Spotify Streaming Before and During the First Wave of the COVID-19 Pandemic in Germany: Compared and Predicted (ID: 659)

Kework K. Kalustian

Music Dept., Max Planck Institute for Empirical Aesthetics, Germany
kework.kalustian@ae.mpg.de

Background

The COVID-19 pandemic is an event with far-reaching consequences. Several negative consequences are observable. Yet, it is also stateable that people tend to use more musical media during nationwide lockdowns (Fink et al., 2021).

Aims

In this vein, I aimed to answer this question: To what extent was music more streamed during the so-called first wave of the COVID-19 pandemic (March 11, 2020, until June 29, 2020) compared with the same period before the pandemic, and how is daily music listening predictable during the COVID-19 pandemic?

To answer this question, I examined these hypotheses:

- H_1: People tend to stream more of Germany's daily top 200 Spotify charts during the COVID-19 pandemic than in a comparable reference period ($r \geq .10$)
- H_2: The stream counts of Germany's daily top 200 Spotify charts during the COVID-19 pandemic are predictable based on daily stream counts from January 01, 2019, to March 10, 2020 ($MAPE \leq 5$).

Methods

To answer that question, I web-scraped (using R) Germany's daily top 200 chart positions, song titles, their stream counts, and track IDs between January 01, 2019, and June 29, 2020 ($N = 109200$) from Spotify's website.

Testing H_1: I split the entire dataset into two periods to obtain reference values. By using a one-way paired Wilcoxon test, I compared the daily median stream counts of the "pre-pandemic period" (March 11, to June 29, 2019; $n_{\text{Pre-Pandemic}} = 111$) and the "pandemic period" (March 11, to June 29, 2020; $n_{\text{Pandemic}} = 111$).

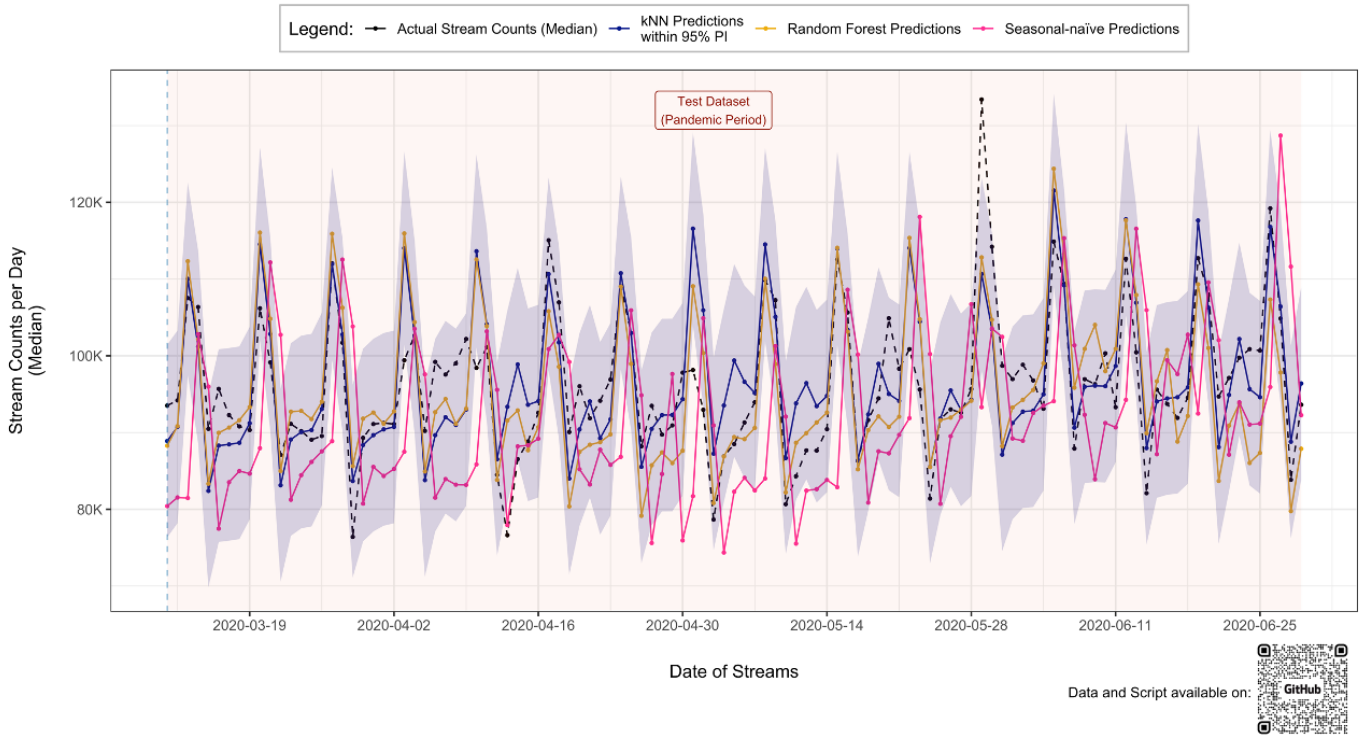
Testing H_2: I defined the prediction task within the scopes of supervised machine learning and time series analysis. Accordingly, I differenced and stationarized the training set (January 01, 2019, to March 10, 2020) to control any trend and seasonal effects. However, the time series was time-dependent (autocorrelations with a lag of 7; equals here to a weekly pattern). Hence, I focused on the time dependence instead of the actual stream counts to achieve reliable predictions at all. By doing so, I embedded the time series of the training set (Takens, 1981) into a lagged matrix to account for that issue. That is, I created a matrix with overlapping time-dependent chunks whose errors (see autocorrelation) were out-cancelable (von Oertzen & Boker, 2010). Simply put, the lagged observations of the time series posed the input variables, whereas the remaining observations posed the predictable dependent variable. Once I prepared the training set in this regard, I run 10-fold cross-validations to optimize hyperparameters of a kNN and a Random Forest model. Then, I trained both models to ultimately predict the 111 daily median stream counts of the test set directly as distinct time steps. Finally, I benchmarked those models against a seasonal-naïve model.

