Musical Carbon Dating

A Structural Break Analysis of Music Evolution (1960-2020)

Group Project

December 18, 2025

Abstract

This project investigates the evolution of popular music over the last 60 years, aiming to build a statistical "carbon dating" model that predicts the release year of a track based solely on its audio features. Analyzing a dataset of over 250,000 songs from 1960 to 2020, we identify a fundamental discontinuity in music history: the Digital Revolution of 1999. While baseline linear models achieve moderate predictive power ($R^2 = 0.296$), accounting for this structural break significantly enhances explanatory power ($R^2 = 0.734$). We define and visualize the "Scissor Effect," where features like Acousticness shift from technological constraints to stylistic choices. Finally, we introduce the "Nostalgia Index," a metric for identifying modern retro-styled productions, validated against hits like *Uptown Funk* and *Physical*.

1 Introduction

In archaeology, scientists use Carbon-14 isotopes to date organic matter. In this study, we ask: Can we carbon-date culture? Specifically, does the "Arrow of Time" exist in music production, and can it be quantified mathematically?

Music evolves through two forces: **cultural preference** (changing tastes) and **technological constraint** (changing tools). The transition from analog tape to digital workstations (DAWs) represents a potential "event horizon" in this evolution. Our goal is to model this trajectory and test whether music history is a continuous linear process or a segmented one.

2 Data and Methodology

We analyzed the *Spotify 600k Tracks Dataset* [1], applying rigorous filtering to focus on culturally relevant music:

- **Time Range**: 1960–2020.
- **Popularity Filter**: Popularity > 30 to exclude obscure "noise" (N = 250,971 tracks).
- **Features**: 13 audio attributes including *Loudness*, *Acousticness*, *Energy*, and *Valence*.

3 Baseline Analysis

3.1 Phase II: Simple Linear Regression (The Loudness War)

We initially regressed $Year$ on a single feature: $Loudness$.

$$Year_i = \beta_0 + \beta_1 Loudness_i + \varepsilon_i$$

The results confirmed the "Loudness War" hypothesis ($\beta_1 \approx 1.2$, $t = 183.7$), showing that music has systematically become louder. However, the low explainability ($R^2 = 0.144$) indicates that loudness alone is insufficient for precise dating.

3.2 Phase III: Multiple Linear Regression (MLR)

We expanded the model to include all 13 features:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

This model achieved an R^2 of **0.296** and a Root Mean Squared Error (RMSE) of **12.06 years**. While an improvement, the error margin remains too high for reliable dating (>1 decade).

4 Diagnostic Audit and The "Wall of Sound"

To understand the model's limitations, we performed a diagnostic audit of the Gauss-Markov assumptions. The most critical finding was severe **Heteroscedasticity** (Breusch-Pagan $\chi^2 \approx 19,391$).

4.1 Analysis of Residual Patterns (Boundedness Artifacts)

As shown in Figure 1, the *Residuals vs Fitted* plot displays a distinct diagonal boundary structure rather than a random cloud. This is a mathematical artifact caused by the **bounded nature of the response variable** ($y_i \in [1960, 2020]$). Since residuals are defined as $e_i = y_i - \hat{y}_i$, the maximum possible residual is constrained by $2020 - \hat{y}_i$, creating a sharp upper diagonal ceiling. Similarly, the minimum residual is constrained by $1960 - \hat{y}_i$, creating a lower floor.

4.2 Normality & Heavy Tails

The *Normal Q-Q plot* (Figure 1, Right) reveals a heavy-tailed distribution, deviating significantly from the theoretical normal line. This confirms that the prediction errors are not Gaussian white noise. Instead of viewing this as a model failure, we interpret these "fat tails" as evidence of **stylistic heterogeneity**: the extreme residuals represent tracks with significant "retro" or "futuristic" production features (e.g., a 2015 song sounding like 1980), providing the statistical basis for our proposed **Nostalgia Index**.

