effects model matrix $N \times q$ for random effects q; b is a vector $q \times 1$ of random effects; $\varepsilon$ is the column vector $N \times 1$ of residues. The assumptions that underlie this class of models can be summarized as follows:

$$Y_{ij} \mid (x_{ij}, z_{ij}, b_i) \sim C.\xi.N(\theta_{ij}, \phi);$$

$$\eta_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i;$$

$$g(\mu_{ij}) = \eta_{ij};$$

$$\mu_{ij} = h(\eta_{ij}) = g(\eta_{ij})^{-1};$$
$$b_i \sim (0, \Sigma_q);$$
$$Y_i \perp Y_j \quad \forall i \neq j.$$

By putting together all the assumptions the conditional distribution is easily derived:

$$f(y_{ij} \mid x_{ij}, z_{ij}, b_i) = exp\left\{ \frac{y_{ij}\theta_{ij} - b(\theta_i)}{\phi} + c(y_{ij} + \phi) \right\};$$

Conditioning on $b_i$, observations from the same cluster are assumed to be independent. In addition, the conditional expected value is related to the linear predictor (containing both random and fixed effects) by the following linking function $g(\cdot)$:

$$g(\mu_{ij}) = x_{ij}^T\beta + Z_{ij}^T b_i.$$

### 4.1.3 The Beta GLMM

In longitudinal analyses or when subjects have any grouping structure, observations related to the same unit will typically be correlated, violating the assumption of independence of observations typical in regression models. The dependence within clusters can be accounted for by adding random cluster or subject effects in the linear predictor (Bonat et al., 2014). Consider the case of longitudinal studies where $j = 1, ..., n_i$ observations are nested within $i = 1, ..., N$ subjects. Let $b_i$ denote the vector of random effects specific to each subject $i$.

Adding random effects to the beta regression model in (3.3) we get the GLMM beta (Bonat et al., 2014) given by

$$log \frac{\mu_{ij}}{1 - \mu_{ij}} = x_{ij}^T \beta + z_{ij}^T b_i \qquad con \quad b_i \sim N(0, G) \qquad (6)$$

where $z_{ij}^T$ is a vector of explanatory variables, and $G$ is the defined positive covariance matrix of random effects. Note that although the assumption of normality for random effects is common and statistically convenient, other distribution hypotheses are also possible. In a longitudinal study, the $b_i$ is typically a scalar (for random intercept models) or a bivariate vector (for random intercept models with a random regression coefficient), i.e. $z_{ij}^T = (1, t_{ij})$, where $t_{ij}$ is the measurement time $j$ for the subject $i$.

In the Beta GLMM the regression parameters have only one specific interpretation per unit and do not describe the effect of the respective variable on the population average; this is due to the non-linear transformation of the average response (i.e. the logit link) as it can be deduced that

$$logit(E(Y_{ij}|b_i)) = x_{ij}^T \beta + z_{ij}^T b_i, \qquad (7)$$

but

$$logit(E(Y_{ij}|b_i)) \neq x_{ij}^T \beta.$$

Model parameters can be estimated by maximizing the marginal probability that is obtained by integrating the joint distribution of $[\boldsymbol{Y}, \boldsymbol{b}]$ on random effects. The contribution to the log-likelihood by each group is as follows:

$$f_i(y_i|\beta, \Sigma, \phi) = \int \prod_{j=1}^{n_i} f_i(y_{ij}|b_i, \beta, \phi)(b_i|\Sigma) db_i. \qquad (8)$$

Assuming independence among the $N$ groups, the full likelihood is:

$$L(\beta, \Sigma, \phi) = \prod_{i=1}^{N} f_i(y_i|\beta, \Sigma, \phi). \qquad (9)$$

