A frontier production function estimation using panel data from music chart artists

April 13, 2023

Contents

List of Figures

List of Tables

1 Introduction

The aim of this research is to investigate the production function of German chart artists using daily Spotify streaming data and musical characteristics from their song catalogue. There is no surefire formula for becoming a superstar, as success in this industry often depends on a combination of factors such as talent, timing and luck. However, there are a number of things that can increase the chances of producing a successful song and with the increasing technological ability to analyse music, it is possible to quantify the impact of certain variables on popularity. Understanding the determinants of success is of great interest to independent artists and labels alike, as it can help reduce the risk of releasing unprofitable albums or singles. According to empirical estimates by Denisoff (1986), only one in ten albums released by artists under contract is profitable. Furthermore, the extraction of musical features and the estimation of success parameters will be necessary in an ever-changing industry that is increasingly driven by technology. Technology already plays a significant role in the production and distribution of music, and this trend is likely to continue in the future. For example, music could be created using artificial intelligence. The basis of this analysis is a huge dataset of the daily Spotify streaming performance and song catalogue variables for all weekly top 100 German chart artists from 2015 to 2023. In numbers, the panel dataset includes 1013 artists (N) with an average of 2226 observed days (T). Using a panel ARDL model, I was able to quantify the short- and long-term impact of different song related variables on streaming growth. By implementing this model in a frontier production function, the technical efficiencies are estimated and heterogeneities of artists and labels are considered.

It is important to note that an artist's income depends on several sources, including income from physical and digital sales, live performances, merchandise sales and music streaming. As it is difficult to collect data on artists' total income and all the factors that influence different types of income, this study focuses only on streaming income and the factors that influence it. Streaming income is not only the largest source of income for many artists, but also a proxy for overall success. The paper is structured as follows: Section 2 presents a literature review of studies based on panel data frontier production functions and discusses the theoretical background. Section 3 explains the market structure. Section 4 provides an overview of the data used, together with descriptive statistics and graphs. Section 5 introduces the developed panel autoregressive distributed lag (ARDL) model and the used frontier production function framework. Finally, Section 6 presents the results and implications.

2 Literature review

2.1 Production function estimation

As an important tool in economics, a production function describes the maximum output that can be obtained from a given set of inputs, taking into account technological and other factors. It provides valuable insights into the contribution of different input factors to the output, as well as the marginal product of an input. In recent years, studies in the field of hit-song science (HSS) have attempted to predict whether a song will become a hit using musical characteristics and machine learning methods. Most of the papers (Herremans et al.,2014; Georgieva et al, 2018; Middlebrook & Sheik, 2019) in the field of HSS focused on the classification and prediction problem, but paid little attention to the effect of the variables.

The estimation of production function parameters has been addressed for a wide range of industries, including manufacturing, agriculture, services or even sports. In his pioneering work, Scully (1974) used an average production function for US baseball teams, and more recent studies have used stochastic frontier analysis to estimate the efficiency of football sports teams (Ruggiero et al., 1996; Dawson et al., 2016). But published work modelling output in the music industry and the creative industries as a whole is rare. This is particularly true for frontier production functions. Among the limited number of studies that have examined the input-output relationships in the music industry, most have analysed the charts and focused on single target variables.

The determinants of success can be roughly divided into socio-economic factors and musical factors. This concept is explained in detail in section four. In particular, the existence of skewness and inequality in the distribution of artists' earnings and the effect of previous success on the success of recently released songs or albums have been discussed in previous studies (Fox & Kochanowski, 2004; Strobl & Tucker, 2000; Towse, 1992). Rosen (1981) describes the superstar phenomenon as the existence of a relatively small group of people who earn enormous amounts of money and dominate activities. Rosen's explanations for the existence of superstars focused on the star's talent and economies of scale. Empirical evidence for Rosen's assumption that even small differences in talent have a large impact on success has not been found for popular music (Hamlem, 1991), but talent is difficult to measure or evaluate empirically (Krueger, 2005). Adler (1985) argues that superstars emerge due to the snowballing effect of positive network externalities, and Towse (1992) also points out the existence of bandwagon effects in the music industry, initially developed by Leibenstein

(1950). Strobl and Tucker (2000) examined how chart success not only reflects previous earnings but also acts as publicity itself. In a study, Pinheiro & Dows (2009) found that jazz musicians with both high social and cultural capital are more successful.

In a parametric survival analysis, Asai (2008) investigated the effect of non-musical factors on the chart position of Japanese hit songs. They regressed the dummy variables "company" (to identify whether a song is produced by a major company), "genre" (whether the song's genre is J-pop), "star" (if the artist's sales have reached a certain value) and "multi-artist" (if the album is a compilation). According to the empirical estimates, the star effect is significant positive. The major label effect is insignificant for singles and positive for albums. Other survival analyses also confirm the positive effects of initial popularity and major label involvement (Bhattacharjee etal, 2005), or provide evidence that instrumental recordings and recordings by female solo artists have significant longer lives at number one (Giles, 2007).

Interiano et al. (2018) trained random forest models on both chart and non-chart songs and found that successful songs differ significantly from the average songs in a number of aspects. They compiled a large dataset of audio features from Echonest and socio-economic variables. Some of the patterns observed are that successful songs are happier, less sad, more danceable and less relaxed than average songs. McKenzie et al. (2020) used Spotify charts and OLS to examine how collaboration in music contributes to demand, and found that songs with other artists generally outperform songs without featured artists. In addition to artist or song related factors, external factors can also influence the popularity of songs. Pettijohn et al. (2009) examined the relationship between song characteristics of the Billboard Top 100 charts and US economic conditions and found that popular songs are longer, slower, more meaningful, more comforting and more romantic when social and economic times are relatively threatening.

The results of these studies suggest that success depends on both socio-economic and musical factors. Studies already provide a detailed insight into the factors behind success, but ever-changing consumer tastes and trends require constantly updated analysis. Although the enormous amount of additional data in recent years has opened up the possibility of comprehensive data-driven analyses, some factors may remain unobserved. A panel data approach can address this issue.

2.2 The concept of technical efficiencies

Before the 1960s, the production function (PF) was usually estimated using traditional least squares, and the resulting PF was an average function that did not take into account firm-specific heterogeneities. The seminal paper by Farrell (1957) stimulated the empirical estimation of production functions with technical efficiency (TE) of individual firms. The introduction of a TE term, also known as non-allocative inefficiency, allows to explain why output varies across firms given the same set of inputs. According to Farrell's concept, TE is considered as the deviation from the most efficient use of inputs, the "production frontier". Productivity is the input-output ratio, while efficiency is the productivity ratio using the most productive unit as a benchmark. The two terms are therefore closely related. However, the existence of TE of firms is highly controversial because TE could also be considered as an effect of omitted variables and their exclusion represents a misspecification of the model. As Müller (1974,p.731) pointed out, "...once all inputs are taken into account, measured productivity differences should disappear except for random disturbances. In this case the frontier and the average function are identical. They only diverge if significant inputs have been left out in the estimation." I will discuss this point for the case in this analysis later. When it comes to the econometric modelling of frontier production functions, there are two basic concepts: The deterministic FPF and the stochastic FPF.

2.3 Deterministic frontier model

The deterministic Frontier model is of the form:

$$y_i = f(x_i, \beta)exp(-u_i),\ u_i \geq 0$$

Where y_i is the maximum production level of the i-th firm, and bounded above by the deterministic quantity $f(x_i, \beta)$. $f(x_i, \beta)$ is a function of the vectors x_i (inputs) and β (parameters) to be estimated and represents the production frontier. u_i is a non-negative random variable (i.e. half-normal distribution) which is associated with the firm specific inefficiency. For the most efficient firm, u_i is equal to 0 and the realised output is equal to the production frontier $f(x_i, \beta)$. Since the production frontier is simply the output of the most efficient firm, all observations must lie on or below the frontier. Given the frontier output of the most efficient i-th firm $y_i^* = f(x_i; \beta)$, the TE is the ratio of $\frac{y_i}{y_i^*}$:

$$TE_i = \frac{y_i}{y_i^*} = \frac{f(x_i, \beta)exp(-u_i)}{f(x_i, \beta)} = exp(-u_i)$$

The predicted TE are obtained by the ratio of the observed production output to the corresponding estimated frontier output $\hat{TE}_i = \frac{y_i}{f(x_i, \hat{\beta})}$. While the straightforward TE estimation in the deterministic model is desirable, a disadvantage is its sensitivity to random events or outliers, which influence the shape of the frontier and consequently the estimated TE values.

2.4 Stochastic frontier model

The (cross-sectional) stochastic frontier PF model is independently proposed by Aigner et al. (1977) and Meeusen and Van den Brook (1977). This model uses a two-part composed error term and is defined by:

$$y_i = f(x_i, \beta) exp(v_i - u_i),\ u_i \geq 0, v_i \sim N(0, \sigma_v^2) \quad (1)$$

where v_i is a two-sided random error (statistical noise) and is assumed to be independently and identically distributed. While in the deterministic model, the production frontier only contains of the function $f(x_i, \beta)$, the production frontier in the stochastic model is bounded above by $f(x_i, \beta) + exp(v_i)$ and is in itself stochastic, including the random error. This allows the frontier to vary randomly across firms and time periods. In other words, each firm has its own frontier which is determined by factors beyond the firm's control such as climate or strikes. The inefficiency is the same as in the deterministic model, namely $exp(-u_i)$. However, the obtained TE values differ from those obtained in the deterministic model because the production frontier output will be in any case smaller than in the deterministic case. In summary, the three approaches differ as follows: The original OLS average function assumes an inefficiency term of zero ($u_i = 0$), and the error term of the average function follows the same symmetric distribution as v_i. In contrast, the deterministic approach assumes that any deviation from the frontier is due to inefficiency (meaning $v_i = 0$). The stochastic approach, on the other hand can explain deviations from the frontier by random events (v_i), and by inefficiencies (u_i). Figure (1) shows the input-output space with the $TE = \frac{y_i}{y_i^*}$ and the differences between these approaches in terms of their distributional properties (Deterministic frontier on the left and stochastic frontier on the right).

