Surveying the loss landscape: using machine learning to improve household survey accuracy

Nikhil Woodruff[1, 2], Max Ghenis[2]

[1] University of Durham, [2] PolicyEngine

Abstract

Microsimulation over survey datasets remains the one of the most widely used methods for analysing and predicting the impact of government policy. However, the accuracy of these models is often limited by the quality of the survey data. In this paper, I present a novel approach to improving the accuracy of survey data by using machine learning to counter both sampling and measurement error. I evaluate this approach on the UK's Family Resources Survey in combination with other datasets and benchmark its performance against other methods currently used within tax-benefit microsimulation modelling. I find that the proposed approach can improve the accuracy of the survey data as a predictor of more trustworthy statistics from administrative sources which are not granular enough to be used for microsimulation.

Introduction

Governments across the world redistribute trillions of dollars every year through tax and benefit programs. On average, countries collect around a third of their gross domestic product in taxes and distribute around a third of tax revenues as cash benefits. These programs are ultimately incident on households, either directly or indirectly. Direct taxes and benefits impact households by direct transfers between households and government entities, and indirect taxes or benefits (subsidies) levied on private suppliers manifest themselves in the altered transactions between households and non-government entities. But whether by indirect or direct means, tax-benefit

policy represents one of the largest single contributors to the state of household finances for billions of people.

The complex rules defining the behaviour of these programs are set in tax-benefit legislation and are continuously amended or reformed. While legislators may have different overall goals for tax-benefit policy outcomes, engineering the tax-benefit system to achieve specific intended outcomes requires being able to predict the impact of individual policy reforms on metrics like the budget, the poverty rate, or the distribution of income. Therefore, the accuracy with which researchers can predict the impact of policy reforms is critical.

Researchers largely achieve this through static tax-benefit microsimulation on survey microdata, a technique which involves simulating the application of tax-benefit program rules on a representative sample of households, using collected data on demographics, household structures, income sources and other features. This technique has been used extensively across both developed and developing countries. Static tax-benefit microsimulation models' accuracy is dependent on the accuracy of the survey data on which it operates. However, there is evidence to suggest that the surveys which are used in microsimulation models today are inconsistent with administrative data and introduce inaccuracy when used to estimate properties of the household population.[1] Research into these errors has found this largely stems from measurement errors combined with (deteriorating over time) survey response biases.[2]

Inaccuracy in survey data has real-world impacts: by negatively affecting the accuracy of tax-benefit policy microsimulation, survey inaccuracy can distort the conclusions of policy evaluations, leading to suboptimal government policy design. Where surveys consistently under-represent subsectors of the household sector, this can lead to a systematic bias in the policy evaluation process and distort public understanding of the impact of government policy.

Most producers of household surveys do not adjust microdata to attempt to counter this,[3] but where statistical agencies do, the methods for mitigating these types of inaccuracy rely on somewhat arbitrary assumptions about the distribution of survey data variables. For example, in correcting for the under-representation of high incomes, a common approach is to match the top income percentiles of a household survey to percentiles from administrative tax datasets.[4] This can achieve exact parity between the two datasets on this very specific target metric, but this introduces significant risk of overfitting: of all the questions that we could ask of the survey data,

[1] https://www.iser.essex.ac.uk/research/publications/working-papers/understanding-society/2020-01
[2] https://www.aeaweb.org/articles/pdf/doi/10.1257/jep.29.4.199
[3] https://onlinelibrary.wiley.com/doi/full/10.1111/1475-5890.12158
[4] https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/compendium/economicreview/february2020/topincomeadjustmentineffectsoftaxesandbenefitsdatamethodology

income percentiles are a small fraction, and other targets could plausibly be thrown off by this adjustment.