## 5 An application: Luciano Ligabue's Audio Spotify features analysis

In order to apply the proposed model to a real case, the whole discografy of Luciano Ligabue has been analized. Luciano Ligabue is one of the most successful Italian artists. He is a famous singer-songwriter, film director and writer. In over thirty years of his career he has won more than sixty awards for his musical activity, five awards for his work as a writer and twelve awards for his film work (Poggini, 2009). In this application 19 albums for a total of 273 songs have been considered. Among the information available on Spotify there is also the year in which the song was published. It's clear that, statistically speaking, the year represents a proxy for the album because there are no different albums released in the same year. So, as the album has been assumed as grouping variable, the year has not been considered among the covariates. However, for completeness of information, the graph below (Fig. 1) shows the average values over time of the examined characteristics:
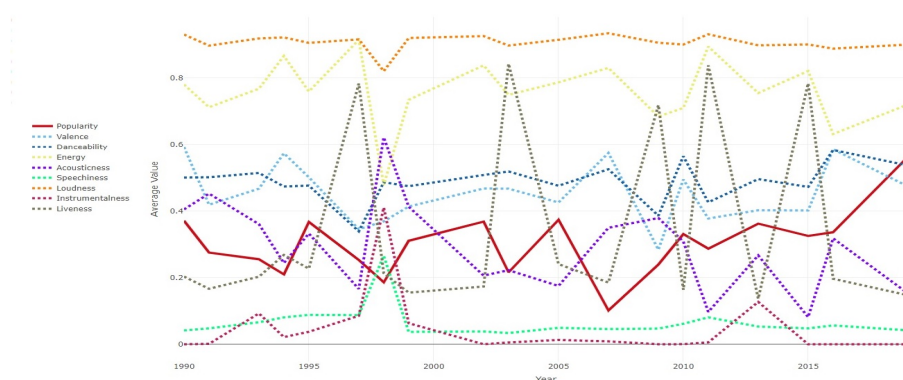


**Fig. 1** Average values of Spotify features over time.

In terms of popularity, it follows that, after reaching the bottom point of popularity compared to the entire career, in 2007, there is an increase in popularity which reaches ever higher values until it reaches its absolute maximum in 2019 with "*Start*". The musical characteristics do not show any particular trend; it is as if the musical genre is constant over time, except for the peaks of Speechness of 1998 given by "*Radiofreccia*". It is important to note that the peak in Speechness is accompanied by a corresponding reduction in Energy, Danceability and Loudness. In addition Speechness seems to have a turning point in 1998: the Speechness index for the years prior to 1998 is always higher than the Speechness index of the subsequent years. From the plot it is possible to distinguish between all the live albums, in which the values of the homonymous index reach abnormal peaks, and then assume values around 0.2 in the case of non live album. Among all the features, the Loudness and the Instrumentalness of the songs seem to be those more stable within the entire discography of Luciano Ligabue.

Focusing on the popularity over the years, plot in Fig. 2 shows that, on average, the most popular album is **Start**, released on March 8, 2019, followed by the album **Buon Compleanno Elvis** in1995, while the least listened to album is **Primo tempo**, probably because it is a collection. In particular, the most streamed songs are "*Polvere di stelle*" and "*Certe Notti*", while the less popular song is