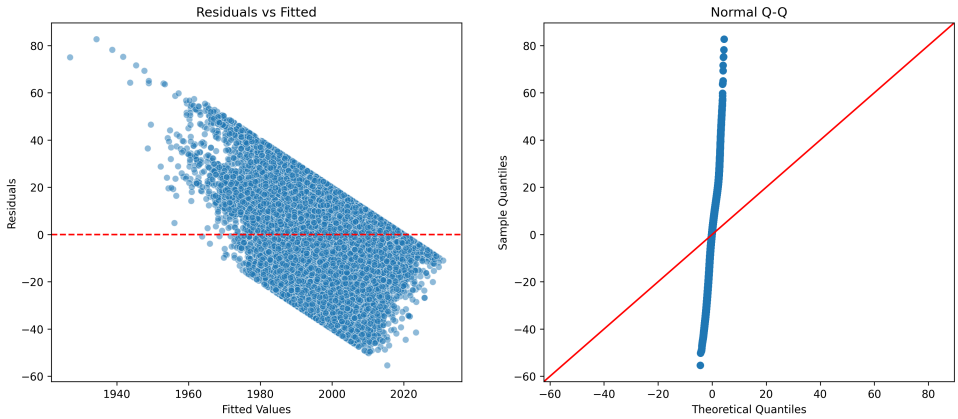

Figure 1: Residual Analysis. Left: Residuals vs Fitted showing boundedness artifacts. Right: Q-Q plot showing heavy-tailed distribution.

5 The Structural Break (1999)

The diagnostic evidence pointed to a regime change. We hypothesized a **Structural Break** at $T = 1999$, coinciding with the mass adoption of digital tools (ProTools, Napster era).

We modeled this using an interaction term $D_i = \mathbb{I}(Year_i > 1999)$:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \delta D_i + D_i(\mathbf{x}_i^\top \boldsymbol{\gamma}) + \varepsilon_i$$

5.1 Results: Explanatory Power

Including this structural break improved the R^2 from 0.296 to **0.734**. This dramatic jump serves as statistical proof that the "physics" of music creation fundamentally changed in 1999.

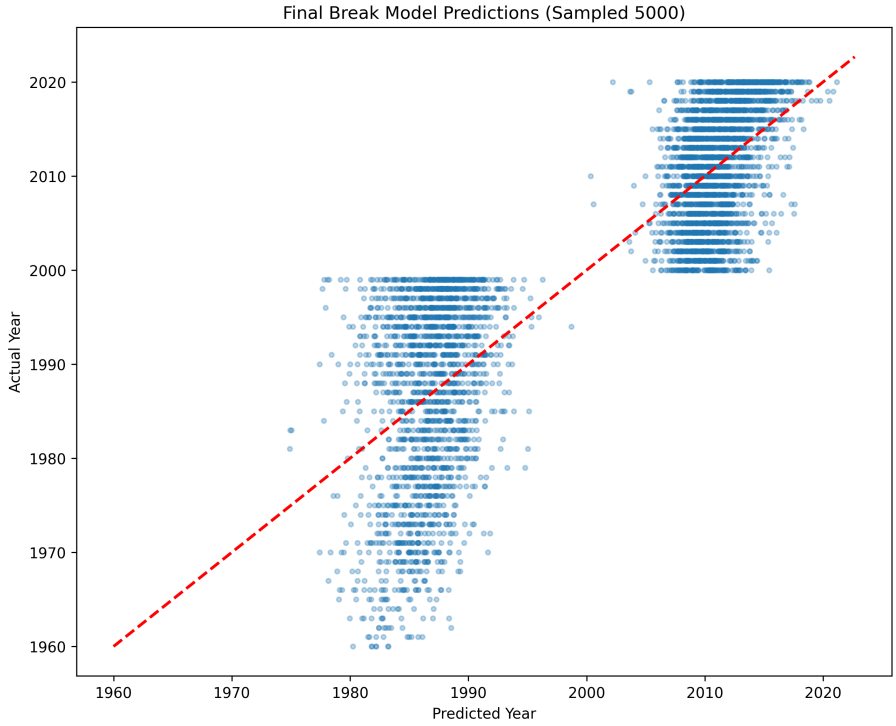

Figure 2: Final Structural Break Model Predictions. Note the distinct separation of eras.