Machine learning has emerged as a powerful tool for data transformation tasks, including synthetic data generation, across many social science subfields. However, the field of tax-benefit microsimulation has been largely untouched by it in the decades that it has been the standard. Additionally, the field of machine learning has a wealth of prior research into the problem of overfitting- where a model trained to perform a specific task over-specialises to the data used to train it and performs worse on generalised tasks. It could be argued that many of the current methods for adjusting survey data show signs of overfitting- household weights (akin to model parameters) are often optimised to hit demographic statistics (the training data) exactly and perform poorly when used to reproduce financial statistics (unseen validation data).

This paper aims to address this gap by proposing a novel approach to improving the accuracy of survey data by using machine learning to counter both sampling error (by gradient descent-based reweighting on a balanced loss function) and measurement error (by random forest model-based synthetic data generation). I evaluate this approach on the UK's Family Resources Survey in combination with other datasets and benchmark its performance against the other methods currently used on the FRS in microsimulation models, finding a sizeable accuracy improvement.

This paper is structured as follows: *Background* sets out the related research on survey inaccuracy and its causes, current applied methods for improving survey performance in microsimulation models, and relevant research in the machine learning field for solving tasks of this nature. *Methodology* specifies the end-to-end pipeline for improving household surveys tested in this paper. The *Results* section contains the main performance assessment of the new approach, and how it fares compared to the current FRS on several tasks that are representative of the usage of tax-benefit microsimulation models. The *Conclusions* section summarises the main findings from these experiments and discusses their applicability to wider policy impact assessment processes.

Background

Over the last few decades in which microsimulation has been used for tax-benefit policy analysis, the accuracy of household surveys, and methods for improving it, have been highly scrutinised.

Inaccuracy in household surveys

Nationally representative household surveys provide the necessary level of detail on respondents to simulate most taxes and benefits and aggregate up to the national level. Yet concerns remain (and have risen over recent decades) about their accuracy, particularly when used for tax-benefit microsimulation.

The basis for these concerns is largely two-fold: *measurement error* in household surveys arising when answers to questions are inaccurate, and *sampling error* where the weighted households in the survey are not perfectly representative of the wider population.

These errors are observable when used to estimate policy impacts for which there is comparable administrative data and are common across country contexts. For example, Cantor et al find the U.S. Current Population Survey (CPS) yields substantially higher estimates of healthcare subsidy enrolment compared with administrative totals.[5] In the United Kingdom's Family Resources Survey (FRS), McKay et al find significantly lower claimant counts than in administrative data for some disability or care-related benefits in the Family Resources Survey.[6]

The reasons for these errors are likely to be specific to the specific error type they appear under. For example, a key explaining factor in the discrepancy between inequality and high incomes statistics depending on the source data is likely to be because of a 'missing rich' problem, where high-wealth households are under-sampled or have financial variables under-reported in surveys.[7] [8] As discussed by Lustig, there are several plausible reasons for this; primarily that the ultra-rich are few enough in number that they are unlikely to be sampled. The FRS does not capture wealth but is still affected through the reporting of capital incomes, particularly at the top of the income distribution, have declined compared to administrative data from both under-coverage and under-reporting.[9] The UK's Wealth and Assets Survey, which does capture wealth properties of surveyed households, is similarly missing around £800bn held by the richest households.[10]

Additionally, high incomes are not the only area where inaccuracy can be identified with reasonable certainty. One method for establishing likely income under-reporting, particularly at the low end of the distribution, is to compare two groups with different income compositions but assumed similar consumption preferences- for example, identifying the share of self-employment income under-reported by matching self-employed households with equivalent employed households.[11] This approach originally identified likely under-reporting of self-employment income among UK households under the reasoning that employed households with the same reported total income tended to spend less (self-employed individuals might be incentivised to under-report income to household surveys if they are also under-reporting income to tax authorities). Applications of this method to

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1955284/

[6] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/222871/WP110.pdf

[7] https://www.oecd-ilibrary.org/sites/9789264307278-5-en/index.html?itemId=/content/component/9789264307278-5-en

[8] https://www.aeaweb.org/articles?id=10.1257/jel.49.1.3

[9] http://eprints.lse.ac.uk/108900/1/Ooms2021_Article_CorrectingTheUnderestimationOf.pdf