"*Radiofreccia*", soundtrack of the movie with the same name. This song is only Instrumentalness and not spoken, so it could have an influence on its low popularity.
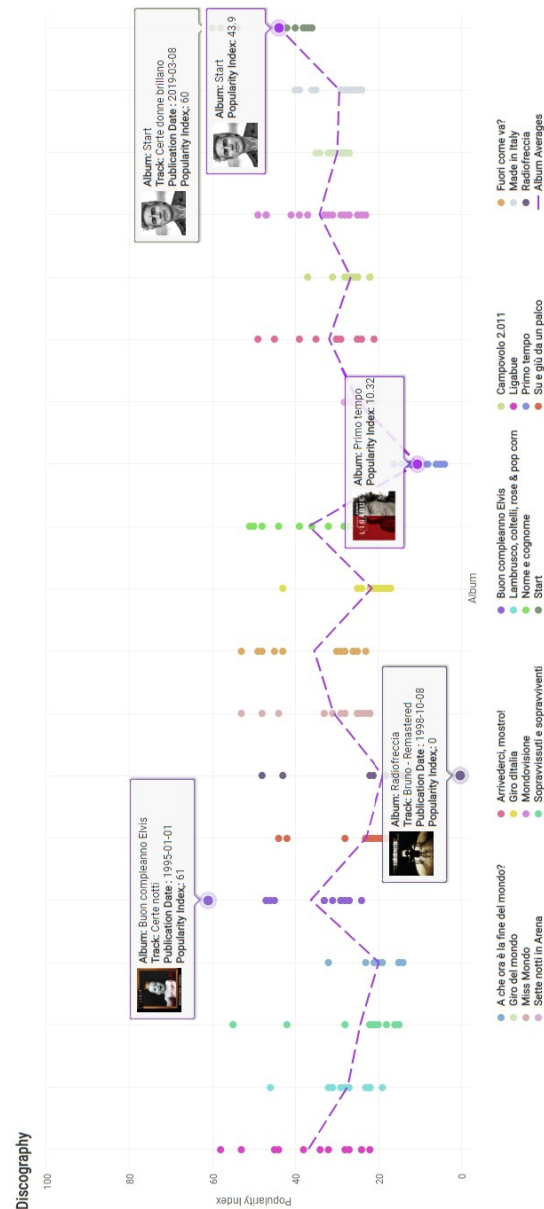


**Fig. 2** Distribution of songs according to the Popularity Index, conditioning to the album (November 13, 2019)

In the light of trends observed in Fig. 1, before fitting the model we want to investigate if there are correlations among these variables in order to avoid problems of multicollinearity. Scatterplots in which the individual audio characteristics are related with the popularity index of the songs are shown in Fig. 3 and Fig. 4.
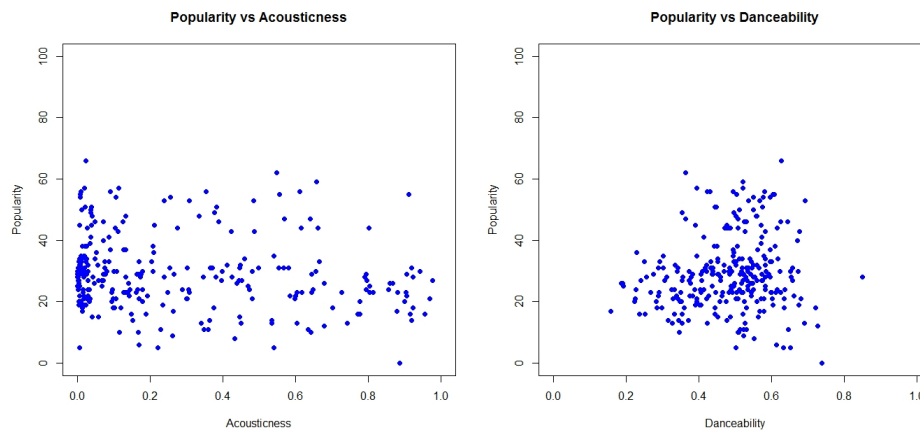
**Fig. 3** Scatterplot of Popularity vs. Acousticness and Danceability.
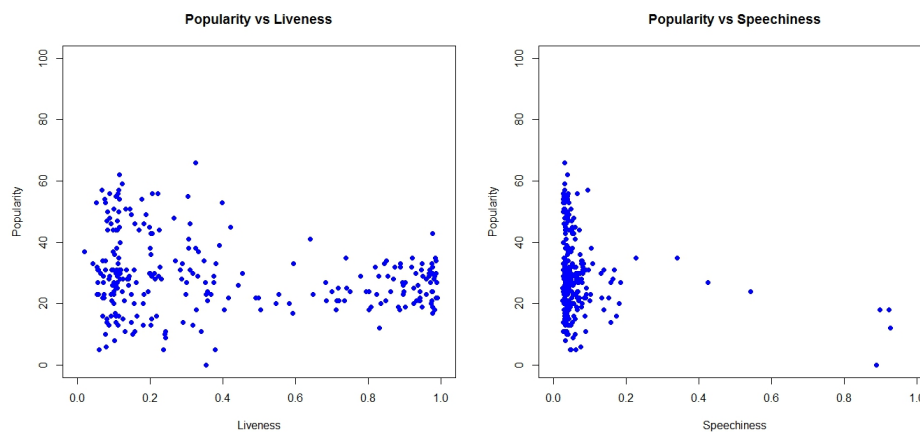