2.5 Stochastic frontier models with panel data

Separating TE from random noise requires strong assumptions about their distribution (iid. $v_i \sim N(0, \sigma_V^2)$, $u_i \sim$ half-normal). Furthermore, the TE measures, although unbiased, cannot be estimated consistently, and cross-sectional

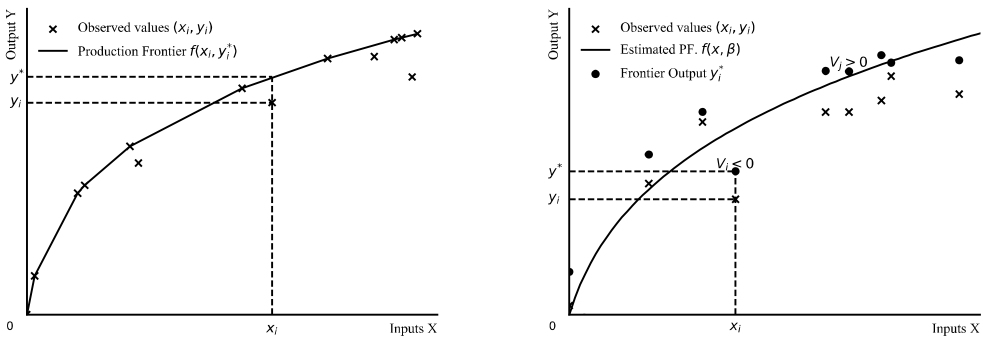


Figure 1: Technical-efficiency in input-ouput space.

models assume that the inefficiency term and the input terms are independent. All of these three difficulties using stochastic frontier models for cross-sectional data can be avoided in the presence of panel data. By considering either fixed effects (FE) or random effects (RE), no distributional assumptions need to be made about u_i and both estimators are constistent for $N, T \to \infty$ (Schmidt & Sickles, 1984). For the frontier case, no further adaption is required. The general stochastic frontier production function, as applied to panel data and first considered by Pitt & Lee (1981), is defined by:

$$y_{it} = f(x_{it}, \beta) exp(v_{it} - u_{it}) \tag{2}$$

with $i = 1, ..., N$ firms and $t = 1, ..., T$ time periods. By taking the natural logarithm of Equation (2), we derive the linear form (Schmidt & Sickles, 1984):

$$y_{it} = \alpha + X_{it}^{'} \beta + v_{it} - u_i \tag{3}$$

We can Furthermore rewrite the model for the FE as follows:

$$y_{it} = \alpha_i + X_{it}^{'} \beta + v_{it}, \; with \; \alpha_i = \alpha - u_i \tag{4}$$

The FE model can be estimated using the common least squares dummy variables (LSDV) procedure or within transformation. In case of a LSDV model, one can simply replace the intercept and add a dummyvariable a_i for every firm. In the case of big datasets with a large N, this method is impractical, and computational limits may reached. Therefore the within estimator is used in most FE approaches and the N intercepts are recovered as the means of the residuals for each firm ($\hat{\alpha_i} = \frac{1}{T_i}\sum \hat{\epsilon_{it}}$). For the RE model, estimation using generalized least squares (GLS) or maximum likelihood (ML) estimation is possible. The RE/GLS specification is derived by the following reformulation. First, let $E(u_i) = \mu$ and define

$$\alpha^* = \alpha - \mu,\ u_i^* = u_i - \mu \tag{5.1}$$

Then in the model

$$y_{it} = \alpha^* + X_{it}^{'}\beta + v_{it} - u_i^* \tag{5.2}$$

both error terms v_{it} and u_i^* have zero mean. By defining $\alpha_i = \alpha - u_i = \alpha^* - u_i^*$, the model becomes:

$$y_{it} = \alpha_i + X_{it}^{'}\beta + v_{it} \tag{5.3}$$

As the FE estmimators, the RE/GLS estimator does not require any distributional assumptions about u_i. In contrast to the FE estimator, the RE estimators assumes, that u_i is uncorrelated with the regressors. If specific distributional assumptions regarding v_{it} u_i are made or known, then ML estimation is feasible, and in the case of correct distributional assumptions, the ML estimator will be the most efficient estimator (Schmidt & Sickles, 1984). Some studies used the ML estimation, by assuming specific distributions such as normal and half-normal in the case of Pitt and Lee (1981) or normal and truncated-normal in the investigation by Battese and Coelli's (1988).

A considerable disadvantage of the within estimator relative to the LSDV and RE/GLS estimator is the impossibility of including other time invariant, but cross sectional heterogeneities. As a consequence, the effects of these omitted time invariant regressors will reappear and contaminate the TE measures. However, if these heterogeneities are observed, they can be considered in the LSDV as $Z_i^{'}\mu$ and the model as written in Greene (2005) is:

$$y_{it} = X_{it}^{'}\beta + Z_i^{'}\mu + v_{it} - u_{it} \tag{6}$$

More problems arise regarding unobserved heterogeneities, even for the LSDV and RE/GLS estimators. Greene (2005) pointed out, that "the treatment of the 'effects' in this models as the inefficiency per se neglects the possibility of other time invariant, unmeasured heterogeneity that is unrelated to inefficiency." To conclude with Greenes words, "Heterogeneities are either absent, included in the model, or absorbed in the TE".

In the panel data case and if N is large, Schmidt and Sickles (1984) proposed to estimate inefficiencies u_i by simply normalizing the firm effects α_i and assuming the largest intercept corresponds to 100% efficiency.

$$\hat{u_i} = \hat{\alpha} - \hat{\alpha_i},\ \ with\ \hat{\alpha} = \max(\hat{\alpha_i}) \tag{5}$$

The panel data models can be further divided into models with timevariant and timeinvariant effects. Early panel data models assumed time invariant effects, which is a questionable assumption for a long panel with large T. It is likely that the TE change over time, which will result in a model where α_{it} depend on both i, t:

$$y_{it} = \alpha_{it} + X_{it}^{'}\beta + v_{it}, \;\; with\, \alpha_{it} = \alpha - u_{it} \tag{6}$$

Cornwell et al. (1990) considered the intercept α_{it} as a function of time with a constant, a time-trend, and a time-squared parameter. They proposed a parametric (deterministic) approach of the TE effect as follows:

$$\alpha_{it} = \alpha_{i1} + \alpha_{i2}t + \alpha_{i3}t^2 \tag{7}$$

The firm specific level of TE needs to be estimated for firm i and time t. Lee and Schmidt (1993) provide a model in which the TE are defined by a product of the individual TE and time effects. They proposed the TE to be

$$u_{it} = \delta_t u_i \tag{8}$$

,where δ_t are the time effects, represented by time dummies (e.g. Month, Day, Year) and the u_i can be individual fixed or random TE effects.

The main advantage of the first is, that it allows efficiency and its functional time form to vary across every firm. This is done by estimating all three parameters for every firm. However, if this deterministic form is not considered realistic or the true time pattern is more complex, the latter model proposed by Lee and Schmidt (1993) can be used. The non-parametric approach is more flexible regarding the functional form, but it assumes the same time effects for each cross-section.

Finally in the alternative model offered by Batesse and Coelly (1992), the structure of the TE to follows an exponential function of time:

$$u_{it} = [exp(-\eta(t - T))]u_i$$

Depending on the estimated time-varying parameter (η), the model allows the TE to be constant ($\eta = constant$) or increase (decrease) monotonically over time ($\eta = positive(negative)$)

3 Market Structure

Streaming revenues contribute a significant share of total income, accounting for 65% of industry revenues worldwide and even more in Europe and the US (Inter-

	FE		RE	
Estimator	LSDV	Within	GLS	ML
Distr. Assumptions about u_i and v_i	no	no	no	yes
Cross sectional X allowed	yes	no	yes	yes
Correlation of TE and X allowed	yes	yes	no	no
Most efficient	no	no	no	yes

Table 1: Comparison of different estimators.

	N	Mean Streams	Median Streams	Skew	VC[1]
Artists	1013	2842.96	669.45	4.09	1.95
Labels	2468	854.73	50.34	13.01	4.71

Table 2: Descriptive statistics of total streams of artists and labels in millions.

national Federation of the Phonographic Industry, 2022). Spotify and most major streaming services, use the pro-rata royalty-allocation system, where artists receive a fixed amount for each stream of at least 30 seconds from a user. The value of a stream is calculated by dividing the revenue of all subscriptions by the number of streams in a given period, making streams a direct proxy for streaming revenue. Songs released in the past contribute significantly to current streaming revenues, making the production of output in the music industry, and especially in terms of streaming revenues an ongoing process. Therefore, a dynamic panel approach is appropriate, taking into account periods without production (i.e. release of a single/album) but with revenues. For the architecture of the production function, this means that the output is the daily streaming revenue, and the input is a set of regressors, which are described below. The results of several studies mentioned in Section (2) suggest a highly right-skewed income distribution in the music industry, as well as a "winner takes the most" or superstardom phenomenon. This is further supported by the right-skewed income distribution of the artists and labels in this dataset (Table 2). The skewness is an indicator for the superstar phenomenon which appears to exists even more for the labels income distribution. The variation coefficient (VC) describes the inequality. [1]

[1]VC = Variation coefficient = $\frac{\sigma^2}{\mu}$

4 Data

Chart position is a good measure of a song's success, as it reflects its monetary earnings from various sources such as physical and digital sales or streaming. The present study is based on daily Spotify streams of 1013 German charting artists with $25,931$ unique releases from the period of 2015 to 2023. In contrast to previous studies that only consider an artist's charting songs, this analysis includes all of an artist's releases, including singles, EPs, LPs, collaborations with other artists and compilations. Using panel data with artist-level earnings avoid some limitations of a chart-based cross sectional PF estimation using song-level earnings. First, in a song-level chart dataset, many artists have only one charting song, and their non-charting releases are ignored. In contrast, the artist-level panel analysis gives the opportunity to estimate artists unobservable heterogeneities, because every song is included. As the charts only represent songs that are already successfull, some important factors of success may remain undiscovered. Estimates using the charts can only calculate the effect of inputs that influence chart positions but cannot analyse factors that influence the chart entry itself. For instance hip hop songs, have become increasingly popular and now account for a large share of the chart songs. Basically, it can be said that songs with a high presence of spoken words or "speechiness" (which is an easily measurable and hiphop-specific characteristic) are popular. However, a pure chart analysis would not identify this, as almost all songs today, both the top positions and the lower ones, have a high speechiness. To overcome this, one possible solution is to use larger chart datasets that include thousands of chart-positions, although these are typically not available. Another option is to include non-charting songs and turn the regression into a classification problem. As a disadvantage of the panel data analysis with whole artist-level earnings, the separation of a songs individual streams from the aggregated artists streams is impossible.

4.1 Set of regressors

There are several factors that contribute to the music-related income of an artist. A first rough function of the income is the interaction between the amount of output produced and the characteristics of the output produced. Only those who produce output make revenue, which can be increased if the output also has popular characteristics. In this analysis the output always refers to musical output, i.e. the release of a Song, EP, LP (Album) or aswell features with other artists and appearings on mixed artists albums (compilations). The first com-

	Name	Description	Var.
Output	Release	Dummy variable, indicating a release of artist i at time t. Could be a Single, EP , LP etc.	r_t
	Release Tracks	The number of total tracks of the release.	$rnum_t$
	Feature	Release of an other artists song at time T with artist i as a Feature.	f_t
	Feature Tracks	Number of total tracks of the feature release.	$fnum_t$
	Compilation	Compilation release at time T where artist i appears on.	c_t
	Compilation Tracks	Number of total tracks of the compilation.	$cnum_t$
	Frequency	The number of releases in the last 365 days. Controls for the growth effect after a release.	$freq_t$
	Frequency Sq.		$freq_t^2$
	Activity	Captures the additional release success (measured by streaming growth) u_t conditionally on the number of last released songs f_t. A positive coefficient indicates, that the effect of a release is even higher with an increasing number of previous releases.	$r_t * freq_t$
	Activity Sq.	Interaction between release and squared Control. Captures the non linear relationship between release success and release frequency.	$r_t * freq_t^2$

Table 3: Set of regressors: Output.

ponent (total produced output) needs to be examined in detail, as the amount of output can have a non-linear, complex effect on the revenue. A high release frequency can increase the effect of a individual release, as bandwagon effects can also play a role here, but a high release frequency above an optimum can also be harmful. Let the data speak. The second component (characteristics) can be further divided into socio-economic and musical factors. The former include the artist's personality, aura, demographics or economic aspects such as promotion and budget. The latter are the genre and song characteristics observed through data-driven audio analysis. Table (3,4) lists the streaming growth factors at time t, divided into the three components of Revenue: output, socioeconomic and musical. While there may be an infinite number of observable and unobservable factors that influence revenue, this list includes the major factors and serves as a foundation for the models estimated in this study. In the following equations for better readability, I summarize all the Output variables as follows:

$$Output'_{it} = r_t + rnum_t + f_t + fnum_t + c_t + cnum_t + freq_{it} + freq_{it}^2 + r_t freq_{it} + r_t freq_{it}^2 \tag{9}$$

Here $i = 1, ..., N$ indexes artists and $t = 1, ..., T$ indexes time periods. For the estimation of technical efficiencies, the variables of interest are the artists fixed effects α_i. These effects capture the differences in average growth after considering the estimated effects of all other variables. But why do artists have different levels of growth? The stochastic frontier model assumes that all firms (artists) have available the same production technology and differences in productivity can be explained by random disturbances or the ability of some firms to use the production technology more efficiently than others. Applied to the music industry, efficiency could mean to increase the release frequency and of course the musical quality by outsourcing processes to composers, ghostwriters or, more recently artificial intelligence. In addition, efficient artists successfully use technologies such as social media to promote releases and are able to identify and capitalise on emerging trends in music, fashion, and personality.

4.2 Data sources

As a reliable indicator of success in the music industry, the charts are a good starting point for an empirical study of popular German music artists. I first collected the set of chart artists from the weekly German charts, published by GfK Entertainment [3]. The artists' release information aswell as the audio features are obtained from the Spotify API and the daily based Spotify streaming performances are scraped from the "Spotify for artists" website. Only artists with at least 5 releases are considered. Because of the musical diversity and in oder to capture unique music styles, Spotify assign a genre combination of a set from over 5000 subgenres to classify artists. Using the artist unique genre combinations as genre fixed effects will cause problems, because of multicollinearity with the artists' fixed effects. As a solution, I regrouped the over 5,000 subgenres into eight main genres based on their musical similarities, as identified in the every noise online database [4]. Figure 2 displays the genre-distribution. In general y axis is Spiky/Bouncy vs. Atmospheric and x axis organic vs. electronic. Some labels only appear in the charts with one artist, wich causes the same problem as with the subgenres. To avoid this, I filtered for genres that released

[1] Own Estimations: Time stamp of the loudest part of the song

[2] descriptions on https://developer.spotify.com/documentation/

[3] www.offiziellecharts.de

[4] https://everynoise.com/

		Name	Description	Var.
Socioeconomic	Artist	Artist Promotion	Captures all artists specific heterogeneities. Econometricially the artists fixed effects. Some factors may be the artists personality, look, aura or promotional factors like social media activity, social network, interviews and economic factors.	α_i
		Spotify career lenght	A artist individual index variable, indicating the accumulated days of an artists career. Starting for every artist from 0 at the day of the first Spotify stream.	t_i
		Spotify career lenght Sq.	Squared artist career lenght in order to capture non linear efficiency effects.	t_i^2
	Label	Artist Promotion	The labels effect on the artists popularity. For example through a marketing budget, social network, professionality. All factors which affects the artists popularity.	β_j
Musical	Song	Audio Features	Vector of release specific audio features or audio-characteristics: Time Signature (four quarter time), Duration, Tempo, Loudness, Mode, Energetic maximum [1], Audio-features estimated by Spotify [2]:Danceability, Energy, Valence, Explicit Lyrics, Speechiness, Acousticness, Instrumentalness, Liveness	$Audio'_{it}$
		Cooperation	Indicates if a release is a cooperation (Through features with other artists).	$Coop_t$
	Genre	Genre	Vector of artist-level genre dummies: Pop, Dancepop, Edm, Rap & HipHop, RnB & Urbancontemporary & Soul, Rock, Urbano & Latino & Reggaeton, Jazz & Blues.	δ_k

Table 4: Set of regressors: Socioeconomic & Musical.

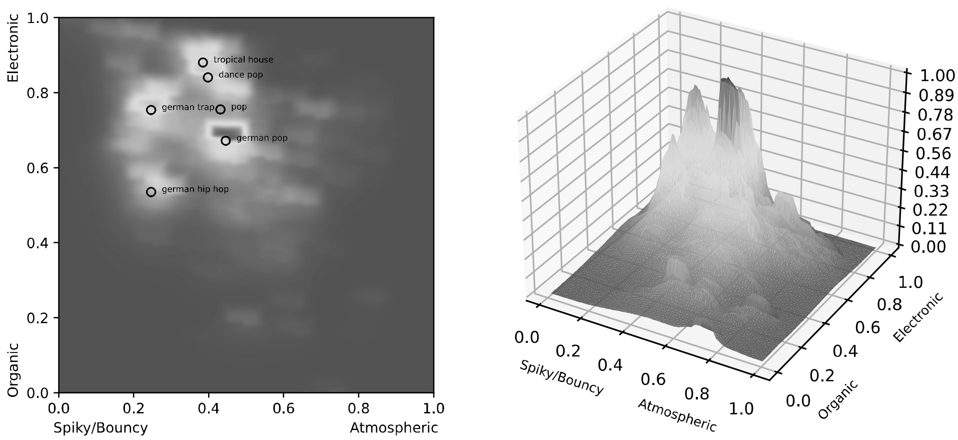


Figure 2: Genre heatmap (left) and genre surface (right).