[10] https://eprints.lse.ac.uk/112698/1/1475_5890.12286.pdf

[11] https://www.sciencedirect.com/science/article/abs/pii/0047272789900522?via%3Dihub

household surveys across Europe have found similar under-reporting, suggesting between 10% and 40% of self-employment income is uncaptured.[12]

McKay et al attempted to quantify the measurement error of the FRS by linking individual households with data from the DWP's administrative records, using non-public identifiers.[13] The process of linking is not perfect: respondents are asked for permission to link their survey data with administrative data, and some (around 30%) refuse. However, for each benefit, the authors were able to find the percentage of reporting adults for whom a link to an administrative data record could be identified, the percentage of reporting adult recipients for whom no link could be found, and the percentage of adults represented only by administrative data.

They found that these splits vary significantly by benefit: recipient data on the State Pension (SP) is highly accurate in the FRS (96% of SP reported recipients were represented by the FRS, 1% were only on the FRS and not on administrative datasets, and 3% were only on administrative datasets). At the same time, around 62% of adults on the FRS who reported receiving Severe Disablement Allowance could not be identified in administrative data. There are multiple possible reasons for this, and they vary by benefit: the recipient population is often confused or mistaken when answering questions about their benefits, and this is more acute for age- or disability-related benefits. This appears to provide additional evidence that measurement error is significant, at least at the low-income subset of the surveys.

Importantly, evidence from both administrative data and other methods like consumption patterns point to household survey inaccuracy. Administrative data alone cannot be taken as a sole source of truth: income tax evasion still occurs,[14] implying administrative data still contains errors, and administrative data often contains different populations than household surveys (for example, by excluding individuals with income too low to pay tax) which means it cannot be directly compared without introducing some degree of inaccuracy. Additionally, administrative data often has a narrower focus in the questions it asks of respondents (for example, excluding common family transfers or capital gains because they are not needed for the specific goal of the administrative dataset, e.g., powering the tax collection process). However, administrative data is still likely to be more trustworthy because individuals have less choice in their data inclusion- it is a highly useful tool in the process of auditing the accuracy of surveys, but not the sole truth.

Although the evidence base from administrative data and other methods is strongest on the existence of under-sampling at the top and under-reporting at the bottom, either error category cannot be ruled out at any point along the income and wealth distributions.

[12] https://link.springer.com/article/10.1007/s10797-019-09562-9
[13] https://www.gov.uk/government/publications/family-resources-survey-data-linking-wp110
[14] https://www.gov.uk/government/statistics/measuring-tax-gaps/tax-gaps-main-findings

These problems have been known while tax-benefit microsimulation models have become widely used to analyse policy impacts. Various government agencies and non-government agencies have developed methods for reducing the inaccuracy issues household survey microdata suffers from.

Household survey weight generation

Survey microdata weights enable each record in a dataset to be aggregated together to form national or subnational statistical estimates. Surveys across countries usually use specific methods to derive these weights, but broadly follow a pattern of using initial design weights (initialising weights according to the probability of selecting a household at random) and applying numerical optimisation methods to calibrate weights to reproduce administrative demographic statistics. In the United Kingdom for the FRS, these demographic statistics include populations of age-sex-region intersections, as well as family and household population sizes by region.[15] In the United States' Current Population Survey, target statistics include race and ethnicity-based population totals, as well as state-level populations.[16] Australia's Bureau of Statistics calibrates weights in the Survey of Income and Housing to targets including state populations by age and sex, as well as labour force category.[17] Some of these procedures also include a penalty for large changes between the design and calibration weights. However, at the time of writing, none include financial statistics such as market incomes or tax-benefit aggregates in the calibration process.

Tax-benefit policy microsimulation

Tax-benefit policy microsimulation has emerged over recent decades as one of the most widely used methods for analysing and predicting the impact of government policy. This technique involves simulating the application of tax-benefit program rules on a representative sample of households, using collected data on demographics, household structures, income sources and other features.

