CSC2552: Review 5, Paper 2

Due on February 28

Ashton Anderson

499 words

Thomas Hollis

Paper 2

This paper, by Blumenstock, Cadamuro and On is an *observational analysis* of *ready-made* mobile phone metadata whose original research question is to find out if it is possible to use predicted attributes from surveys to infer asset distribution. The main result of this paper is the revelation that it is not only possible to accurately forecast asset distribution using this method but that it also offers many advantages over the alternative approach of a census.

The main weakness of this paper stems from the use of the *partner-with-the-powerful* approach. Indeed, purchasing data from big mobile phone providers can help improve *external validity* as the large amount of data will help with generalisation, however this comes at a cost. The main compromise here arises from the ethical repercussions of having access to sensitive user data. While the paper claims that this data was anonymised, many subsequent papers have shown that full anonymisation of this type of data is almost impossible [1]. An alternative approach could have been the *do-it-yourself* approach which would have allowed a much more customisable dataset to be acquired without public exposure, perhaps collecting even more relevant data for the second step of machine learning inference. Another notable weakness of this paper is its vulnerability to *non-response bias* in the survey methodology. Indeed, by only considering 75 respondents of the original sample of 856 subscribers there may be *systematic bias* (e.g only the wealthy have time for surveys). However, the study has taken all the steps it could to mitigate this and has shown through its comparison with census data that while the outcome is not as accurate, the systematic bias is not large enough to invalidate the conclusions of the paper.

Conversely, a significant strength of this paper is the *amplified asking method* pioneered and used by the paper's authors, as explained in Bit By Bit [2]. Such a strategy leverages benefits of big data with benefits of surveys without encountering the major time and cost overheads of the later. However, this method does have its drawbacks. Most notably, it requires secondary supporting data to validate its results and there is no strong theoretical framework to estimate prediction uncertainty. Thus, this missing framework reduces the *construct validity* of the approach. Another limitation of the method used in this paper is the significant *coverage bias* that arises from the possibility that people with mobile phones could systematically differ from people without.

The implications of this paper were key in the formal development and demonstration of the *amplified asking method* which was shown to be around 10 times faster and 80 times cheaper than the census alternative [3]. My interpretation of the results does however differ from those of the paper's authors. While I agree with some of the major benefits of this method, I don't think enough evidence has been presented to suggest combinations of big data and survey data will work as successfully for other applications in developing countries.

[1] Narayanan, A. and Shmatikov, V. (2006). Robust De-anonymization of Large Datasets. arXiv preprint cs/0610105.

[2] Salganik, M. J. (2017). Bit By Bit: social research in the digital age. Princeton University Press.

[3] Jerven M. (2014). Benefits and costs of the data for development targets for the post-2015 development agenda. Data for Development Assessment Paper, Copenhagen Consensus Center.