**Fig. 4** Scatterplot Popularity vs. Liveness and Speechness.
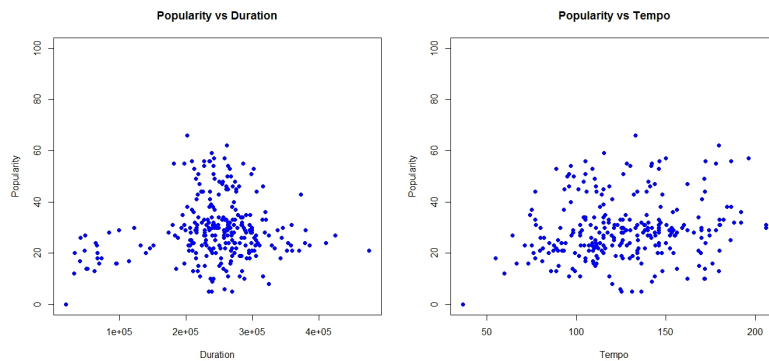
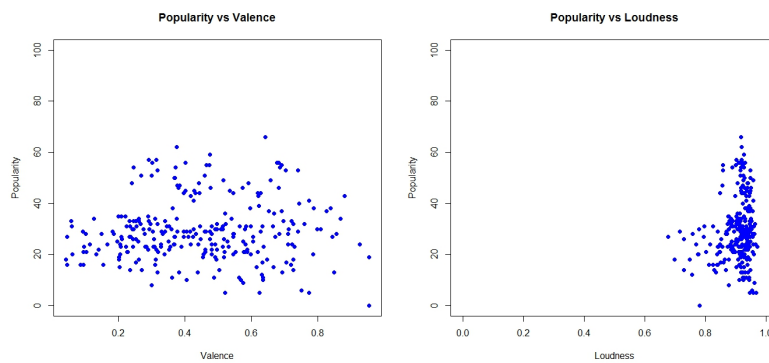**Fig. 5** Scatterplot Popularity vs. Duration and Tempo.



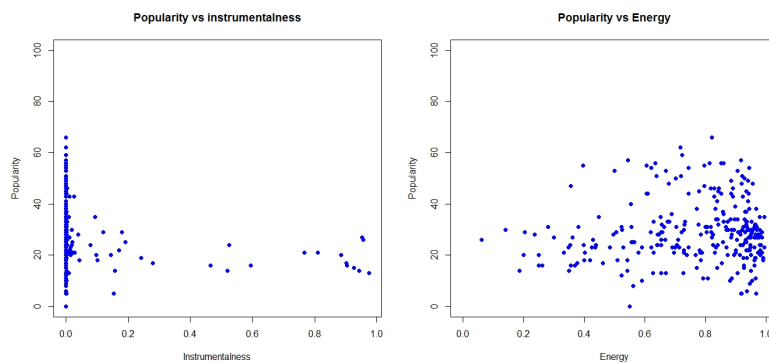**Fig. 6** Scatterplot Popularity vs. Valence and Loudness.



**Fig. 7** Scatterplot Popularity vs. Instrumentalness and Energy.

Scatterplots show that most of the songs in Luciano Ligabue's discography are not very Speechness, not very Instrumentalness, with a lot of Energy and have Loudness values close to zero. In particular, there is a downward trend in Popularity compared to Live, Speechness, Instrumentalness and Acoustics, i.e. as these audio characteristics increase, Popularity decreases. The Valence, Rhythm and Danceability charts do not show any particular trend, while Energy seems to have a positive trend,

i.e. Energy increases and the Popularity of the song increases as well.