In the United Kingdom (on which this paper will test its proposed survey improvement methods), one of the most comprehensive household surveys is the FRS. This survey is collected annually by the Department for Work and Pensions under its responsibility to produce poverty and inequality statistics (Households Below Average Income) and includes approximately 20,000 households each year. Estimates for population-level features (such as the median income) can be derived using individual weights for household records which indicate how many UK households each respondent is

[15] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/321820/initial-review-family-resources-survey-weighting-scheme.pdf

[16] https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/weighting.html

[17] https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/6553.0~2017-18~Main%20Features~Weights~25

representative of. These weights are calibrated to hit population totals and other demographic statistics, and not financial statistics. The types of errors common across household survey in general are similarly observable in the FRS.[18]

Survey improvement methods

Various methods have been developed to attempt to counter the issues of survey inaccuracy over recent decades. These methods largely can be categorised into two groups: reweighting-based methods (that adjust survey weights to alter the properties of the aggregated dataset) and imputation-based methods (that change the values of the survey respondents' answers to questions).

In the UK, the Department for Work and Pensions (the government body responsible for the FRS) developed a process aimed at countering the under-coverage of high incomes within the FRS, called the 'SPI adjustment' (also referred to as 'percentile adjustment' in the generalised form). This adjustment essentially corrects the distribution of total (gross) income within the FRS to match the equivalent population within the Survey of Personal Incomes (a sample of HMRC's administrative tax dataset). Examined by Burkhauser et al, this method is straightforward: identify high-income individuals (with income above a threshold) for adjustment, replace incomes with equivalent SPI percentiles, and recalibrate weights, including a penalty for the total number of high-income individuals in addition to the demographic targets (separately for pensioners and non-pensioners).[19]

Burkhauser et al, although noting that this approach is successful at its core purpose and that no other country outside the UK employs such a method, raise several issues with this approach: firstly, it imputes total income values, rather than the components of total income (which tax-benefit microsimulation requires to simulate policy impacts). The parameters for the method are also arbitrary: applying to the top half-percent of the income distribution, and separating pensioners and non-pensioners. Additionally, SPI data is usually released at least a year after the FRS microdata for the same time period, which means that SPI data need to be adjusted for the relevant data lag (for example, the 2021-21 FRS would need to be adjusted using the 2019-20 SPI).

The issue of income decomposition remained largely untackled until Ooms et al attempted to improve the reporting of a specific component of gross income which is more severely under-reported in the FRS than others: capital income.[20] They first establish that income under-reporting is mostly due to

[18] https://www.gov.uk/government/statistics/family-resources-survey-financial-year-2020-to-2021/family-resources-survey-background-information-and-methodology
[19] https://onlinelibrary.wiley.com/doi/full/10.1111/1475-5890.12158#fisc12158-bib-0013
[20] https://doi.org/10.1007/s11205-021-02644-4

this particular category by comparing individual income sources between the FRS and SPI, finding that the aggregates of non-capital income are around 100% of the totals for the SPI, while capital income is only around 40% as represented. The authors present a novel observation about the instances where capital income is under-reported: the capital share of income in individuals is far less represented in the FRS than in the SPI (specifically, the number of individuals with a high capital share'), rather than simply a lack of high-capital-income individuals.

They introduce a new method to correct for this under-capture: adjust the weights of high-capital-share individuals to match the totals in SPI data, finding that the new method is largely successful at correcting for under-capture of capital income, and increases the Gini coefficient of FRS data by between 2 and 5 percentage points (applying the methodology to historical FRS data releases). However, they do not measure the changes to how well the FRS ranks against other aspects of the SPI.

Machine learning

Optimisation methods play a key role underpinning many of the approaches in improving survey microdata, particularly in reweighting. Household survey weights are largely processed by applying numerical optimisation algorithms penalised by deviations from demographic statistical targets. Linear programming methods are used to determine the optimal weights for the Family Resources Survey, according to limits on how far apart the FRS aggregates can be from national and regional population estimates.[21] Multiple U.S. federal tax models apply a linear programming algorithm to solve for weight adjustments satisfying a combination of tax statistic deviation constraints, and weight adjustment magnitude limits.[22] [23]

There are several reasons why machine learning techniques are well-suited to the task of survey imputation. The most fundamental justification is in its context-agnostic nature: machine learning approaches do not require assumptions specific to the field they are applied in, unlike the current approaches to survey accuracy improvement (for example, percentile adjustment which explicitly partitions households into 'rich' and 'non-rich' using arguably arbitrary definitions). In other domains, for example image classification, a move away from prescriptive methods towards loss function minimisation has seen substantially improved accuracy and robustness.[24]

Gradient descent, a technique for finding parameters which minimise a loss function, iteratively updates the parameters in the direction of the steepest

[21] https://www.gov.uk/government/publications/initial-review-of-the-family-resources-survey-weighting-scheme
[22] https://www.taxpolicycenter.org/resources/brief-description-tax-model
[23] https://github.com/pslmodels/taxdata
[24] https://www.researchgate.net/publication/209804567_A_Survey_of_Image_Classification_Methods_and_Techniques_for_Improving_Classification_Performance

negative gradient.[25] This is a highly common technique in machine learning, and is used in a variety of contexts, most notably as the foundation for training artificial neural networks. It relies on no domain-specific assumptions other than those present in the definition of the loss function, enabling it to be applied to a wide range of problems. Several variations of gradient descent have emerged over the years which achieve more efficient training procedures: stochastic gradient descent steps in the direction of an estimate of the gradient using individual training examples, rather than loading the full dataset.[26] Mini-batch gradient descent represents a compromise between batch (full-dataset) and stochastic gradient descent, by iterating parameters using fixed-size subsets of the training data.[27]

As well as gradient calculation methods, optimisation algorithms have revealed significant accuracy and efficiency improvements by defining behaviours for hyper-parameters such as the learning rate (the velocity at which parameters follow the gradient). These include Adam,[28] AdaGrad,[29] and the (unpublished) RMSProp optimiser. Gradient descent could feasibly be applied to survey accuracy problems, since it requires only a loss function that is differentiable with respect to the parameters being optimised. In the context of survey accuracy, a loss function could be defined as the squared errors of individual aggregate statistics between official sources, and a survey, which would be continuously differentiable over the weights of individual household records.

Relevant to the problem of countering measurement error are machine learning methods which generate new realistic values given context. Random forest models are a type of ensemble learning technique, which combine the predictions of multiple decision trees to produce a more accurate prediction than any individual tree.[30] The decision trees are trained on a subset of the training data, and the predictions of each tree are combined using a voting system. Although its introduction is far less recent than more modern innovations in the field of neural networks (for example, artificial neural network variants[31] or transformers[32]) random forest models have shown consistently high accuracy across a wide range of domains, remaining competitive with the most recent techniques. This type of model has been

[25] https://ieeexplore.ieee.org/document/363438
[26] https://proceedings.mlr.press/v108/wen20a.html
[27] https://ieeexplore.ieee.org/document/8264077
[28] http://arxiv.org/abs/1412.6980
[29] http://jmlr.org/papers/v12/duchi11a.html
[30] https://doi.org/10.1023/A:1010933404324
[31] https://www.researchgate.net/publication/314457741_A_Survey_on_Various_Applications_of_Artificial_Neural_Networks_in_Selected_Fields_of_Healthcare
[32] https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

applied (to a limited extent) in the context of policy analysis and have shown superior performance in prediction tasks to logit and other model types.[33]

There are several reasons why random forest models might outperform neural networks in predicting survey microdata values from other attributes (for example, predicting employment income from demographic variables), but the most natural reason is that tax-benefit law, which heavily influences financial decisions, is more similar in structure to a random forest than a neural network. For example, in Dowd et al found that capital gains variables are ùnnaturally' distributed in order to respond to incentives set by particular tax law parameters.[34] Methods used currently in surveys such as percentile adjustment often use 'matching', in which a record in a survey has its value replaced with the closest record in another according to a specific criteria (for example, in the SPI adjustment, records within the same percentile group are replaced with the exact same mean income value as exists in the SPI). This removes heterogeneity, which may have adverse effects when used later in a microsimulation model.