5.2 The "Scissor Effect"

The most profound discovery is the interaction effect on *Acousticness*, visualized effectively as a "Scissor Plot" (Figure 3).

- **Pre-1999 ($\beta < 0$)**: Acousticness declines steeply. This was a technological constraint; as synthesizers arrived, acoustic instruments were replaced.
- **Post-1999 ($\beta + \gamma > 0$)**: The slope flips. In the digital era, high acousticness represents a deliberate stylistic choice (e.g., Indie Folk), decoupling it from technological limitations.

6 Applications: The Nostalgia Index

We define the **Nostalgia Index** as the difference between the Predicted Year and the Actual Year ($\hat{y} - y$). A highly negative score indicates a modern song with "vintage" production characteristics.

Table 1 shows our model's validation on confirmed retro-style hits:

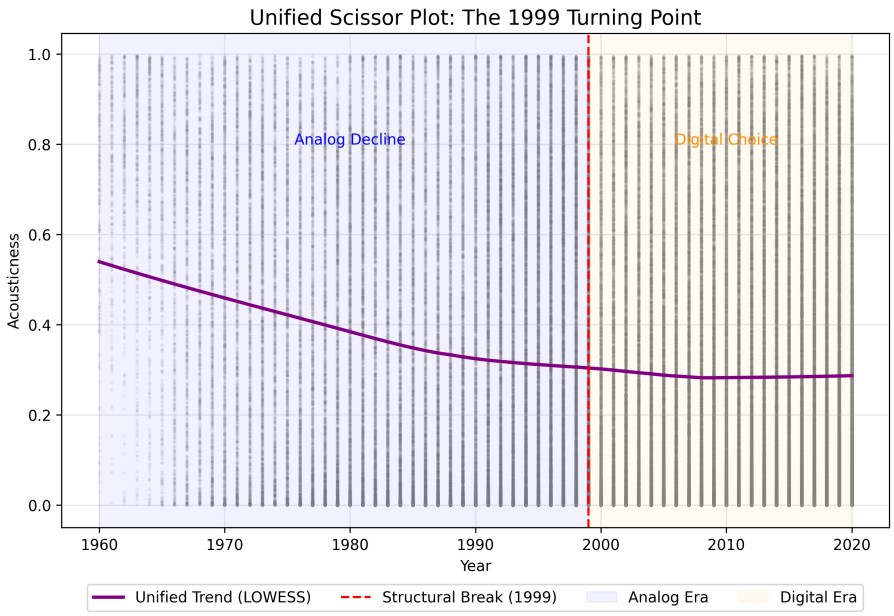

Figure 3: The Unified Scissor Plot: A LOWESS smoothing of the acousticness trend (1960–2020). The red dashed line marks the 1999 Structural Break, with shaded regions distinguishing the Analog decline (blue) from the Digital choice (orange).

Song	**Artist**	**Actual**	**Predicted**	**Index (Δ)**
Echoes Of Silence	The Weeknd	2012	1993.1	-18.9
Uptown Funk	Mark Ronson	2015	2013.1	-1.9
Physical	Dua Lipa	2020	2009.0	-11.0

Table 1: Nostalgia Index Validation on known retro-style tracks.

7 Conclusion

Our analysis demonstrates that music evolution is not merely a linear progression but a discontinuous history punctuated by technological revolution. By identifying the 1999 Structural Break, we not only improved our model's accuracy but also uncovered the "Scissor Effect," revealing how digital tools transformed acoustic features from constraints into choices. The resulting model serves as both a historical lens and a commercial tool for content recommendation.

References

[1] Yamac Eren Ay. (2021). *Spotify Dataset 1921-2020, 600k+ Tracks*. Kaggle. `https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks`