9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Representativeness Explained: Uncovering Hidden Biases in Human Mobility Data

Keywords: Human mobility, Biased data, Fairness, Equitable policies, Social systems

Extended Abstract

Our capabilities to collect, store and analyze vast amounts of human mobility data have greatly increased in the past decades [1]. Today these datasets play a critical role in a majority of algorithmic systems [2], business processes [3], and policy decisions [4]. While lots of progress has been made in developing new models to analyze the data, there has been much less focus on understanding the fundamental shortcomings of these big datasets [5]. Here we focus on understanding the representativeness of high-resolution human mobility datasets. This data contains information about people's travel patterns, inferred home/work-locations, and social lives, and is increasingly used to make decisions regarding millions of people's health and well-being.

Large-scale mobility is often measured using digital tools such as smartphones. However, it remains an open question how truthfully these digital proxies represent the actual travel behavior of the general population. The literature shows that smartphones ownership is unequally distributed across society. For instance, in the US only 81 out of 100 people own a smartphone [6], with younger and wealthier groups being more likely to own one [7]. As such, it is a well-established fact that not everybody is properly represented in digital datasets. Here, we show that this type of bias is not the only one to be mindful of. Understanding *how* people are represented in big digital datasets is of equal importance.

We study which demographic factors determine how certain groups are represented in mobility data. This aggregated mobility data is provided by Spectus, a location intelligence platform. Data is collected from anonymized users who have opted-in to provide access to their location data anonymously, through a CCPA and GDPR-compliant framework. Through its Social Impact program, Spectus provides mobility insights for academic research and humanitarian initiatives. We focus on mobility data generated in big cities in the United States and compare travel patterns on census-tract level to demographic data on poverty, sex, age, education and ethnicity, obtained from the United States Census Bureau. Individuals are linked to census tracts based on their inferred home locations. To understand the effects of poverty we divide users into 5 equally sized groups (quintiles), with Q1 containing the poorest 20% and Q5 containing the wealthiest 20%. Comparing how many datapoints each group generates reveals noticeable differences. Fig. 1a shows the results for New York City for data from the month of April 2021. The poorest 20% of individuals generate around 17% of all data, while the wealthiest group generates approximately 22%. These findings are robust for other months (April 2019 and April 2020) and for other cities, for instance for Chicago (Fig. 1b). Imbalances in data production rates can ultimately result in the travel behavior of poorer individuals being less detailed. However, imbalances in data production are not the only issue we identify. We further find that data production has different temporal fingerprints, depending on whether data comes from poorer or richer individuals. Fig. 1c shows differences in when data is generated (averaged into an average week). We see a consistent pattern of relatively more activity during night-time for the low-income group, and a mobility behavior which is shifted towards later

hours. This will undoubtedly impact efforts to correct, or de-bias data, as simple under and over-sampling techniques might not correct for this.

Lastly, we find that data imbalances are not only affected by poverty. To untangle which factors impact mobility data we built a model to predict the average number of datapoints a *regular* person living in different census tracts would generate (determined from the median person). Fig. 1d, shows the effect different demographic features have on the number of generated datapoints. For instance, a one percentage point increase in poverty, on average, results in approximately 250 fewer datapoints. Similarily, Fig. 1d shows that other demographic variables, like ethnicity and education, also have a large influence on data production. A remarkable finding is the big negative association with academic education. Our model reveals that more educated regions (higher number of people with university degrees) have lower data production rates. What makes this observation remarkable is that poverty and academic levels are negatively correlated (rank correlation $r = -0.52$, $p \ll 10^{-5}$), but have almost similar influence on data generation. This can potentially be the result of algorithmic confounders [8]. Our mobility dataset is created by pooling data from multiple smartphone apps. Each of these may be using different social engineering mechanisms to collect data, which will invisibly nudge their users towards specific behaviors—future work will involve untangling these effects.

Biased data leads to biased algorithms. Our analysis shows how certain demographic factors can severely impact data representation. Our work is a first step towards understanding how to develop techniques that correct for biases in widely used human mobility datasets, and to provide actionable guidelines for how to make algorithmic systems more equitable.

References

[1] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[2] C. Wagner, M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor, and T. Eliassi-Rad, "Measuring algorithmically infused societies," *Nature*, vol. 595, no. 7866, pp. 197–204, 2021.

[3] C. Demunter, "Tourism statistics: Early adopters of big data," *Publications Office of the European Union*, 2017.

[4] J. E. Blumenstock, "Estimating economic characteristics with phone data," in *AEA Papers and Proceedings*, vol. 108, pp. 72–76, 2018.

[5] K. Crawford and R. Calo, "There is a blind spot in AI research," *Nature*, vol. 538, no. 7625, pp. 311–313, 2016.

[6] Pew Research Center, 2019. Smartphone ownership is growing rapidly around the world, but not always equally.

