Inaccurate Prediction or Genre Evolution?

Rethinking Genre Classification

Ke Nie

Department of Sociology, UC San Diego

Research Question: Does genre evolution affect MIR-based genre classifier performance?

- Genres are cultural constructs, whose boundary depends on subjective judgments by musicians, audiences, critics, etc.
- Genres may evolve over time as the type of music style associated with a particular genre mutates.
- As genre evolves, genre classifiers trained on songs from different year-cohorts might will impact how the classifier perform on the songs from other year-cohorts.
- If this is true, then we can use genre classifiers to detect genre evolution by looking at change in classifier performance over the years.

Key takeaway: Genre evolution does affect MIR-based genre classifier performance.

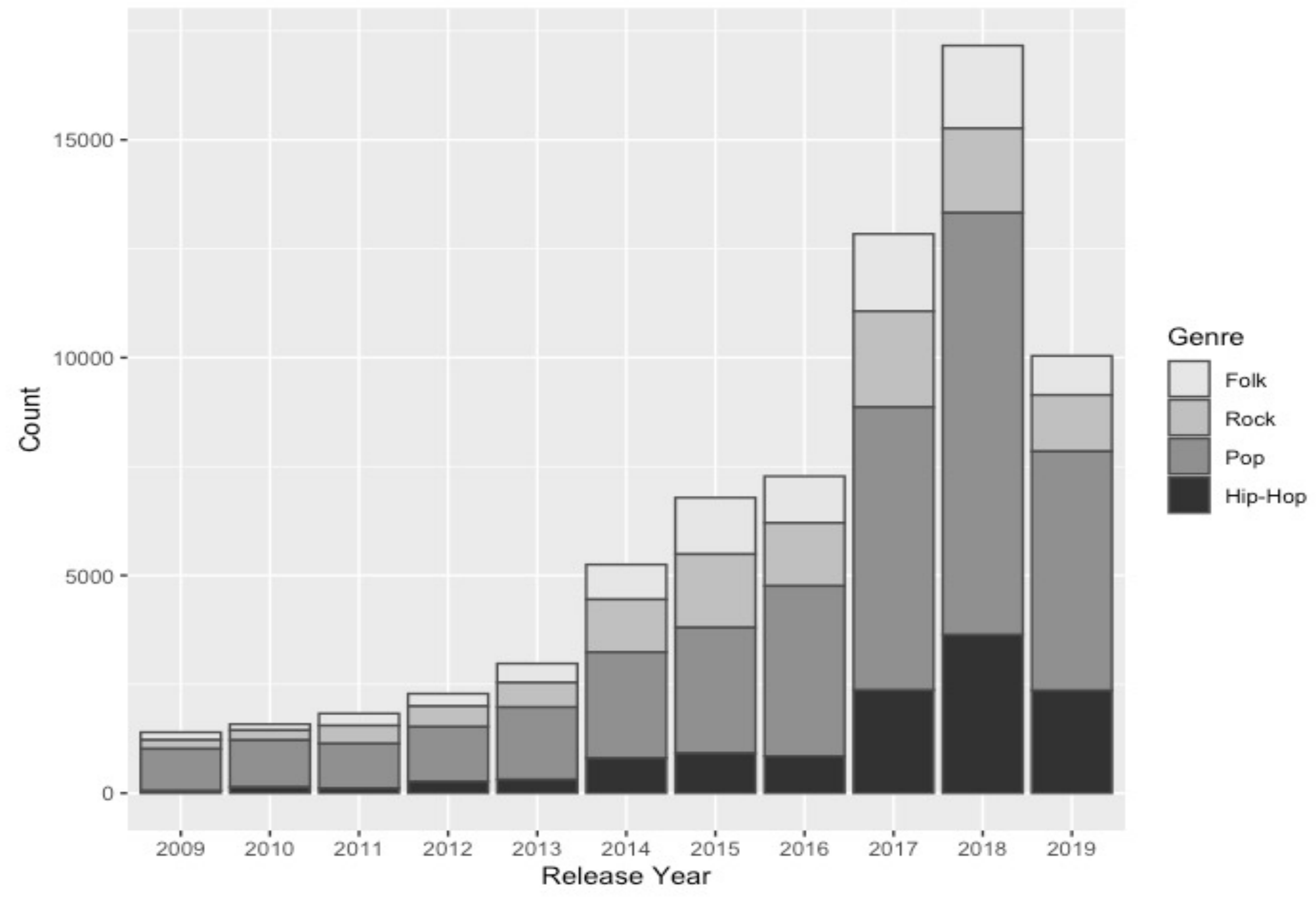
- Genre classifiers trained on older songs do not always correctly predict the genre of newer songs (and vice versa); performance depends on genre evolution including genre-crossing and subgenre salience.
- But this does not mean the classifiers were defectively trained; the drawback can only be spotted post-hoc when new songs are released.
- For the same reasons, we can thus use genre classifiers to detect genre evolution when trained properly. It is difficult, though, to separate genre evolution from flaws in algorithmic design; therefore, it is important to supplement the analysis with more detailed investigations.