Methodology

This section presents a new proposed integrated pipeline of methods to enhance survey microdata. The main novel additions in this method are the use of a balanced loss function (a set of statistical targets that does not exclude financial statistics, and with weights similar to the target uses of the survey microdata in tax-benefit microsimulation models) with gradient descent-powered reweighting to correct sampling error, and the combination of that method with random forest-based imputation models to correct measurement error.

Balanced loss

The loss function is the function that is minimised by the optimisation algorithm. In the context of survey imputation, the loss function is the difference between the survey aggregate statistics and the official aggregate statistics. The loss function is defined as:

$$L(S) = \sum_{c \in C} L_c(S)$$

where $L_c(S)$ is the loss function for a particular aggregate statistic c, C is the set of all aggregate statistics and S represents a given household survey (a collection of relational databases). The loss function for a particular aggregate statistic c is defined as:

[33] https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2408~aa6b05aed7.en.pdf?9551c7c6e8e8fdbd35e5512b5afcf097

[34] https://doi.org/10.17310/ntj.2019.2.02

$$L_c(S) = w\left(\frac{\sum_i^N (X_i \cdot \max(W_i, 0))}{y} - 1\right)^2$$

where X_i is the value (of a particular variable) of the i −th household record, W_i is the weight of the i −th household record, y is the official aggregate statistic for c, and w is a weighting factor for the loss function. The weighting factor w is used to prioritise certain aggregate statistics over others (for example, budgetary impact size is used to comparatively weight different financial aggregate statistics). The loss function is also hierarchical, in that each loss category contains a weighted sum of other (normalised) loss functions. For example, the loss function for demographic performance might contain subcategories measuring performance over household population targets as well as individual population targets. Note that the weight W_i is constrained to be non-negative, since negative weights do not have a meaningful interpretation in the context of survey imputation.

There remains an issue in how to constrain the relative sizes of different loss values. For example, we might have many more detailed statistics over which we can evaluate the survey in its representation of Income Tax (revenues by income band, taxpayer counts) than we do for Child Benefit (only aggregate revenue and total claimants). Naively summing the relative error comparisons for each category would give Income Tax targeting a much higher weight in the optimisation process purely because of our access to statistics (which is no indication of a program's importance). Simply dividing by the number of comparisons would be inaccurate too, given some of those comparisons might be more important than others. Therefore instead, a more neutral assumption is to normalise each loss category by dividing by its initial value:

$$L(S) = \frac{\sum_{c \in C} L_c(S)}{L(S_0)}$$

The question of how to determine the weighting factor for each loss function is also arbitrary, but the most neutral assertion could be to use the aggregate financial size of a program, or the size of the population concerned. For example, if we have one loss category measuring how well the survey reproduces Income Tax statistics and another measuring Child Benefit, we could reasonably consider that the Income Tax loss category is approximately twenty times more important than the Child Benefit loss category.

This paper benchmarks the performance of the survey accuracy improvement pipeline by focussing on UK survey data, which requires an implementation of the loss function for the UK context. The UK loss function is defined using the following categories:

1. Demographics
    a. Households
        i. Region-Council Tax Band intersections*
        ii. Region-tenure type intersections*
    b. Populations

    i. Age-sex-region intersections*
2. Programs
    a. Universal Credit
    b. Child Benefit
    c. Child Tax Credit
    d. Working Tax Credit
    e. Pension Credit
    f. Income Support
    g. State Pension
    h. Housing Benefit
    i. Income-based Employment Support Allowance
    j. Income-based Jobseeker's Allowance
    k. Council Tax
    l. National Insurance
    m. Employment income
    n. Self-employment profit
    o. Private pension income
    p. Savings interest income
    q. Property income
    r. Dividend income
    s. Income Tax
        i. Taxpayers by UK nation
        ii. Tax liability by 10 income bands