`Filling technique`	`Variables`
Forward Filling[1]	β_j
Cross Section Mean[2]	$rnum_t, fnum_t, cnum_t, Audio'_{it}, Coop_t$
Zero Filling	r_t, f_t, c_t

Table 5: Missing values filling techniques.

songs with at least 10 different artists. All other labels were grouped together and labelled as small labels. At this point, the artist release information and the release characteristics are mapped to the streaming timeseries on the release day and on all other days we have missing values. These missing values are filled using the techniques in Table (5). Interaction-terms are not listed, because they rely on the missing value techniques of the main variables.

5 Econometric Architecture and Methodology

In the following section, the econometric model is build and the theoretical factors are combined with the observed variables. The focus is on modelling and

[1]Missing values are filled with the last non missing value

[2]Missing values are filled with the cross-section mean value of the Variable.

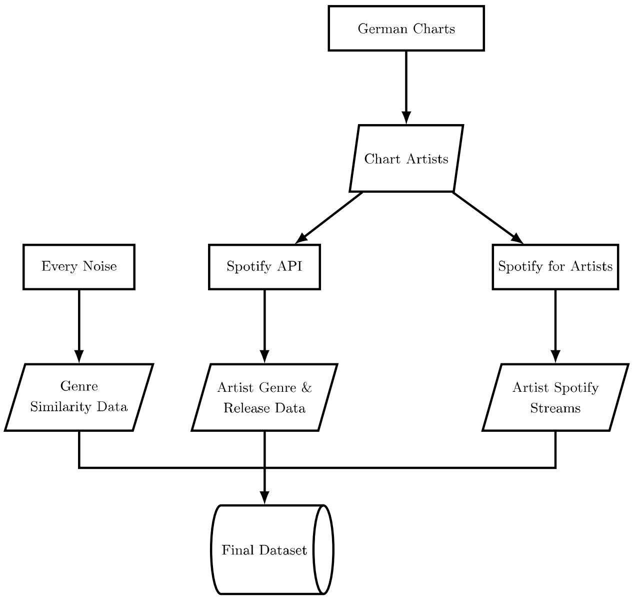


Figure 3: Data generation flowchart. Trapezium = Data, Blocks = Sources.

explaining the streaming revenue as the exogenous variable and investigating individual technical efficiencies. For this task, a stochastic frontier PF approach is appropriate. A first visual observation of the absolute streaming time series indicates non-stationarity (I(1)). The used log-differentiated time series is stationary and represents the growth. The advantages of a panel data set have already been mentioned above, namely the possibility of modelling unobserved heterogeneity and the inclusion of lags. In terms of a comparative analysis, ML estimation drops out, because of the need to make specific distributional assumptions. Because there are likely regressors which are correlated with the TE, the fixed Effects approach is preferred (over the random effects approach). More concrete, I decided to use the FE approach proposed by (Schmidt & Sickles, 1984). Furthermore the existence of time invariant and cross sectional regressors such as label- and genre-effects, which can not be included in the within estimator, forces me to use the FE/LSDV approach, even though this means generating a large number of dummy variables.

5.1 Model 1: Stoch. frontier panel model with time constant efficiencies

First, I introduce the baseline FE panel model without an autoregressive part, distributed lags and heterogeneity in intercepts only. The advantage of this model is its simplicity and the interpretation of the coefficients is straightforward The model 1.1 has the form:

$$y_{it} = \alpha_i + Output'_{it}\lambda + Audio'_{it}\omega + Coop'_{it}\gamma + z_t + \epsilon_{it} \qquad \text{M. 1.1}$$

where y_{it} is the growth rate of daily streams, estimated from the log-differenced absolute Streams. z_t are time dummy variables, indicating each month from 2015-2023 and the weekdays. They capture seasonal shifts and trends affecting the streaming growth of all artists in a given month and weekday. This may be the Spotify's popularity or the growth shifts on specific weekdays (i.e Weekends). In $Output'_{it}\lambda$, r_t, f_t, c_t are dummy variables, indicating if a single/album, a feature track or a compilation is released at time t with a total of $rnum_t, fnum_t, cnum_t$ tracks. The variables are expected to be positive. $freq_{it}$ is the release frequency of an artist i at time t, and the interaction between both r_t and f_{it} is the effect of release frequency on the release effect. A positive interaction coefficient means, that an additional release in the last 12 months increases the release effect. The squared term $r_t freq^2_{it}$ captures the possible non linear relationship. While the growth rate of the streams is positive on the day of a release, the growth is usually negative in the days or weeks following the release as the streams return to an equilibrium. The effects $f_{it}, freq^2_{it}$ capture the following negative growth and are expected to be negative. Without $f_{it}, freq^2_{it}$, the negative growth after a release would be absorbed by α_i, which would therefore be smaller the more uploads the artist had. The variables are explained in detail in the Tables 3 and 4.

Song related exogenous variables are represented by the vector $Audio'_{it}\omega$. α_i are the artists FE. There is no intercept α_0, and all artists are represented by a set of i artist dummies. The artists TE can be estimated with Equation (5), but for the baseline model this seems inappropriate for a lot of reasons. As a first problem, the artists FE contains a wide range of unobserved or not included effects which are not related to the artist's own productivity or technical efficiency. For example, artists with a major label contract are generelly more popular as showed in papers presented in section 2. Furthermore, the artists popularity is affected by external industry-wide musical trends not influenced by the artist itself. The popularity of the artists' genre, affect the artists' popularity. Without additional control variables, the effect of the label involvement and genre trends contaminate the artist's FE. Based on Table (4), artists' FE in Model (1.1) contains all socio-economic variables and the genre effects, which is not a desirable conglomerate.

In order to distinguish the true artists effects, which are displayed in Table (4) (*Socioeconomic* $\rightarrow$ *Artist*), the inclusion of label FE and genre FE is a reasonable next step and transforms Model (1.1) into a multidimensional panel Model (1.2):

$$y_{ijkt} = \alpha_i + \beta_j + \delta_k + Output'_{it}\lambda + Audio'_{it}\omega + Coop_{it}\gamma + z_t + \epsilon_{ijkt} \qquad \text{M. 1.2}$$

The parameter α_i now contains clear artists effects, β_j is the Label effect with j label dummies and δ_k is the genre effect with k genre dummies.

However, it is important to note the relationship between labels, artists and musical trends. Profit-oriented labels choose artists based on their expectations of future popularity, and musical trends play a major role in this equation. If a label has a good intuition (or forecasting quality) about the trends, the labels artists will be popular. This forecasting quality is represented in the label's FE. Since the aim of this analysis is to explore the determinants of the artists' popularity, the label FE should only represent the labels' actions on the artists' popularity, and not the forecasting quality. The addition of a genre FE δ_k can control for this effect by absorbing musical trends. This absorption allows to interpret the labels coefficient exclusively as the labels actions influencing the artists popularity.

Model (1.2) is a good starting point for estimating a frontier production function, but the investigation of short-term effects (the effects only on the release day) is not enough. A significant amount of the streams from releases are realised in the time horizon after the release day. Fortunately, we have time series and the inclusion of lags allows us to estimate the long-run effects of all release-related variables $Output'_{it}, Audio'_{it}, Coop_{it}$.

5.2 Dynamic panel models

By including lags of the dependent and independent variables, Model (1.2) is transformed into a dynamic panel model. More precisely, it is a panel ARDL model. Various methods can be used to estimate long-run relationships in time series. If (1) all variables are $I(1)$ and (2) a cointegration relationship exists, then the error correction model (ECM) can be derived. To determine whether the variables are cointegrated, a Johansen cointegration test, which requires all variables to be I(1), or an ARDL bounds test introduced by Pesaran and

Shin (1999), can be conducted. With only non-stationary (or at least mixed order of integration) variables, the ECM is not an option. The standard ARDL can be transformed into an ARDL-ECM re-parameterisation. Both models, the ARDL and the re-parameterisation model are useful for analysing short-run and long-run relationships. The ARDL-ECM model is estimated with OLS and is applicable to non-stationary (I(1)), stationary (I(0)) or mixed order integration time series (Pesaran and Shin, 1999). The standard ARDL model is also appropriate when all variables are stationary (I(0)), which is the case. To test for a unit root, the panel unit root test of Im, Pesaran and Shin (2003) (IPS) is applied. The IPS test allows for heterogeneity, unbalanced panels and for some, but not all of the individual series to have unit roots. The IPS $\bar{t}$ statistic is the average of the cross-sectional Augmented Dickey Fuller (ADF) test statistic ($\bar{t}_{NT} = N^{-1} \sum_{i=1}^{N} t_{iT}$). The test results indicate that there is no unit root and the null hypothesis is rejected for both absolute streams (s) and stream growth (g) for the 1% W-stat critical value. The distributions of the ADF test-statistics for each artist are shown in Figure 4. The present form of the ARDL is in fact

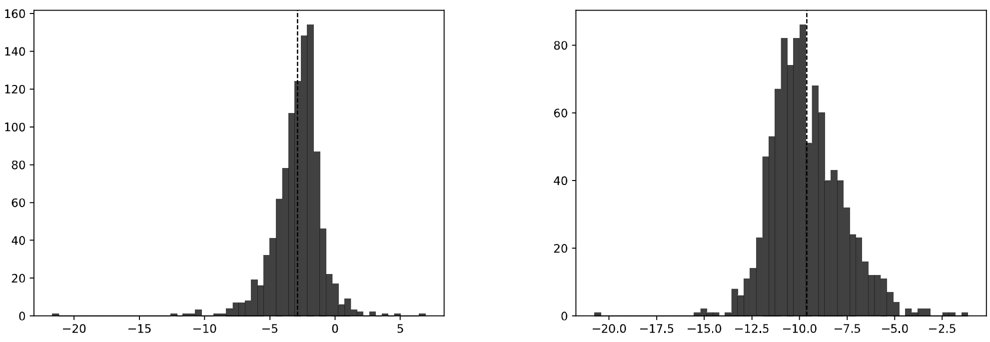


Figure 4: ADF unitroot test-statistics for absolute streams-timeseries (left) and log-differentiated streams growth rates (right).

