Enhancing household survey microdata accuracy using machine learning: project plan

Nikhil Woodruff

1 Project outline

Tax-benefit policy is the set of rules that determine tax liabilities and benefit entitlements, deciding the allocation of nearly half of GDP in the United Kingdom alone. When policymakers, think-tanks or academic institutions propose changes to tax-benefit policy, they often use mathematical models to estimate the impact of such changes. The dominant approach to modelling tax-benefit policy is *microsimulation*. Microsimulation models are combined of two parts: a model of policy logic that can compute taxes and benefits for a given household under current and proposed changes, and dataset representative of a target population, on which the policy model is executed. Given that the most common population of interest is the private household population in a country, most microsimulation models use a household survey dataset as their input. The most common of these surveys is the Family Resources Survey (FRS).

The accuracy of microsimulation models is crucial to their utility. Policy logic is easier to verify, since it is specified in detail by legislators. However, household surveys can often fail to be truly representative of the population. We know that this is the case, because we have access to other datasets which are both quantitatively different (in the distributions of shared variables) and more likely to be correct (due to better administration processes or higher population coverage). For example, the Survey of Personal Incomes (SPI) is a random sample of HMRC's Real Time Information computer-based database of personal tax and income data. Both the FRS and SPI contain the annual dividend income of respondents, but the total dividend income in the FRS is aound 40% of the total in the SPI.[2] Similarly, the Department for Work and Pensions estimated in 2008 that the FRS excluded 26% of the Severe Disablement Allowance recipient population that should have been included in a representative sample of the UK household sector.[1]

Approaches to improving the accuracy of household survey data (up to the accuracy of administrative data) as used in microsimulation have been limited and largely unsuccessful. This can lead analyses to use separate datasets in order to answer separate policy impact questions. For example, if a tax policy raises taxes on dividends to fund a universal cash transfer, analysis which uses the SPI to estimate tax revenues, and the FRS to estimate the percentage of households better off under the policy, will be biased because the FRS under-represents the dividend-holding population, while the SPI accurately represents it.

This project will aim to answer the question of *how much can machine learning improve the accuracy of household surveys*. It will survey the current approaches used to improve the accuracy of household surveys (not limited to the U.K.). It will implement various different methods (including the current state-of-the-art) and evaluate their performance at reducing the gap between the FRS and administrative data when used to answer questions about the U.K. household population. These methods will include:

- Percentile-based adjustment of data values.
- Gradient-descent based methods to reweight household records.
- Matching-based methods to impute data across

surveys.

- Random forest-based imputation of data across surveys.
- Distributional random forest-based imputation.

2 Project plan

The project has been split into several sections for simplicity. Each task, numbered below, is shown in the Gantt chart in Figure 1.

2.1 Literature survey

1. The literature survey will analyse current approaches to improving survey microdata accuracy, primarily from across the United States, Europe and the U.K.

2.2 Calibration

This section will evaluate the extent by which reweighting (using gradient descent) can outperform changing data values according to percentiles in adminstrative data.

2. Implement percentile adjustment on the FRS using the SPI.
3. Implement gradient-descent-based reweighting.
4. Benchmark methods against each other.
5. Write up validation results and conclusions.

2.3 Imputation

In this section, I will evaluate the performance of novel and SOTA methods for imputation, the process of adding in data in a survey with data from another survey.

6. Implement matching.
7. Implement random forest imputation.
8. Implement distributional random forest imputation.
9. Benchmark methods against each other.
10. Write up validation results and conclusions.

2.4 Application

Under this section, I will apply the calibration and imputation methods together on the FRS, assessing any change in accuracy.

11. Gather calibration data for the FRS.
12. Apply a calibration method to the FRS.
13. Apply an imputation method to the FRS from other household surveys.
14. Apply reweighting together with imputation for the SPI income variables.
15. Evaluate any change in accuracy on SPI income variables.
16. Benchmark the enhanced FRS against the original on unseen targets.
17. Write up the FRS enhancement process and performance.

2.5 Review

This phase is the final review stage, ensuring the project report is complete.

18. Review the final report.

References

[1] Stephen McKay. Evaluating approaches to Family Resources Survey data linking.

[2] Tahnee Christelle Ooms. Correcting the underestimation of capital incomes in inequality indicators: with an application to the uk, 1997–2016. *Social Indicators Research*, 157(3):929–953, 2021.

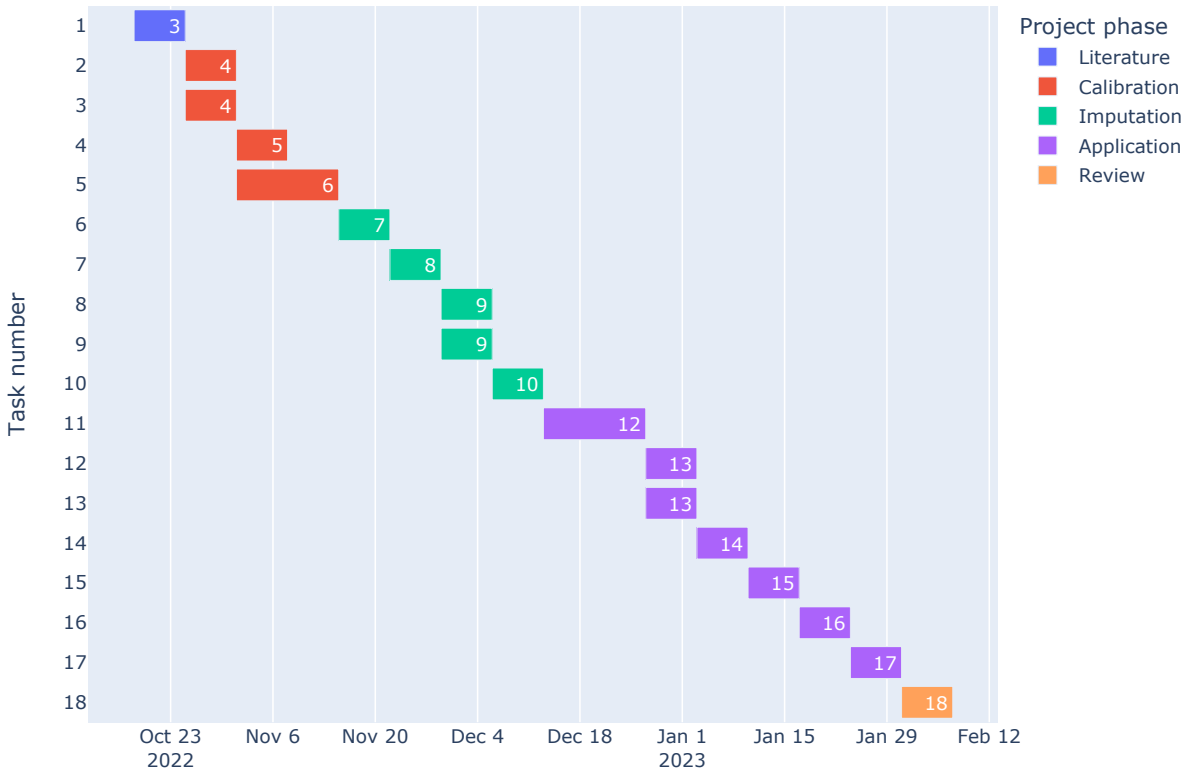

Figure 1: Gantt chart with project milestones