In order to identify which are the musical characteristics influencing popularity, and if this relationship is similar for all albums (and therefore over time), a *Beta model with random effects* has been estimated, since the available data has a cluster structure given by the music albums, and it has been assumed that the response variable is distributed according to a Beta distribution, as the popularity index is a continuous and limited variable in an interval the $(0, 1)$. In addition, since to the presence of exact zeros, the index is rescaled in the bounded interval.

Through the *glmmTMB* library (Brooks et al., 2017), different models are estimated and based on the *Akaike Information Criterior* (AIC), the selected model is considered as the right compromise between parsimony and goodness of fit; the results of the "chosen" model are shown below:

**Table 1** Selected Beta GLMM

| **Random effects** | | | | | |
|---|---|---|---|---|---|
| **Groups** | **Variance** | **Std. Dev.** | | | |
| Album name (Intercept) | 0.1926 | 0.4389 | | | |
| | | | | | |
| **Fixed effects** | | | | | |
| | **Coeff.** | **Std. Error** | **z value** | **Pr($>$ \|Z\|)** | |
| *Intercept* | -1.18120 | 0.21596 | -5.470 | 4.51e-08 | *** |
| *Speechness* | -1.52742 | 0.33042 | -4.623 | 3.79e-06 | *** |
| *Instrumentalness* | -0.34554 | 0.18122 | -1.907 | 0.05655 | . |
| *Liveness* | -0.27380 | 0.12229 | -2.239 | 0.02516 | * |
| *Duration* | 0.08419 | 0.03061 | 2.751 | 0.00594 | ** |
| *Energy* | 0.66948 | 0.26992 | 2.480 | 0.01313 | * |
| *Valence* | 1.16800 | 0.47151 | 2.477 | 0.01324 | * |
| *Energy:Valence* | -1.46232 | 0.58209 | -2.512 | 0.01200 | * |
| **Overdispersion parameter for the beta family** | | | | $\phi = 38.7$ | |

The results of the selected model show that Speechness, Instrumentalness and Live are the features that negatively affect the Popularity Index, while Energy, Valence and Duration of the song are the ones that positively affect it. It should also be noted that the interaction between Energy and Valence has a particularly negative effect on the considered index.

Table 2 shows the changes, in percentage terms, in the Popularity Index of songs, for unitary changes of the audio features.

**Table 2** Variations in % in Popularity Index

| Audio features | Variation in Popularity Index |
|---|---|
| Speechness | -32% |
| Instrumentalness | -9% |
| Liveness | -7% |
| Duration | 2% |
| Energy | 16% |
| Valence | 26% |
| Energy:Valence | -31% |

Through the *Caterpillar plot* let's try to understand whether the clustered structure in is relevant for obtaining better, in inferential terms, predictions.
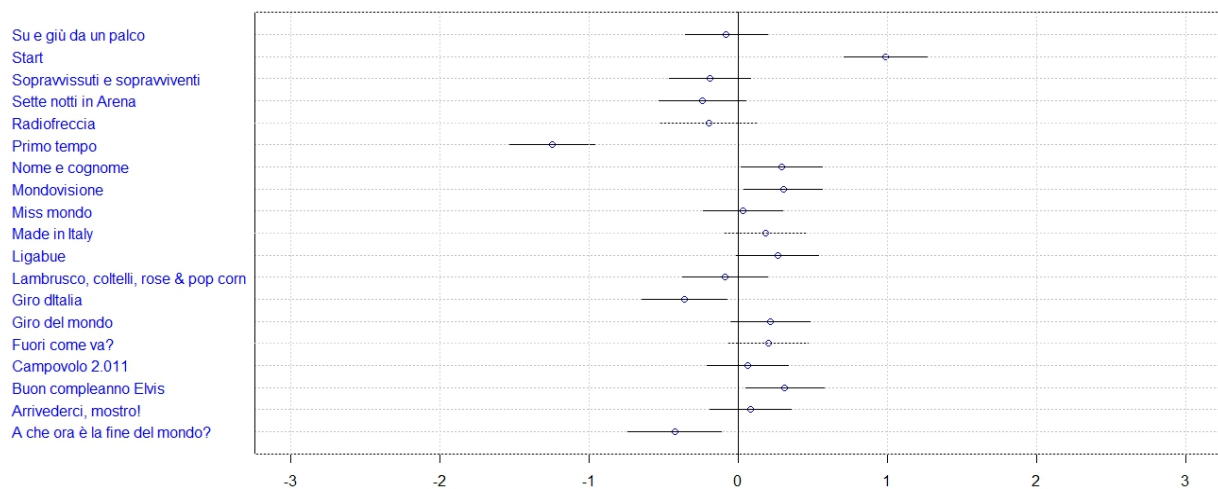


