Universidad de
SanAndrés

Big Data - Propuesta de Trabajo Final

Ritmo y recesión:

La relación entre variables macroeconómicas y las preferencias musicales

Renato Albareti, Federico López y Franco Mettola

Fecha de entrega: 5 de julio de 2024

1 Introducción

Este trabajo investiga cómo se relacionan ciertas variables económicas con bienestar de las personas y, en particular, con el bienestar emocional. Para ello, utilizamos la música como reflejo y medición de la felicidad de las personas porque es un bien cultural de fácil acceso y costo bajo. Consideramos que la música popular es un reflejo del bienestar emocional de las personas para tratar de analizar si, por ejemplo, cuando la economía está peor (desempleo alto, inflación alta, entre otras medidas) la gente busca escuchar música más alegre. Por lo tanto, este trabajo propone expandir la literatura sobre la conexión entre la cultura y la economía.

Para hacer más explícita esta relación hipotética, proponemos un modelo microeconómico de preferencias que refleja la suposición expresada anteriormente. Vamos a considerar que el agente deriva un valor de utilidad por consumir una cantidad de música C (que podría ser medido en minutos) que es del tipo M y en una situación económica S.

La variable S es algún indicador económico normalizado al rango $[0, 1]$ tal que 0 sea una mala situación económica. La variable M puede ser discreta o continua y básicamente representa el *sentimiento* de la música que se está consumiendo. Una medida comúnmente usada en esta literatura es la 'valencia', que mencionaremos luego, la cual es continua en $[0, 1]$ y pensada tal que 1 sea muy alegre. Si es discreta puede ser una categoría como el género o ser binaria (eg. $\{\text{triste} = 0, \text{alegre} = 1\}$).

Por ejemplo $U(C = 40, M = 0.9, S = 0.2)$ es la utilidad de consumir 40 minutos de música alegre en un contexto de dificultad económica. La función tendría esta forma

$$U(C, M, S) = C[\alpha M + \beta \cdot (1 - M) \cdot f(S) + B]$$

B es la utilidad base que le proporcionaría escuchar musica sin importar el género o la situación económica. α y β escalan la contribución de la situación económica y del tipo de musica a la utilidad. $f(S)$ es una función de la situación económica. Por ejemplo, puede ser lineal $f(S) = \gamma(1 - S)$, puede ser logarítmica $f(S) = \log(1 + \gamma(1 - S))$, puede tener más parámetros que γ. Lo principal es que sea creciente en S.

La forma de la función de utilidad la pensamos de tal manera que capture el efecto descrito previamente. Si S es bajo (la situación económica es mala), entonces $f(S)$ es mayor, la contribución marginal de escuchar música alegre aumenta. Es decir, la musica alegre es relativamente más valiosa en momentos económicos adversos.

2 Literatura previa

En Anglada-Tort, Masters, Steffens, North, and Müllensiefen (2023), los autores proponen un sistema *Behavioural Economics of Music* o BEM, que busca estimular la investigación en la intersección entre la música y el modelaje del comportamiento de los agentes. Mediante este sistema basado en economía del comportamiento, el BEM sienta bases en la teoría del comportamiento de los agentes para pensar cómo interactúan los factores macroeconómicos con las decisiones relacionadas a la música. ¿Pueden cambiar las decisiones de consumo músical en medio de una recesión? Si bien no hay un contrafáctico para establecer un enlace causal, si podemos utilizar la alta frecuencia de datos disponibles para encontrar una relación de manera descriptiva.

Sabemos que la economía del comportamiento es una disciplina en crecimiento al buscar relajar supuestos de la teoría neoclásica de, por ejemplo, expectativas racionales. De esta manera las políticas públicas pueden empezar a pensar más en la interacción *macro-behavioural* en lugar de las recomendaciones basadas en teoría neoclásica. Esto no quiere decir que las formulaciones anteriores no pueden ser aplicadas, sino que hay lugar para repensar la posición de los agentes en las políticas y cómo afectan las alteraciones macroeconómicas en el comportamiento (Baddeley, 2017).