a short-run model and the temporary short-term effect at time t is dircetly the estimated coefficient. To derive the long-rung impact of a variable, the size and order of the dependent variable must be considered. The long-run cumulative dynamic multiplier is given by:

$$\frac{\sum_{j=0}^{q} \beta_j}{1 - \sum_{i=1}^{p} \phi_i} \tag{10}$$

where β_j are the distributed lags of a variable and ϕ_i is the coefficient of the dependent variable with p lags. Alternatively, the cumulative dynamic multipliers can be estimated directly using a modification of the ARDL model as suggested in Stock and Watson (2020)(eq. 16.7). The accumulation or addition of growth rates, used in the well-known cumulative dynamic multiplier of equation (10) is not appropriate because growth rates (g) must be considered multiplicatively over time:

$$y_{t+q} = y_t(1+g)(1+g_1)(1+g_2)...(1+g_q) \tag{11}$$

or

$$Multiplier = \frac{y_{t+q}}{y_t} = (1+g)(1+g_1)(1+g_2)...(1+g_q) \tag{12}$$

By considering the autoregressive effects, the long run multiplier for growth rates is given by:

$$= g_0 * g_1 * g_2 * g_3 * g_4 * g_5 * g_6 * g_7 * g_8 * g_9 * g_{10} = \prod_{j=0}^{q} g_j \tag{13}$$

whereby g_j is estimated with the recursive equation

$$g_j = (1+\beta_j) * \prod_{n=0}^{j-1} ((g_n - 1)\phi_{|n-j|} + 1) \tag{14}$$

where β_j are the estimated coefficients with q lags and ϕ_i are the estimated autoregressive coefficients with p lags. Detailed explanation can be found in Appendix. A common problem in distributed lag models is autocorrelation of the error term u_t and the regressors, resulting in inconsistent standard errors and misleading hypothesis tests. For this reason, I used heteroskedasticity- and autocorrelation- consistent (HAC) standard errors with a truncation parameter $m = 9$, estimated as suggested by Stock and Watson (2020) with $m = \lfloor 0.75\bar{T}^{\frac{1}{3}} \rfloor = 9$. The inclusion of a dynamic part in a panel model introduces two main econometric complications. First, pooled models (FE/RE) with a lagged dependent variable generate biased parameters because the lagged dependent variable $y_{i,t-1}$ is correlated with the error term ϵ_i. Conventional methods to avoid this error are the instrumental variables (IV) estimator proposed by Anderson and Hsiao (1981) and generalized methods of moments (GMM) estimators discussed in Arellano and Bond (1991). Fortunately, the bias is of order $O(T^{-1})$ (Nickel, 1981) or $O(N^{-1}T^{-1})$ (Kiviet 1995) and can be ignored for sufficiently large panels. Using a Monte Carlo approach, Judson and Owen (1999) find that when $T = 30$, LSDV performs just as well or better than the alternatives. The panel used in this analysis is unbalanced and has the size

$\bar{T} = 2226$, $N = 1013$, why the bias can be ignored, and LSDV/FE estimation is appropriate. However, even when T and N go to infinity there is another problem with traditional pooled estimators in dynamic models because they implicitly assume and require homogeneous parameters across groups. Paresan and Smith (1995) showed that pooled models (such as FE, RE, GMM, IV) can produce inconsistent estimates unless the slope coefficients are in fact identical. As a solution, estimators which impose weaker homogeneity assumptions are employed. On the one extreme, the dynamic fixed effects (DFE) estimator constrains all other coefficients (besides the intercepts) to be the same for all cross sections. At the other extreme, the mean Group estimator (MG) by Pesaran and Smith (1995) allows the long- and short run parameters to vary across groups. As an intermediate estimator, the pooled mean group estimator (PMG) proposed by Pesaran et al. (1999) allows the short-run coefficients to vary freely across groups, but constrains the long-run coefficients to be the same. While the DFE estimator requires conveniently only one pooled estimation, for MG and PMG, separate equations for each group are estimated and in a second step the average parameters are calculated. To avoid the strong assumption of homogeneity of parameters, MG and PMG are the usual choice when a sufficiently large panel is available. In this analysis, separate cross section estimates require not only a large T, but also a hight number of releases per artist. Reliable estimates can only be obtained for artists with a correspondingly high number of total releases, (i.e. 20 releases per artist) which would reduce the number of artists enormously. I decided to use the pooled DFE estimation and assume homogeneous parameters. This decision is supported by the homogeneous nature of the sample, which consists of only successful artists with similar musical genre and comparable fan bases (Figure 2). For example, it's unlikely that the songs duration effects the streaming growth positive for one artist and negative for another. At most, a gradual but not strong or even opposite variation in the parameters is realistic. While Peterson and Berger (1975) argued that the music markets oligopolization will reduce the diversity of popular music, more recent studies (Lopes 1992; Mauch et al. 2015) found no evidence for a homogenization of music. However, the results of Mauch et al. 2015 do not provide evidence that diversity is higher in 2010 than in 1980.

5.3 Model 2: Stoch. frontier panel ARDL model with time invariant efficiencies

Now we can focus on the specific case. In the panel ARDL(p,q) model, every Variable in the vectors $Output'_t$, $Audio'$ and $Coop'$ as well as the dependent variable has lagged components and the model is formulated as follows:

$$y_{ijkt} = \sum_{i=1}^{p} \phi_i y_{t-i} + \alpha_i + \beta_j + \delta_k + \sum_{j=0}^{q} \left(Output'_{it-j}\lambda + Audio'_{it-j}\omega + Coop_{it-j}\gamma \right) + z_t + \epsilon_{ijkt} \quad \text{M. 2}$$

where $\sum_{i=1}^{p} \phi_i y_{t-i}$ is the AR part with p lags of y and variables in the Vectors $\sum_{j=0}^{q} (Output'_{it-j}\lambda + Audio'_{it-j}\omega + Coop_{it-j}\gamma)$ have q Lags. The model is estimated with three different lag lenghts ($p = q = 5, 10, 15$). Of these three models, the model with $p = q = 15$ have the smallest the Bayesian Info Criterion (BIC). Because of the limited computational power required for such large models, Lag lenght $p = q = 10$ is used for the Models (2) and (3).[5].

5.4 Model 3: Stoch. frontier panel ARDL model with time varying efficiencies

The time-invariant technical efficiencies (as used in the Models 1 and 2) assume that the efficiencies does not change over time. This may be the case for panels with relative small time periods, However for larger panels, time variant efficiencies are more realistic, because (1) artist's learn from previous periods and adjust their behaviour and (2) because of industry wide technological changes which affect the efficiencies over time. The Cornwell et.al (1990) model assumes a parametric functional form but allows efficiency to vary across artists and is suitable for the first (1) cause. The intercepts are defined as in Equation (7). In contrast, the Lee and Schmidt (1993) model is flexible in the time domain but restricts the efficiency to vary across artists, and can therefore capture industry wide changes (2). Assuming artist individual temporal variation in efficiency is reasonable and probably more important than industry wide effects. Furthermore, the estimated non-linear artists efficiency parameters provide interesting information about an artists efficiency curve over the career in time. In practice, the discussed three element artist individual time varying intercept require a large number of parameters (three per artists $\alpha_i, \alpha_i t, \alpha_i t^2$). By slightly modify the model and use a cross section invariant parametric solution as already used

[5]BIC values: ARDL(5,5) $= -3.922e+06$, ARDL(10,10) $= -4.099e+06$, ARDL(15,15) $= -4.192e+06$.

in the literature (Cornwell et al., 1990, Table 2), Model (3) can avoid this difficulty. The Model (3) still allows for heterogeneous intercepts α_i but restricts the parametric temporal pattern to be common across artists ($\sum_{n=1}^{4} t_i^n$). The common nature defines the parameters $\sum_{n=1}^{4} t_i^n$ as productivity terms rather than efficiencies. In contrast to Lee and Schmidt's (1993) time (t) based proposition, in Model (3) the temporal pattern is based on the career lenght (t_i). Without the productivity term, the estimated efficiencies in the first three models capture also the productivity. The argument in favour of including the productivity term is that the career lenght should not effect the estimated efficiencies of an artist.

$$y_{ijkt} = \sum_{i=1}^{p} \phi_i y_{t-i} + \alpha_i + \beta_j + \delta_k + \sum_{j=0}^{q} \big(Output'_{it-j}\lambda + Audio'_{it-j}\omega + Coop_{it-j}\gamma\big) z_t + \sum_{n=1}^{4} t_i^n + \epsilon_{ijkt} \tag{M. 3}$$

6 Results and Conclusion

6.1 Discussion of long- and short-run parameter estimates

Table 6 presents the estimated short-term- and multiplier- effects of the four models. The inputs are absolute values and the coefficients are semi-elasticities. The interpretation of the audio features is as follows: an increase from the minimum value (0) to the maximum value (1) leads to a growth by the value of the corresponding coefficient. In all models, the Durbin-Watson test statistic is near 2 (2-2.14) and indicates that the residuals are not first order autocorrelated. According to the Jaque-Bera test, the residuals are not normally distributed. The distribution of the residuals has a skewness between 3 and 4 and a very high kurtosis of 133 to 183, which indicates long-tails. This is mostly because of extreme outliers in the dependent variable that cannot be explained by the model.

It is important to consider lag effects when analyzing the impact of variables on streaming growth. For instance, in Model (3), the effect of danceability at time t is small and insignificant (0.0078), but the effect of the first lag at time t+1 is much larger and significant (0.0196***). Figure 5 illustrates selected variables and their distributed lags structure. The first graph (”ar”) plots the autoregressive lags. Overall, most of the coefficients in each of the competing models are robust and differ only slightly.

The estimated short-run and multiplier- parameters of Model (3) imply that songs that are shorter, more danceable and releases in minor keys with four-

	Dependent variable: streams-growth				$Multiplier_1$	$Multiplier_2$
	(1.1)	(1.2)	(2)	(3)		
r	0.3031***	0.3032***	0.3061***	0.3081***	0.2254	0.3452
rnum	0.0091***	0.0091***	0.0089***	0.0086***	0.0048	0.0060
f	0.1625***	0.1624***	0.1624***	0.1622***	0.1091	0.1647
fnum	0.0063**	0.0063**	0.0062**	0.0065**	0.0055	0.0073
c	0.0611***	0.0609***	0.0629***	0.0027***	0.0019	0.0053
cnum	0.0014***	0.0014***	0.0014***	0.0015***	-0.0007	-0.0008
freq	-0.0006***	-0.0006***	-0.0007***	-0.0007***	-	-
freqsq	1.2e-05***	1.183e-05***	2.09e-05***	1.883e-05***	-	-
activity	-0.0208***	-0.0208***	-0.0211***	-0.0256***	-0.0191	-0.0242
activitysq	0.0004***	0.0004***	0.0004***	0.0005***	0.0003	0.0004
acousticness	-0.0308**	-0.0308**	-0.0324**	0.0278***	0.1080	0.1358
danceability	0.1037***	0.1037***	0.0854***	0.0078	0.0633	0.0901
duration	-0.0246***	-0.0246***	-0.0266***	-0.0295***	-0.0083	-0.0111
energetic-max.	-3.704e-05	-3.7e-05	-2.414e-05	-7.272e-05	-1.87e-04	-2.224-04
energy	0.0231	0.0230	-0.0024	-0.0058	0.0181	0.0280
explicit	0.0813***	0.0813***	0.0845***	0.1684***	0.1116	0.1651
instrumentalness	-0.0885***	-0.0885***	-0.0832***	-0.0174***	0.0221	0.0449
liveness	-0.1402***	-0.1402***	-0.1522***	0.0038	-0.0043	-0.0070
loudness	0.0022*	0.0022*	0.0024*	0.0031***	0.0031	0.0049
mode	-0.0280***	-0.0280***	-0.0298***	-0.0706***	-0.0682	-0.0900
speechiness	0.3117***	0.3118***	0.3101***	0.0111***	0.0195	0.0265
tempo	0.0001	0.0001	4.048e-05	4.297e-05	1.5e-05	3.2e-05
timesignature44	0.0386***	0.0386***	0.0395***	0.0657***	0.0291	0.0404
totalartists	-0.0303***	-0.0303***	-0.0299***	-0.0271***	-0.0117	-0.0157
valence	-0.0704***	-0.0704***	-0.0773***	-0.0098	0.0200	0.0293
trend	-	-	-	-2.32e-06	-	-
$trend^2$	-	-	-	-1.505e-09	-	-
$trend^3$	-	-	-	1.291e-12	-	-
$trend^4$	-	-	-	-2.384e-16	-	-
monday	0.0925***	0.0868***	0.0279	0.0383***	-	-
tuesday	0.0358	0.0300*	-0.0201	-0.0102***	-	-
wednesday	0.0259	0.0201	-0.0020	0.0082***	-	-
thursday	0.0271	0.0214	-0.0099	-0.0005	-	-
friday	0.0713**	0.0656***	0.0239	0.0336***	-	-
saturday	-0.0475	-0.0532***	-0.0432*	-0.0337***	-	-
sunday	-0.1241***	-0.1298***	-0.1047***	-0.0943***	-	-
month dummies	Yes	Yes	Yes	Yes	-	-
artist FE	Yes	Yes	Yes	Yes	-	-
label FE	-	Yes	Yes	Yes	-	-
genre FE	-	Yes	Yes	Yes	-	-
ARDL	-	-	Yes	Yes	-	-
R^2 (Adj. R^2)	0.349(0.349)	0.349(0.349)	0.445(0.444)	0.442(0.442)		
Log-Likelihood:	1.8955e+06	1.8956e+06	2.0735e+06	2.0675e+06		
BIC:	-3.774e+06	-3.774e+06	-4.126e+06	-4.135e+06		
Skew	3.060	3.059	4.098	4.067		
Kurtosis	133.069	133.069	184.714	183.325		
Durbin-Watson:	2.138	2.138	2.003	2.004		
Prob(Jarque-Bera):	0.00	0.00	0.00	0.00		

Note: Observations: 2245130 ,Cov. Type: HAC using 9 lags, *p<0.1; **p<0.05; ***p<0.01
$Multiplier_1$ estimated with equation (13), $Multiplier_2$ estimated with equation (12)

Table 6: Model results.

four beats, acoustic instruments and explicit lyrics have a positive impact on streaming growth. In the "mode" variable, major is represented by 1, minor is 0. Major keys are generally recognized as happy and minor keys as sad. The positive speechiness coefficient (0.0111***) indicates that a high presence of spoken word (i.e. Rap) have a positive effect. These coefficients are robust in each of the competing models and differ only slightly.

As expected for growth rates, an initially significant effect is usually neutralized by an opposite effect in the following periods. The column $Multiplier_1$ shows the estimated multipliers using Equation (13), while the last column $Multiplier_2$ shows the multipliers estimated using Equation (12). These multiplier effects represent growth after 10 periods and are derived from the parameters of Model (3). Demand for music is lowest on weekends (Saturday and Sunday) and highest on Fridays. Releasing music on a Thursday or Friday, as is currently the case in the music industry, seems to be a good strategy. Compared to $Multiplier_2$, the estimated effect of $Multiplier_1$ is generally reduced to zero due to negative autoregressive components.

The release types r, f have a significant positive effect on the streams at the release day. As expected a release r has the largest effect (0.3081^{***}), followed by feature f (0.1622^{***}) and the effect of compilations c (0.0027^{**}) is much smaller. The estimated parameters $rnum, fnum, cnum$ suggest that an additional track on the release increases the effect of r by 0.0086^{***} and f by 0.0065^{**}. Although the effect of an additional track of a release is positive, in terms of total tracks, more small releases are more effective than a few large ones. For example two EP's with six tracks each would have a short-run (long-run) effect of $2 * (r + rnum * 6) = 0.7194\,(0.5084)$ while a LP with 12 tracks would only have a short-run (long-run) effect of $1 * (r + 12 * rnum) = 0.4113\,(0.283)$.

In the case of a cooperation (feature), the estimates indicate that the feature artist can benefit, but the main artist cannot. It depends on whether the artist appears as a feature on another artists track (f), or whether another artist appears as a feature on the artists own track ($totalartists$). In the first case, the artist can benefit by attracting new fans from the other artist and the streams of the own catalogue will increase by 16.22% (10.91%). Contrary to other studies (which only analysed chart songs) mentioned in Section (2), an own song with another artist (second case) performs worse than songs without feature. Each additional artist on an own song reduces the streams on the release day by 2.71% (1.17%). However, the characteristics of the featured artist such as popularity can also play a significant role.

The estimated release-frequency parameters ($activity_t, activitysq_t$) provide

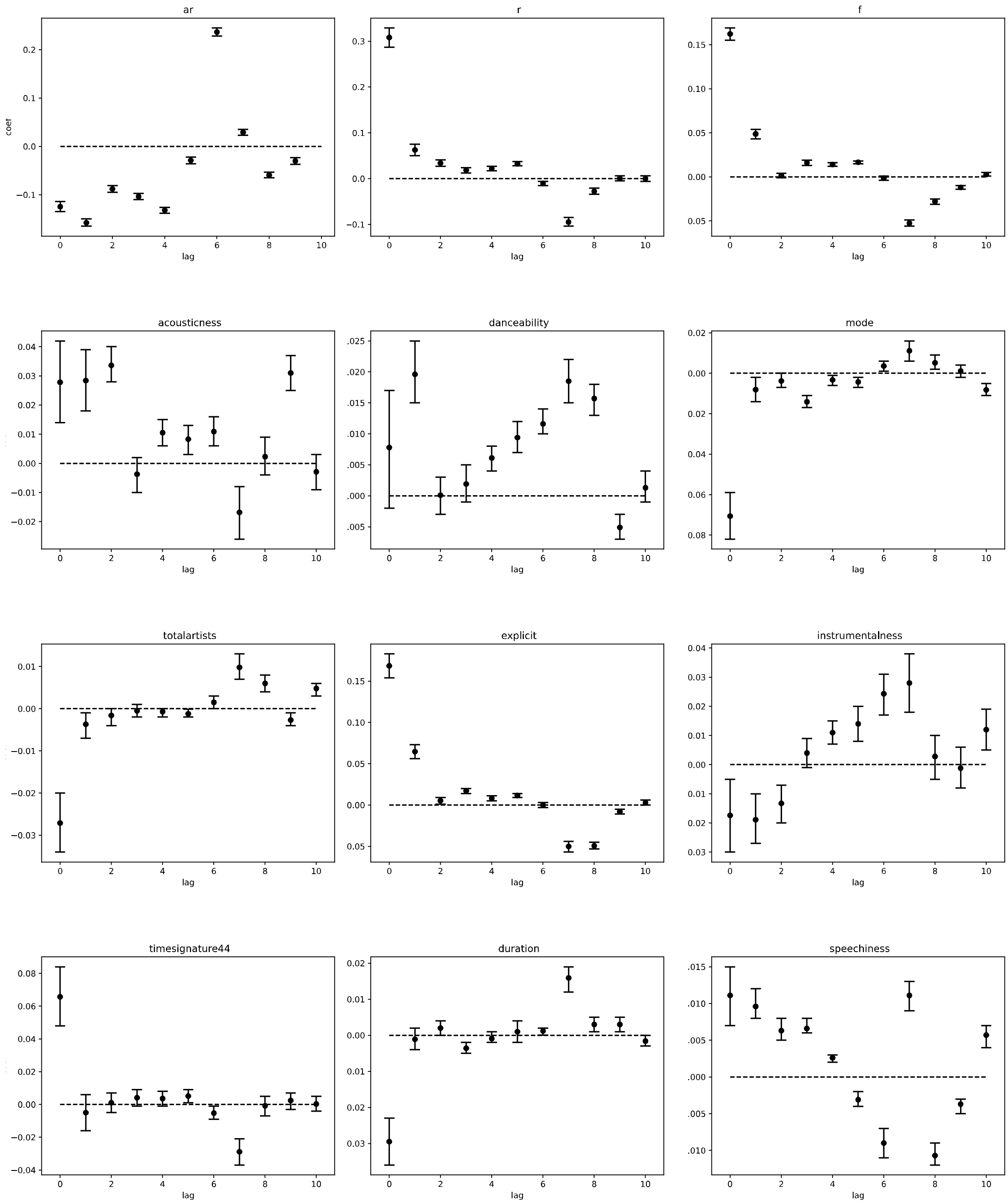


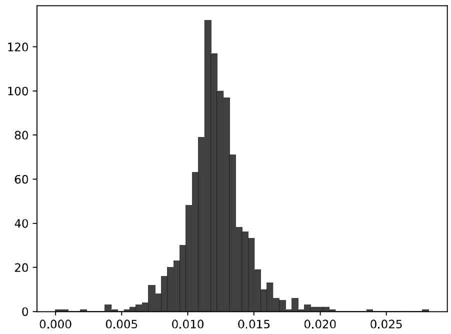
Figure 5: (Lag-)Effects of selected variables.

further interesting insights. The non-linear effect suggests an initial negative effect with a minimum at around 26 (32)[6] releases per year and a break-even at around 51 (64). If the artist has only one release per year, the effect of a release on the streams is more positive than if the artist has up to 26 (32) releases. With a release frequency of 1-26 (1-32), each additional release reduces the release effect. With a release frequency of 26-51 (32-64), each additional release increases the release effect, although the effect is still smaller than with only one release. If the artist has a release frequency of more than 51 (64), the release effect is even higher than with only one release. This pattern can be explained by two opposing effects. (1) When there are very few releases, each release can be marketed more intensively and the streams are more concentrated. As the release frequency increases, the effects are distributed over several releases and the effect of a single release is reduced. (2) On the other hand, a high release frequency triggers bandwagon effects, which exceed the first effect in absolute terms from a release frequency of 51 (64). Streaming services in particular repeatedly point out the importance of a permanent presence and constant releases, which are also considered positive by the recommendation algorithms.

6.2 Technical efficiencies in detail

Efficiency is a relative term and depends on the model specification. It is therefore more appropriate to compare efficiency ranks rather than absolute values. The top twenty most efficient artists are listed in Table 6, the top five most efficient labels are listed in Table 7, and Table 8 reports the calculated Spearman rank correlation coefficients for the artist efficiencies of the models. The rank correlation between the model pairs of (1.1), (1.2) and (2) are all above 82.5%. However, there is less association between the efficiency ranks of the first three models and Model (3), where none exceeds 76.2%. This is due to the time-varying joint productivity term t, t^2, t^3, t^4 in Model (3). The estimated productivity parameters indicate that an artist's productivity (growth rates) decreases over time. In the first three models, productivity is captured in efficiency and therefore the estimated efficiencies are pushed down for artists with long careers. Or the other way around, efficiencies are positively biased for artists with short careers because the usually higher growth rates at the beginning of the career are absorbed in the efficiencies. This is confirmed by the estimated correlation coefficients between efficiencies and career length. The coefficients reveal a strong negative correlation for the first three models and a moderate

[6] $Minimum = \frac{activity_t}{2*activitysq_t}$, Multiplier effects in parantheses.



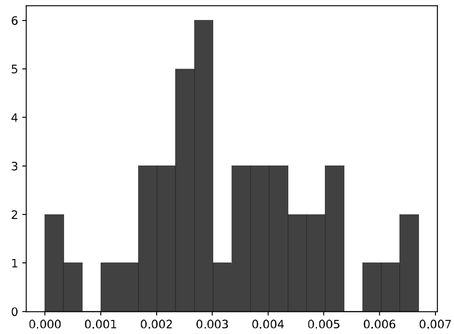


Figure 6: Estimated TE (intercepts) of artists (left) and labels (right) from Model (3).

negative relationship for Model (3).[7] Figure 6 illustrates the distribution of efficiencies of Model (3).

6.3 Conclusion

This study uses a new large dataset of 1013 chart artists to analyse the dynamics of success on streaming platforms (Spotify). This is one of the first studies to estimate the efficiency of artists in popular music. Due to the size of the dataset and limitations of computing power, some methods such as optimal lag length selection, individual time-varying intercepts or non-linear squared effects could not be implemented. The examination of various song characteristics and socio-economic variables has yielded interesting new insights or confirmed the results of previous studies. Releases can have a positive effect on Spotify streams if they are shorter in duration, more acoustic, without a featured artist and with a high presence of vocals and explicit lyrics. The success of an artist depends positively on the release frequency from above 51 (64) releases per year and more small releases (Single/EP) are better than a few larger ones (LP/Album). The results can be used by profit-oriented labels and artists to increase the probability of success in the music industry. The focus of this study is on the calculation of these parameters, but not on the presumed negative impact of this optimisation and technologisation on the quality of art.

[7] Pearson correlation coefficients between efficiencies and career length: (1.1) = -0.178, (1.2) = -0.277, (2) = -0.347, (3) = -0.0164

	Model			
Rank	(1.1)	(1.2)	(2)	(3)
1	DJ Robin	DJ Robin	DJ Robin	Natalie Jane
2	NoD	NoD	NoD	Emmy Meli
3	ARY	ARY	ARY	Romero
4	Natalie Jane	lityway	lityway	Tiësto
5	lityway	Natalie Jane	Natalie Jane	Jnr Choi
6	Emmy Meli	Romero	Romero	NoD
7	Romero	Emmy Meli	Silva	CamelPhat
8	Amir Tataloo	Paula Hartmann	Emmy Meli	Öwnboss
9	John Williams	CIVO	Kordhell	Polo G
10	ACRAZE	Amir Tataloo	MADE	01099
11	Kordhell	BoyWithUke	Paula Hartmann	CIVO
12	CIVO	Silva	CIVO	Maestro Chives
13	Paula Hartmann	01099	BoyWithUke	Fousheé
14	Maestro Chives	John Williams	ACRAZE	Julian Sommer
15	BoyWithUke	Jnr Choi	Jnr Choi	Matt Sassari
16	01099	ACRAZE	Maestro Chives	ClockClock
17	Julian Sommer	Maestro Chives	01099	Bobby Vandamme
18	Jnr Choi	Kordhell	flowerboii	MEDUZA
19	CKay	Bobby Vandamme	Bobby Vandamme	Moby
20	Moby	CKay	Julian Sommer	Joel Corry

Table 7: Twenty most efficient artists by model.

	Model			
Rank	(1.1)	(1.2)	(2)	(3)
1	-	Virgin	BMG Rights Management (UK)	Columbia/B1 Recordings
2	-	Atlantic Records UK	Virgin	BMG Rights Management (UK)
3	-	Columbia/B1 Recordings	Spinnin' Remixes	Musical Freedom
4	-	Spinnin' Remixes	Columbia/B1 Recordings	Epic Local
5	-	recordJet	Atlantic Records UK	Four Music Local

Table 8: Five most efficient labels by model.

Model	(1.1)	(1.2)	(2)	(3)
(1.1)	1			
(1.2)	0.846	1		
(2)	0.825	0.969	1	
(3)	0.631	0.742	0.762	1

Table 9: Spearman rank correlation coefficients of artists efficiency ranks, Model vs. Model.

7 Appendix

Cumulative ARDL long-run multiplier for growth rates: For growth rates, the Effect of an exogenous variable in an ARDL(10,10) should be estimated with

$$= g_0 * g_1 * g_2 * g_3 * g_4 * g_5 * g_6 * g_7 * g_8 * g_9 * g_{10} = \prod_{j=0}^{q} g_j \tag{15}$$

whereby g_j is estimated with the recursive equations:

$$\begin{aligned}
g_0 &= (1+\beta_0) \\
g_1 &= (1+\beta_1) * ((g_0 - 1)\phi_1 + 1) \\
g_2 &= (1+\beta_2) * ((g_1 - 1)\phi_1 + 1) * ((g_0 - 1)\phi_2 + 1) \\
g_3 &= (1+\beta_3) * ((g_2 - 1)\phi_1 + 1) * ((g_1 - 1)\phi_2 + 1) * ((g_0 - 1)\phi_3 + 1) \\
g_4 &= (1+\beta_4) * ((g_3 - 1)\phi_1 + 1) * ((g_2 - 1)\phi_2 + 1) * ((g_1 - 1)\phi_3 + 1) * ((g_0 - 1)\phi_4 + 1) \\
g_5 &= (1+\beta_5) * \prod_{n=0}^{4}((g_n - 1)\phi_{|n-5|} + 1) \\
g_6 &= (1+\beta_6) * \prod_{n=0}^{5}((g_n - 1)\phi_{|n-6|} + 1) \\
g_7 &= (1+\beta_7) * \prod_{n=0}^{6}((g_n - 1)\phi_{|n-7|} + 1) \\
g_8 &= (1+\beta_8) * \prod_{n=0}^{7}((g_n - 1)\phi_{|n-8|} + 1) \\
g_9 &= (1+\beta_9) * \prod_{n=0}^{8}((g_n - 1)\phi_{|n-9|} + 1) \\
g_{10} &= (1+\beta_{10}) * \prod_{n=0}^{9}((g_n - 1)\phi_{|n-10|} + 1)
\end{aligned} \tag{16}$$

g_j can be written in a more compact form as:

$$g_j = (1+\beta_j) * \prod_{n=0}^{j-1}((g_n - 1)\phi_{|n-j|} + 1) \tag{17}$$

where β_j are the estimated coefficients with q lags and ϕ_i are the estimated autoregressive coefficients with p lags.

8 References

Adler, M. (1985). Stardom and Talent. *The American Economic Review*, 75(1), 208-212.

Aigner, D. J., Lovell, C.A.K., Schmidt, P. (1977). Specification and estimation of frontier production, profit and cost functions. *Journal of Econometrics*, 6, 21-37.

Anderson, T.W., Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, 76, 589-606.

Arellano, M., Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58, 277-297.

Asai, S, (2008). Factors Affecting Hits in Japanese Popular Music. *Journal of Media Economics*, 21(2), 97-113, https://doi.org/10.1080/08997760802069895

Battese, G. E., Coelli, T. J. (1988). Prediction of firm-level technical efficiencies with a generalised frontier production function and panel data. *Journal of Econometrics*,38, 387-399.

Battese, G. E., Coelli, T. J. (1992). Frontier production functions, technical efficiency and panel data: With application to paddy farmers in India. *Journal of Productivity Analysis*,3, 153-169.

Bhattacharjee, S., Gopal, R. D., Lertwachara, K., Marsden, J. R. (2007). The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. *Management Science*, 53(9), 1359-1374.

Cornwell, C., Schmidt, P., Sickles, R. C. (1990). Production Frontiers with Cross-Sectional and Time-Series Variation in Efficiency Levels. *Journal of Econometrics*, 46, 185-200.

Dawson, P., Dobson, S., & Gerrard, B. (2000). Stochastic Frontiers and the Temporal Structure of Managerial Efficiency in English Soccer. *Journal of Sports Economics*, 1(4), 341-362. https://doi.org/10.1177/152700250000100402

Denisoff, R.S. (1986) *Tarnished Gold: The Record Industry Revisited.* Transaction Books, New Brunswick.

Farrell, M.J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, 120(3), 253-290. https://doi.org/10.2307/2343100

Fox, M. A., Kochanowski, P. (2004). Models of Superstardom: An Application of the Lotka and Yula Distributions. *Popular Music and Society*, 27(4).

Fox, M. A., Kochanowski, P. (2004). Models of Superstardom: An Application of the Lotka and Yula Distributions. *Popular Music and Society*, 27(4).

Georgieva, E., Sutam, M., & Burton, N. (2018). Hitpredict: Predicting Hit Songs using Spotify Data. *stanford Computer Science 229: Machine learning.* https://cs229.stanford.edu/proj2018/report/16.pdf

Giles, D. E. (2007). Survival of the hippest: life at the top of the hot 100. *Applied Economics*, 39(15), 1877-1887. https://doi.org/10.1080/00036840600707159

Hamlem, W. A., Jr. (1991). Superstardom in Popular Music: Empirical Evidence. *The Review of Economics and Statistics*, 73(4), 729-733.

Herremans, D., Martens, D., & Sörensen, K. (2014). Dance Hit Song Prediction. *Journal of New Music Research*, 43(3), 291-302. DOI: 10.1080/09298215.2014.881888

Im, K.S., Pesaran, M.H., Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, 115(1), 53-74.

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., Komarove, N. L. (2018). *The Royal Society Open Science*, 5(5). https://doi.org/10.1098/rsos.171274

International Federation of the Phonographic Industry (2022). *Global Music Report 2022.* *https://www.musikindustrie.de/fileadmin/bvmi/upload/06_Publikationen/GMR/IFPI_Global_Music_Report_2022-State_of_the_Industry.pdf*

Judson, R.A., Owen, A.L. (1999). Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters*, 65, 9-15.

Kiviet, J.F. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics*, 68, 53-78.

Krueger, A. B. (2005). The Economics of Real Superstars: The Market for Rock Concerts in the Material World. *Journal of Labor Economics*, 23(1), 1-30. https://doi.org/10.1086/425431

Lee, Y. H., Schmidt, P. (1993). A production frontier model with flexible temporal variation in technical efficiency. In H. Fried, C.A.K. Lovell, S. Schmidt (Eds.),*The measurement of productive efficiency: Techniques and applications.* New York: Oxford University Press.

Leibenstein, H. (1950). Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand. *The Quarterly Journal of Economics*, 64(2), 183-207.

Lopes, P.D. (1992). Innovation and Diversity in the Popular Music Industry, 1969-1990. *American Sociological Review*, 57(1), 56-71.
Mauch, M., MacCallum, R.M., Levy, M., Leroi, A.M. (2015). The evolution of popular music: USA 1960-2010. *Royal Society open Science*, 2(150081).
McKenzie, J., Crosby, P., Lenten, L. J. A. (2020). It takes two, baby! Feature artist collaborations and streaming demand for music. *Journal of Cultural Economics*, 45, 385-408.
Meeusen, W., Van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18(2), 435-444.
Middlebrook, K., Sheik, K. (2019). Song Hit Prediction: Predicting Billboard Hits Using Spotify Data. *arXiv e-prints.* DOI: 10.48550/arXiv.1908.08609
Müller, J. (1974). On sources of measured technical efficiency: the impact of information. *American Journal of Agricultural Economics*, 56, 730-738. https://doi.org/10.2307/1239302
Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49, 1417-1426.
Pesaran, M.H., Shin, Y., Smith, R. (1999). Pooled Mean Group Estimation of Dynamic Heterogeneous Panels. *Journal of the American Statistical Association*, 94(446), 621-634.
Pesaran, M.H., Shin, Y. (1995). An Autoregressive Distributed Lag Modelling Approach to Cointegration Analysis *Cambridge Working Papers in Economics 9514*, Faculty of Economics, University of Cambridge.
Pesaran, M.H., Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels.*Journal of Econometrics*, 68(1), 79-113.
Peterson, R.A., Berger, D.G. (1975). Cycles in Symbol Production: The Case of Popular Music. *American Sociological Review*, 40(2), 158-173.
Pettijohn, T. F. II, Sacco, D. F., Jr. (2009). Tough times, meaningful music, mature performers: Popular Billboard songs and performer preferences across social and economic conditions in the USA. *Psychology of Music*, 37(2), 155–179. https://doi.org/10.1177/0305735608094512
Pinheiro D. L., Dowd T. J. (2009). All that jazz: The success of jazz musicians in three metropolitan areas. *Poetics*, 37, 490-506.
Pitt, M. M., Lee, L. F. (1981). The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry. *Journal of Develioment Economics*, 9, 43.64.
Rosen, S. (1981). The Economics of Superstars. *American Economic Review*, 71(5), 845-858.
Ruggiero, J., Hadley, L., & Gustafson, E. (1996). Technical efficiency in Major League Baseball. In J. Fizel, E. Gustafson, & L. Hadley (Eds.), *Baseball economics: Current research*(pp. 191-200). Westport, CT: Praeger.
Schmidt, P., Sickles, R. C. (1984). Production Frontiers and Panel Data. *Journal of Business & Economic Statistics*, 2(4), 367-374.
Scully, G. W. (1974). Pay and performance in Major League Baseball. *American Economic Review*, 64, 915-930.
Stock, J.H., Watson, M.W. (2020). *Introduction to Econometrics*(fourth Edition). Global Edition.
Strobl, E. A., Tucker, C. (2000). The Dynamics of Chart Success in the U.K. Pre-Recorded Popular Music Industry. *Journal of Cultural Economics*, 24, 113-134.
Towse, R. (1992). The Earnings of Singers: an Economic Analysis. in Towse and Khakee (Eds.) *Cultural Economics*, 209-217. Springer Verlag, Berlin.