PIC Math professor: Laura Smith
CSU Fullerton
2015 Spring

RAND Corporation
www.rand.org
RAND is a nonprofit institution that helps improve policy and decision making through research and analysis.

An Analysis of Proxy Population Mismatch in Social Media Data

Social media platforms give an opportunity to sample the opinions and responses of large sections of society. This is a treasure trove of data relevant to many policy questions. But a number of analytic problems make the principled use of such social media data complicated. One such problem is the "proxy population mismatch" problem.

Traditional polling/statistical research methods try to infer information about general populations using samples drawn from these populations. The validity of these inferences typically relies on the assumption that the samples are *representative* of the population of interest. The easy availability of social media data makes it attractive to use social media users as a proxy for general populations. But recent studies indicate that there are strong enough differences between social media users and the general population to upset that assumption of representativeness.

We propose to do a quantitative study of such a proxy population mismatch. A simple set up would be to pick a population-specific variable/dimension for which there exists verifiable "ground truth" from a different source. Derive an estimate of that variable on a social media proxy population. Then compare the ground truth to the social-media-derived estimate.

An example of such an approach could focus on the characteristics of social media users active around midnight on New Year's Eve. We could look at the reported spoken language of these users. The simplest sources of expected representation levels may be the most recent US census data about reported spoken language. The comparison could be between census-reported-proportion of US population speaking Spanish versus a social-media-informed (e.g. from twitter) estimate of the same proportion. We could also look at how these users breakdown in attributes when engaging with specific topics of students' choice.

The goal is to develop a general sense of what proxy population mismatch looks like in concrete terms. Once there is formal procedure in place for a few variables, it might be interesting to build up a database of such mismatches for different variables (e.g. ethnicity, education level, political affiliations) without topic restrictions. Such a database of mismatches may be useful for informing future predictions about proxy population mismatches for different variables or different populations.