Data: 67,427 songs from Chinesemusic.com (anonymized)

- Songs performed by Chinese musicians released on the platform between 2009 and 2019 (until July) from 4 primary genres: Pop, Hip-Hop, Rock, Folk. A song can claim only one primary genres but can claim multiple subgenres.

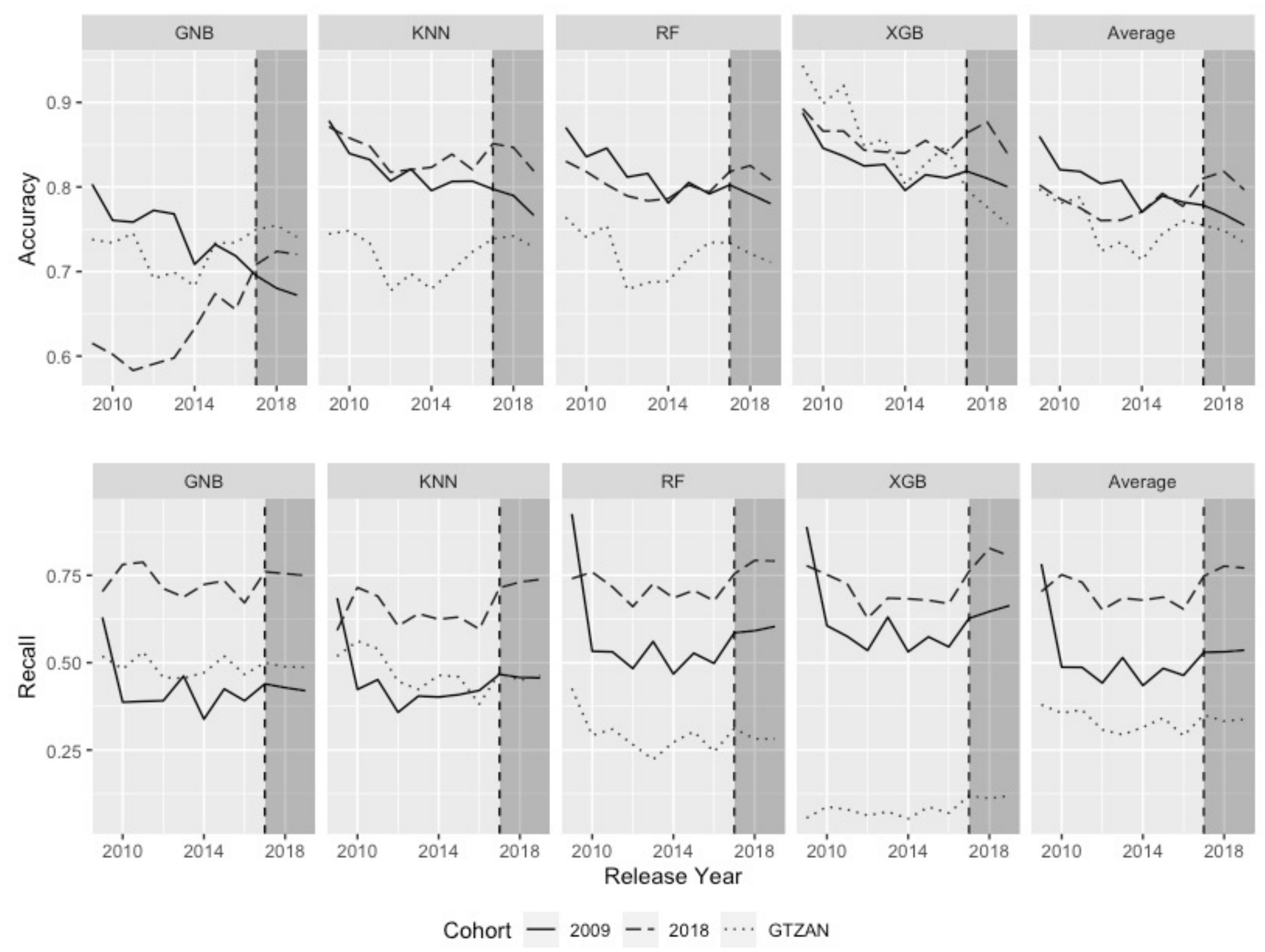
Method: Use genre classifiers trained on songs from different year-cohorts and predict the genre of all songs

- 3 different training sets: 2009 songs, 2018 songs, GTZAN (corrected)
- 4 different machine learning approaches: Gaussian Naïve Bayes, K-Nearest Neighbors, Random Forests, eXtreme Gradient Boosting, plus their average.
- Focus on Hip-Hop as it experienced a series of dramatic events in recent years
- Key metrics: accuracy (correctly predict Hip-Hop vis-à-vis non-Hip-Hop); recall (correctly predict Hip-Hop among all true Hip-Hop songs)