Otros autores incursionaron en el campo, con la diferencia que relacionan los géneros que ingresaron en el billboard 100 comparando con la inflación, el desempleo, el PBI y la media de la opinión política en Estados Unidos, medida por medio de la encuesta *General Social Survey* (Petitbon & Hitchcock, 2022). También se estudió el tipo de música que se escuchaba en relación al desempleo, la inflación, la tasa de interés y el nivel de intercambio de activos. De esta manera de Lucio and Palomeque (2023) buscan aproximar el ciclo económico y por medio de un *Sentiment Analysis* separan las canciones del billboard 100 entre felices y tristes, encontrando que en tiempos de mala macro, escuchan canciones más felices.

Consideramos entonces que nuestra investigación aporta a la literatura al ampliar el set de variables macroeconómicas; aunque utilicemos algunas que son similares, aplicamos otras que también son del interés de la sociedad argentina. Además, los trabajos previos se centraron en países que, a diferencia de Argentina, experimentan economías estables. De esta manera, consideramos que realizar esta investigación en un país afectado por la volatilidad macroeconómica es un aporte valioso para este campo de estudio.

3 Base de datos

3.1 Bases musicales

Para obtener las canciones que fueron más populares en cada año vamos a tener que utilizar distintas fuentes, ya que, a diferencia de otros países como España o México, solo recientemente se compila un ranking general de las canciones más escuchadas. Estos son los datos que vamos a buscar usando *web-scraping* y la API de Spotify ya que algunas radios los compilaron en esa aplicación: Desde el 1985 tenemos los Ranking de la Rock and Pop. Desde 1991 tenemos el rankings de la radio FM 105.5 (LOS40). Desde 1999 está el Ranking de MTV. Desde 2018, Billboard publica el top 100 Argentina. Desde 2021 CAPIF publica las canciones mas escuchadas de argentina.

Para muchos años hay *overlap* y vamos a tratar de combinar los datos de estas distintas fuentes para obtener las 50 canciones más populares de cada año. De esta manera tendríamos 39 años de datos para utilizar en la segunda etapa. La base de datos sería un diccionario con cada año (key) asociado a una lista de canciones las cuales vamos a referenciar con el `id` de esa canción en Spotify para que sea más fácil referenciar diferentes datos de esas canciones.

Con eso a armar algunas bases de datos: Para la primera vamos a usar la API de Spotify para obtener las características de las canciones. En esta base de datos cada observación es una canción y cada canción va a tener asociadas estas columnas: {`"acousticness": , "danceability":, "duration_ms": , "energy": , "instrumentalness": , "key": , "liveness": , "loudness": , "mode": , "speechiness": , "tempo": , "time_signature": , "valence":`}. Esas son las features principales, pero hay muchas

más que describen el contenido y la estructura musical más en detalle que también podemos usar.

La más importante de estas que vamos a usar es una que cuantifica el `mood` (qué tan negativa o positiva una canción) con la variable `valence` que está entre 0 y 1. La idea es que un valor más alto sería una canción más alegre. Con esta vamos a construir una variable `average_valence` por cada año.

Otra base de datos la vamos a armar con *web-scraping* de páginas web con letras de canciones como genius.com, letras.com y musixmatch.com. Esto nos da más libertad para explorar distintos aspectos de las canciones. Vamos a usar métodos de Natural Language Processing y la librería `nltk` y `vader` para clasificar las letras como positivas o negativas. En terminos del modelo simple que propusimos al inicio, diriamos que M es una función de L, las letras, R, el ritmo, entre muchas otras. De esta manera incorporariamos a distintos tipos de demandantes que ponderan diferente las caracteristicas de la musica que escuchan. En Lugos Abarca (2023) hay una muy completa exposición de todas las variables que componen una cancion que podrian afectar una funcion de utilidad.