Results

The employed Wilcoxon test showed a significant moderate difference effect in the central tendencies of the stream counts ($p < .001$, $z = -3.61$, $r = .343$, $n = 111$). Hence, people tend to listen to more music during the pandemic. Regarding the prediction task, the kNN model yielded the best results ($RMSE_{\text{kNN}} = 6400.86$, $MAPE_{\text{kNN}} = 5.05$). Therefore, I calculated prediction intervals for the respective estimates. The Random Forest model yielded less acceptable predictions ($RMSE_{\text{RF}} = 6808$, $MAPE_{\text{RF}} = 5.58$). Furthermore, the predictions of the kNN model yielded a 53.2% higher accuracy ($UMBRAE_{\text{kNN:SN}} = 0.468$) compared to the seasonal-naïve model (see Figure 1).

Figure 1

Comparison of actual stream counts, kNN, Random Forest estimates, and seasonal-naïve fits



Note. The purple-colored ribbon represents the 95% PIs based on the SD of the kNN model residuals.

Conclusions

Finally, the data processing and analyzing strategies turned out to be appropriate for examining the research question, for dealing with autocorrelations in the time series, and for using supervised machine learning methods. Accordingly, I update the hypotheses as takeaways:

- H_{1*}: People streamed during the COVID-19 pandemic more of Germany's daily top 200 Spotify charts than in a comparable period.
- H_{2*}: The stream counts of Germany's daily top 200 Spotify charts during the COVID-19 pandemic are predictable based on daily stream counts from January 01, 2019, to March 10, 2020, when using time-delay embedding and direct kNN predictions.

Using these methods to build a date-sensitive music recommender system where, for instance, stream counts of mood clusters are focused can foster research at the intersection of music psychology and data science.

References

Fink, L., Warrenburg, L., Howlin, C., Randall, W., Hansen, N., & Wald-Fuhrmann, M. (2021). Viral tunes: Changes in musical behaviours and interest in coronamusic predict socio-emotional coping during COVID-19 lockdown. *PsyArXiv*. Preprint. 10.31234/osf.io/7mg2v

Takens, F. (1981). Detecting strange attractors in turbulence. In D. Rand & L. Young (Eds.), *Dynamical Systems and Turbulence, Warwick 1980, Lecture Notes in Mathematics*, Vol. 898 (pp. 366-381). Springer. 10.1007/BFb0091924

von Oertzen, T., & Boker, S. (2010). Time Delay Embedding Increases Estimation Precision of Models of Intraindividual Variability. *Psychometrika*, *75*(1), pp. 158-175. 10.1007/S11336-009-9137-9

Keywords: Spotify streaming, Music and COVID-19, Supervised Machine Learning, Time Delay Embedding.