Finding #1: Classifiers have a fuzzy U-shaped performance over the years, particularly on recalls

- The trend of recalls fit a polynomial regression on year and its quadratic term, where all coefficients are statistically significant
- This indicates Hip-Hop deviated from 2009 releases in the middle years but slowly bounced back as the decade concludes



Finding #2: Classifiers perform worse on Hip-Hop-crossing non-Hip-Hop songs

- Hip-Hop-crossing non-Hip-Hop songs are those who claim a genre other than Hip-Hop as their primary genre but also claim subgenres explicitly related to Hip-Hop (e.g., "Rap Rock")
- Two sample t-tests on classifier accuracy and false negative between Hip-Hop-crossing non-Hip-Hop songs and other non-Hip-Hop songs indicate significantly worse performance when there are more genre-crossing songs
- U-shaped recall in Finding #1 is thus partly driven by the fact that there are proportionally more Hip-Hop-crossing non-Hip-Hop songs in the middle years

	GNB		KNN		RF		XGB		Average	
	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN
2009	-0.063*** (0.015)	0.087*** (0.016)	-0.020 (0.017)	0.163*** (0.015)	-0.089*** (0.015)	0.162*** (0.016)	-0.076*** (0.016)	0.181*** (0.016)	-0.062*** (0.008)	0.148*** (0.008)
2018	-0.090*** (0.012)	0.217*** (0.017)	-0.125*** (0.016)	0.270*** (0.017)	-0.131*** (0.016)	0.245*** (0.017)	-0.161*** (0.016)	0.261*** (0.017)	-0.127*** (0.008)	0.249*** (0.008)
GTZAN	-0.079*** (0.011)	0.182*** (0.017)	-0.120*** (0.016)	0.134*** (0.017)	-0.086*** (0.015)	0.263*** (0.015)	-0.081*** (0.014)	0.073*** (0.010)	-0.092*** (0.007)	0.120*** (0.008)

Note: Standard errors are in parentheses. Diff. Acc/FN refers to the difference between Hip-Hop-crossing non-Hip-Hop songs and other non-Hip-Hop songs in terms of the accuracy/False Negative rate of the classifiers in predicting their genre.
*p < .05; **p <.01; ***p<.001

Finding #3: Classifiers perform worse in years where there are fewer songs from salient subgenres in the year in question

- Salient subgenres: subgenres that are robustly represented in numbers in the training set and accurately predicted by the classifier
- Overall pattern suggests a positive relationship between performance metrics and the size of the salient subgenres' proportions. E.g., the classifier trained on 2009 songs perform better when there are more Hip-Hop songs that are Old School Hip-Hop, Instrumental Hip-Hop, Conscious Hip-Hop, Alternative Hip-Hop, or Cloud Rap.

		GNB			KNN			RF			XGB			Average		
		Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec
2009	Top1	OS	0.236* (0.103)	0.625*** (0.095)	OS	0.188* (0.066)	0.636** (0.149)	OS	0.165* (0.067)	1.010*** (0.182)	OS	0.180** (0.050)	0.773** (0.165)	OS	0.154* (0.062)	0.761*** (0.139)
	Top5	OS; I; Con; A; CR	0.143* (0.045)	0.265** (0.075)	OS; I; Con; A; CR	0.123*** (0.021)	0.329** (0.072)	OS; A; U; Pop; I	0.090 (0.065)	0.001 (0.314)	OS; A; U; Pop; I	0.082 (0.058)	-0.049 (0.251)	OS; I; Con; A; CR	0.138*** (0.029)	0.301 (0.168)
2018	Top1	HH	0.211** (0.045)	0.065 (0.264)	HH	-0.001 (0.029)	0.208** (0.055)	HH	0.033 (0.023)	0.149* (0.048)	HH	0.003 (0.028)	0.206* (0.071)	HH	0.071* (0.022)	0.157** (0.047)
	Top5	HH; Pop; T; OS; CU	0.245** (0.058)	0.016 (0.070)	HH; Pop; T; OS; CU	-0.022 (0.034)	0.150 (0.093)	HH; Pop; T; OS; CU	0.245** (0.058)	0.106 (0.074)	HH; Pop; T; OS; CU	-0.013 (0.033)	0.136 (0.109)	HH; Pop; T; OS; CU	0.082** (0.028)	0.102 (0.077)

Note: Standard errors are in parentheses. Among the subtitles, Subg refers to subgenre, Acc refers to accuracy, and Rec refers to recall. Abbreviations in the Subg column: OS for Old-School Hip-Hop; I for Instrumental Hip-Hop; Con for Conscious Hip-Hop; A for Alternative Hip-Hop; CR for Cloud Rap; U for Underground Hip-Hop; Pop for Pop Rap; HH for Hip-Hop; T for Trap; CU for Chinese Underground Hip-Hop.
*p < .05; **p <.01; ***p<.001

Code available at: https://github.com/norzvic/ismir_2022_codes

Contact author at: knie@ucsd.edu