[7] Pew Research Center, "Mobile fact sheet," 2021. [Last accessed 2023-02-22] https://www.pewresearch.org/internet/fact-sheet/mobile/.

[8] M. J. Salganik, *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

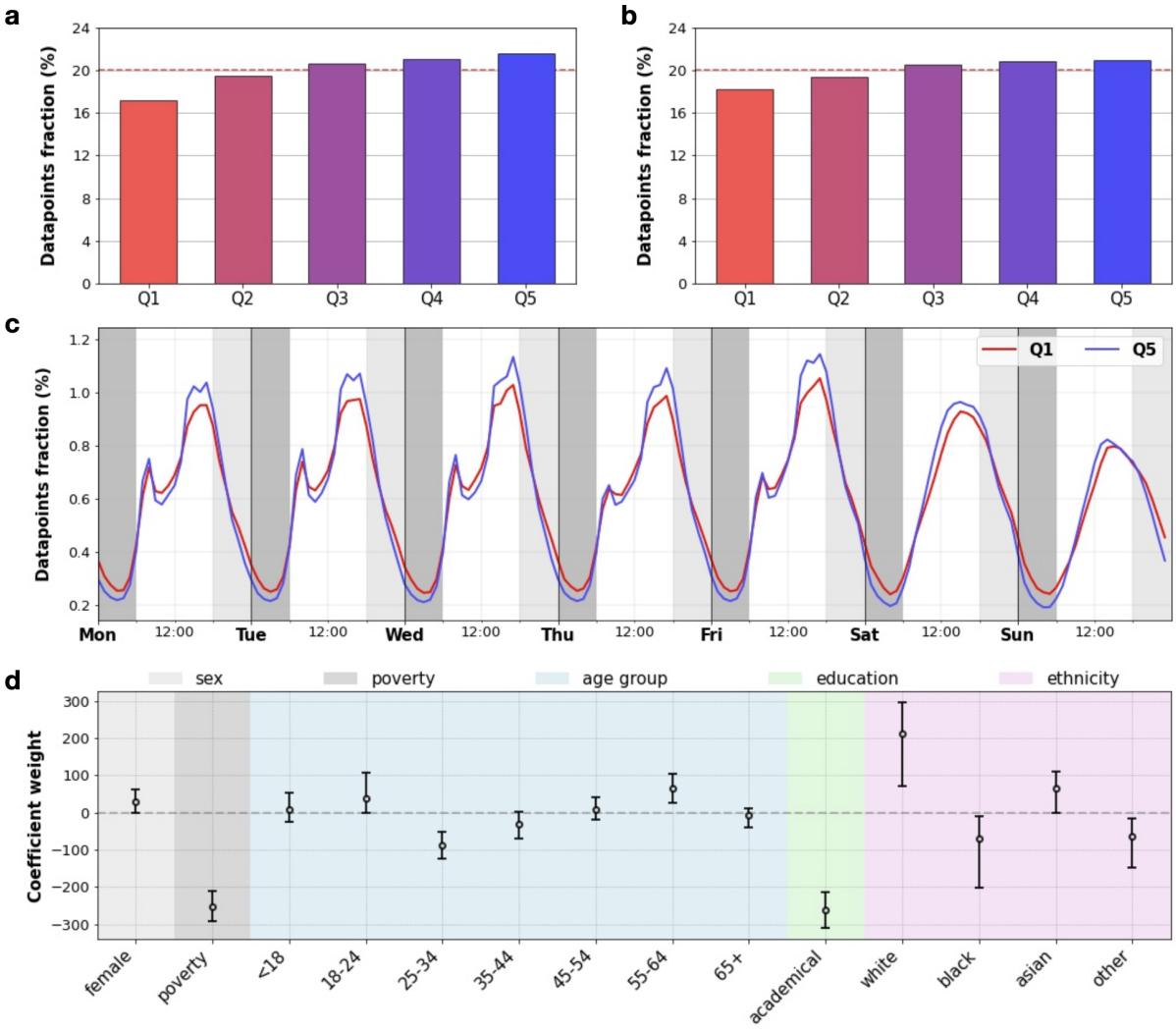

Figure 1: **Data representation for April 2021. a,** Datapoint contribution per poverty quintile in New York City (NYC). Poorest to richest quintiles are indicated with Q1 to Q5 respectively. **b,** Datapoint contribution per poverty quintile in Chicago City. **c,** Temporality of data production in NYC per hour. One month of data is collapsed onto the corresponding weekday and hourly aggregated counts are normalized per poverty quintile. **d,** Coefficient weights of model (LASSO regression, adjusted $R^2 = 0.31$) for median aggregated NYC census tract datapoints per person. Error bars indicate 95% confidence intervals, estimated from 1000 bootstrapped samples.