Podríamos realizar otros análisis con las letras por ejemplo, también construir un índice de agresividad o de obscenidad y de esa manera ver si, quizás, en los momentos de crisis las canciones más escuchadas expresan más o menos sentimiento de enojo, por ejemplo.

3.2 Bases de indicadores macroeconómicos

Para construir nuestra base, utilizamos la API disponible en el sitio web Estádisticas BCRA donde obtuvimos los siguientes indicadores estimadores detallados a continuación. El resto fueron obtenidos de Datos Argentina

Variable	Descripción	Fuente
PBI	Producto Nacional	Subsecretaría de Programación Macroeconómica
Reservas	Stock de reservas en el BCRA	Estadísticas BCRA
Desempleo	Desempleo en aglomerados urbanos	Subsecretaría de Programación Macroeconómica
MERVAL	Principal índice bursátil nacional	Estadísticas BCRA
Dolar blue	Dólar en el mercado paralelo	Estadísticas BCRA
Pobreza	Tasa a nivel nacional	Subsecretaría de Programación Macroeconómica
Inflación anual	Variación de precios anual	Estadísticas BCRA

Table 1: Variables del set macroeconómico

4 Metodología

4.1 Organizando los datos

Tenemos una impresionante cantidad de características de cada canción gracias a Spotify. Vamos a explorar estas canciones, primero, sin un enfoque supervisado. Esto nos permitiría llegar a la obvia categorización de canciones ya conocida: en géneros.

Hay un módulo de la API Spotify que asocia un genero automáticamente decidido por su algoritmo, pero estos suelen ser extremadamente granulares y nos interesaría dividir el espacio de canciones en 7 u 8 géneros nada más; aquellos que consideramos serían los más importantes y prevalentes en Argentina, que es donde estamos haciendo el análisis. Podrían ser, por ejemplo, Rock, Pop, Folclore, Trap, Cumbia, Cuarteto, Tango; aunque eso va a depender del análisis en particular de los clusters y un encuadre subjetivo que nosotros le demos.

Usando el algoritmo de Lloyd, el cual comienza con un con k puntos aleatorios y va

iterando los centroides hasta hallar estabilidad. De esta manera nos regresará los k clusters cuyas observaciones están mejor asociadas por su modo, timbre, modo, instrumentalidad y todas las otras características. Cuáles géneros nos quedamos va a depender de los resultados del proceso y deberíamos probar más de un k para ver que division tiene más sentido y se ajusta a la realidad.

4.2 Los modelos

Para recapitular las ideas que hemos mencionado a lo largo del trabajo, tenemos como regresores las variables macroeconómicas mencionadas el la sección previa. Vamos a tener diferentes modelos para cada una de las medidas musicales en la variable dependiente:

Valencia: Continua, $[0, 1]$. Es provista por default por Spotify y es usada por exelencia en la literatura para medir el sentimiento de una canción. Usamos regresión lineal

Letras: Continua, $[0, 1]$. Análisis de sentimiento de las letras obtenido con vader. Usamos regresión lineal

Genero: Discreta. k categorías obtenidas con clustering de K-medias. La idea acá es analizar si la situación económica se relaciona con un género en particular u otro. En cuanto al modelo que ajustaríamos en este caso, usamos regresión logística multinomial y no un análisis discriminante ya que consideramos que el supuesto de normalidad para estos regresores es demasiado fuerte.

Todos los géneros son inherentemente diferentes; tal vez la cumbia es siempre más alegre para el criterio de valencia mientras que quizás el trap obtiene sistemáticamente una valencia menor. Por esto pensamos que también sería valioso controlar los primeros

dos modelos por el género. Esto nos permitiría ver para cada género como va variando el sentimiento de las canciones del género sin el ruido al ajustar una valencia alta solo porque el género suele ser más alto.