**Fig. 8** Random intercepts with 95% confidence intervals.

Fig.8 confirms that the random part in the model cannot be ignored, as there is a cluster effect due to the album entitled **A che ora è la fine del mondo?**, **Buon Compleanno Elvis**, **Giro d'Italia**, **Mondovisione**, **Nome e Cognome**, **Primo tempo** and **Start**; for these albums estimate confidence intervals do not intersect the zero, this means that the hypothesis of homogeneity of the albums is rejected because the albums are different from each other. This is also confirmed by the estimate of the variance of random effects which is 0.1926 and therefore not negligible.

## 6 Conclusions

In this work a new class of models for dealing with songs popularity index has been introduced. The use of a Beta GLMM allows to account for clustering structure of data from music album and results on the real case example show that this structure cannot be ignored. The Spotify Web API audio features, used as covariates, have shown that not all the Spotify characteristics have a high explanatory power for a higher stream count but some of them are actually important. Significant relationships were found, which lays a promising foundation for the research in prediction with these variables. In particular, Speechness, Instrumentalness and Live are the features that negatively affect the Popularity Index, while Energy, Valence and Duration of the song are the ones that positively affect it.

This research contributes to further understanding in the field of HSS and the new product success prediction. Creating effective prediction models is an interesting next step to this research and so next step would be to expand on the variables used. We hope that this paper will have practical implications also on Spotify, suggesting for example interesting ideas to further develop its database with the hope that data of increasing quality can lead to interesting discoveries and added value to the world of HSS.

## References

Bonat, W. H., P. J. Ribeiro, and W. M. Zeviani (2014, Aug). Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics 42*(2), 252–266.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal 9*(2), 378–400.

Charlie, T. Spotify for developers. `https://www.rcharlie.com/spotifyr/`.

Dhanaraj, R. and B. Logan (2005). Automatic prediction of hit songs. In *ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings*, pp. 488–491.

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics 31*(7), 799–815.

Giles, D. E. (2006). Superstardom in the us popular music industry revisited. *Economics Letters 92*(1), 68 – 74.

Lee, J. and J.-S. Lee (2018, Nov). Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia 20*(11), 3173–3182.

Lerch, A. (2018). *The Relation Between Music Technology and Music Industry*, pp. 899–909. Berlin, Heidelberg: Springer Berlin Heidelberg.

Lovison, G., M. Sciandra, A. Tomasello, and S. Calvo (2011). Modeling posidonia oceanica growth data: from linear to generalized linear mixed models. *Environmetrics 22*(3), 370–382.

Middlebrook, K. and K. Sheik (2019). Song hit prediction: Predicting billboard hits using spotify data.

Nasreldin, M. (2018). Song popularity predictor. https://towardsdatascience.com/song-popularity-predictor-1ef69735e380.

Ni, Y., R. Santos-rodríguez, M. Mcvicar, and T. D. Bie (2015). Hit song science once again a science?

Nijkamp, R. (2018, July). Prediction of product success: explaining song popularity by audio features from spotify data.

Pachet, F. (2011). Musical metadata and knowledge management. In *Encyclopedia of Knowledge Management*.

Poggini, M. (2009). *Liga: la biografia.* Bur.

Sloboda, J. A. (2011). Music in everyday life: The role of emotions. In P. N. Juslin and J. Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press.

SpotifyWebAPI (2019). Spotify for developers. https://open.spotify.com/.