Which is your favourite music genre? A validity comparison of Facebook data and survey data

Zoltán **Kmetty**
Renáta **Németh**

Email: kmetty.zoltan@tatk.elte.hu

Introduction

The validity of surveys has been criticised for a long time as they create an artificial environment while collecting data with some pre-specified purpose. In contrast, Facebook (FB) yields 'organic data,' that is, observational data of users' behaviors. However, validity is a concern in the case of FB. In our study, we investigate whether there are transition paths between the two data sources and whether we can overcome some validity issues and operationalization-related questions by using the two information sources in parallel to cross-validate them by each other.

Data and methods

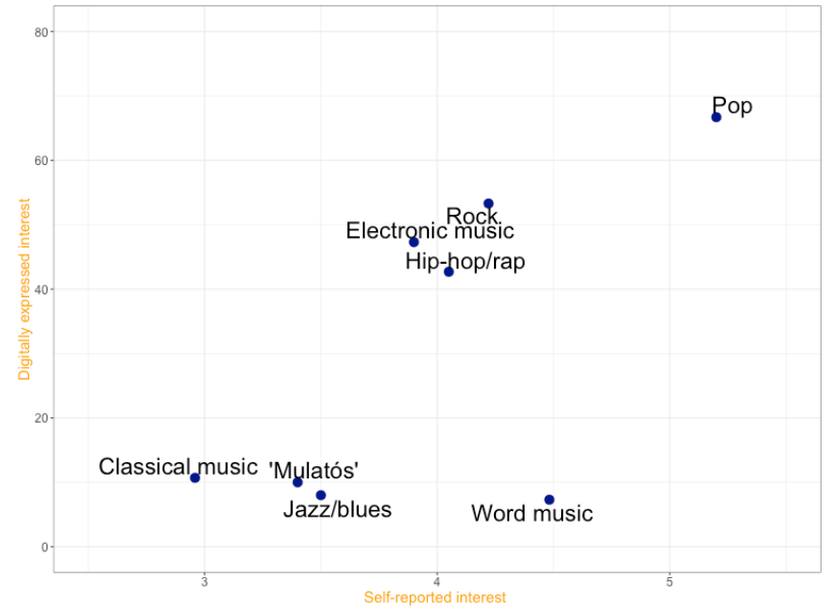
Our study uses a novel joint data source of combined Facebook and survey data. After an informed consent obtained, respondents were asked to log-in to FB on the interviewers' notebook and to download their FB profile archive. 150 respondents took part in our study. the data covers a wide range of Facebook activities: posts, comments, likes and reactions, pages, friends, profile, and ads data. Besides sharing their Facebook data, participants had to fill out an online questionnaire. Questions about politics, media usage, self-representation. spare-time activities and music preferences were asked from the participants.

We measured music interest in three different ways. (1) *Self-reported interest*. In the survey, we measured nine music genres, using a 1-7 scale of liking the given genre. (2) *Digitally expressed interest*. FB's page like data contains the name of the Facebook page. We used those pages which were categorized as music pages by FB. Our coders categorized these pages manually into genre categories that were identical to the typology what we had in the survey. (3) *Algorithmically inferred interest*. Here we used ads interest data. Facebook categorizes every user for sales for advertising. This is an algorithmic classification of the users based on their own likes, activities, and used keywords and also based on their friends' preferences. The algorithm is a black-box; we can only observe the result of the categorization.

Selected results

11 percent of the sample did not have any music-related page likes. It is an important validity question why don't they have any page likes. One possible explanation is that they do not have any interest in music, and this behavior is expressed in their (non)-activity. However, it is also possible that they don't use this functionality of Facebook, and/or they are just using Facebook lightly. We tested all of these explanations. In the analysis, we concluded that people without music page like, have overall strong music preference, and they use Facebook actively, but they don't use page-like functionality.

One of the most important questions of this study is the relationship between the self-reported interest (survey) data and the Facebook based data. Most of the genres are in a similar position in the two approaches. 'Mulatós' and Jazz are evaluated higher in self-reported data, while rock, electronic music, and hip-hop/rap categories are a little bit lower. There is one significant difference, the "world music" category.



The digitally expressed and the inferred data showed a strong correlation with each other (above 0.9). However, it was not expected that the self-reported music taste would correlate stronger with the page-likes than the ads interest.

An important validity problem arose in the case of the algorithmically inferred data. We did not find any ad interest category which fits the 'mulatós' genre. The category selection of Facebook limits the measurability of certain interest groups. Although big data theoretically makes it possible to reach smaller subpopulations, it is not obvious how we can find measures to analyse these groups if this group is not classified by the Facebook algorithm.

Full paper with further results is available here:
https://arxiv.org/pdf/2002.00501.pdf

IC2S2 2020, *July 17-20. 2020*

This work was supported by the Hungarian Research Fund (NKFIH / OTKA, grant number FK 128981.