Dicho simplemente, los primeros dos tienen la forma:

$$\mathrm{y}_t = \beta_0 + \beta_1 \mathrm{Desemp}_t + \beta_2 \mathrm{Inflación}_t + \beta_3 \mathrm{PBI}_t + \beta_4 \mathrm{Gen}_t + \beta_5 \mathrm{Pobreza}_t + \mu$$

y el tercero:

$$P(G = g|C = c) = \frac{\exp\{\beta_{g0} + \beta_{g1}\mathrm{Desemp} + \beta_{g2}\mathrm{Infla} + \beta_{g3}\mathrm{PBI}\}}{1 + \sum_{l=1}^{G-1} \exp\{\beta_{l0} + \beta_{l1}\mathrm{Desemp} + \beta_{l2}\mathrm{Infla} + \beta_{l3}\mathrm{PBI}\}}$$

No hay que olvidar que cada observación tiene un valor temporal: el año del ranking al que pertenece. (el año en el que la canción fue muy popular). Esto es porque dijimos que estábamos trabajando con datos anuales. Un problema muy grave e insalvable, que probablemente condena toda la viabilidad del trabajo, es que vamos a tener grupos de 10/15 observaciones con el mismo valor en los regresores y distinta respuesta. Una solución sería asociar el mes de cada canción en el que fue publicada. Obviamente encontramos luego que no hay datos certeros de los regresores macroeconómicos por mes, mucho menos hace 40 años, asi que deberíamos limitarnos a estimaciones imprecisas. El supuesto entonces va a ser que la canción se hizo popular apenas salió al público. Luego de eso, tratamos todo como un problema de series de tiempo con unidades mensuales. Para seleccionar el modelo de serie de tiempo eligiríamos rezagos de las variables dependientes económicas usando el criterio AIC y BIC.

La validación de modelos se realizará mediante técnicas de remuestreo: K-Fold Cross Validation, para garantizar la robustez de los resultados. En particular k=5 ya que el numero de observaciones no es tan grande.

4.3 Posibles ajustes alternativos

Nos inclinamos por modelos clásicos lineales como regresión lineal o logística porque lo interesante del problema en particular es interpretar la relación entre las variables mas que predecirlas. De todas maneras, hay otros modelos que podemos probar que tal vez provean un mejor ajuste.

Por ejemplo, quizás usando splines o métodos basados en árboles llegamos a un menor error cuadrático medio. Un árbol simple puede ser una buena idea ya que es muy fácil de interpretar. Métodos basados en árboles más elaborados como Random Forest no serían apropiados ya que no tenemos un problema de dimensionalidad. Un modelo no lineal con splines puede capturar dinámicas no lineales que ignoramos de otra manera.

5 Conclusiones y Limitaciones

En esta propuesta discutimos una forma de analizar cómo la economía afecta las preferencias culturales y las emociones de las personas. Una decisión que tomamos que puede ofuscar nuestros resultados es el largo de tiempo: comenzar desde mitades de los 80. La ventaja es que hay mas variedad de situaciones económicas agregadas y de géneros populares. Además, antes de 2010 son pocos los datos de alta frecuencia y existen solo anuales. Al reducir la cantidad de años podríamos tener rankings mensuales o semanales pero solo para los últimos 15 años en lugar de 40 y por eso preferimos esta especificación temporal.

References

Anglada-Tort, M., Masters, N., Steffens, J., North, A., & Müllensiefen, D. (2023). The behavioural economics of music: Systematic review and future directions. *Quarterly Journal of Experimental Psychology*, *76*(5), 1177–1194.

Baddeley, M. (2017). Keynes' psychology and behavioural macroeconomics: Theory and policy. *The Economic and Labour Relations Review*, *28*(2), 177–196.

de Lucio, J., & Palomeque, M. (2023). Music preferences as an instrument of emotional self-regulation along the business cycle. *Journal of Cultural Economics*, *47*(2), 181–204.

Lugos Abarca, J. A. (2023). ¿qué emociones provoca una canción? sobre un modelo probabilístico emocional – musical. *Ricercare*(16), 59—119.

Petitbon, A. M., & Hitchcock, D. B. (2022). What kind of music do you like? a statistical analysis of music genre popularity over time. *Journal of Data Science*, *20*(2).