Loss categories marked with an asterisk (*) contain statistics that are also used to determine the original FRS survey weights. Categories within the *Programs* include aggregate financial size and non-zero counts by UK nation, weighted by the aggregate financial size of the program. For example, the program loss category is approximately defined by:

$$\begin{aligned} \mathrm{L}_{\mathrm{Programs}}(\mathrm{S}) = 45 \times 10^{9}\mathrm{L}_{\mathrm{UniversalCredit}}(\mathrm{S}) \\ +11 \times 10^{9}\mathrm{L}_{\mathrm{ChildBenefit}}(S) \\ + \ldots \\ +200 \times 10^{9}\mathrm{L}_{\mathrm{IncomeTax}}(\mathrm{S}) \end{aligned}$$

This is necessary, since loss categories in themselves are normalised and expressed as a percentage of the first loss value.

A given survey S is itself a set of variables $\mathrm{X}_{\mathrm{i,j}}$ (where i is the household record and j is the variable), as well as household weights W_{i}.[35] We can therefore split up the loss function to be a function of the variables and weights separately (and implementing this split is achievable in the underlying algorithm code) as $\mathrm{L}(\mathrm{S}) = \mathrm{L}(\mathrm{X}, \mathrm{W})$. Our loss minimisation task therefore becomes finding the solution to the equation:

[35] Although most household surveys also include personal and family weights, only the household weights are optimised in this project.

$$\frac{\partial L(X, W)}{\partial W} = 0$$

The loss function for a specific household survey will be a large set of composite functions incorporating hundreds of individual targets, but the gradient function can be analytically calculated using automatic differentiation packages such as PyTorch.[36] Under the gradient descent algorithm,[37] the weights are iteratively updated in the direction of the steepest negative gradient, until the loss function is minimised.

Imputation

There are several reasons why reweighting alone will likely not be sufficient to eliminate certain types of error in the survey. For example, suppose that one of the income tax targets involves the revenue from certain high-income tax filers. A survey which does not include any instances of these filers will categorically be unable to make any progress towards this target. This case does occur frequently in practice: the highest taxable income in the Family Resources Survey (2020-21) is less than £1m, but HMRC reports aggregate tax revenues from filers with incomes over this level in the order of £1bn.

This problem manifests as a global minimum floor in the loss landscape over the space of the original survey weights, below which no optimisation improvement can reach. To circumnavigate this barrier, we must add new records and weights to the parameters optimised by the gradient descent algorithm.

Synthesising new records to add to the existing survey brings risks: we could decrease the accuracy of the survey by adding new records which are unrealistic. This is not a concern: it can be avoided entirely by adding new records with weight values set to zero, since this does not actually change the result of anything produced with the survey, and instead just provides the optimisation algorithm with more parameters to change. However, while synthesising implausible records cannot harm the survey, it is likely that the more plausible the new record, the better it can aid the optimisation routine.

There are a variety of machine learning-based methods for synthesising new data points from a learned distribution. Ghenis benchmarked several of these methods against each other and found that a random forest model-based approach minimised *quantile loss*, an indicator of how well the distribution of generated values aligns with the prior distribution, the most.[38] Preserving heterogeneity in distributions is important here: microsimulation modelling's

[36] http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[37] https://www.researchgate.net/publication/314457741_A_Survey_on_Various_Applications_of_Artificial_Neural_Networks_in_Selected_Fields_of_Healthcare

[38] https://towardsdatascience.com/quantile-regression-from-linear-models-to-trees-to-deep-learning-af3738b527c3

core strength comes from its ability to simulate independent outcomes across a highly diverse sample of the population.

Ghenis' findings reproduce well in a new experiment with the Survey of Personal Incomes: random forest models show much lower loss on SPI holdout sets than other imputation methods, including matching, OLS, quantile regression and gradient boosting.[39] Figure 1 shows the results of this experiment, which trained each of five models to predict employment income given age and then measured total quantile loss over seven quantiles.

Add comment emphasising it's assessed on a holdout set

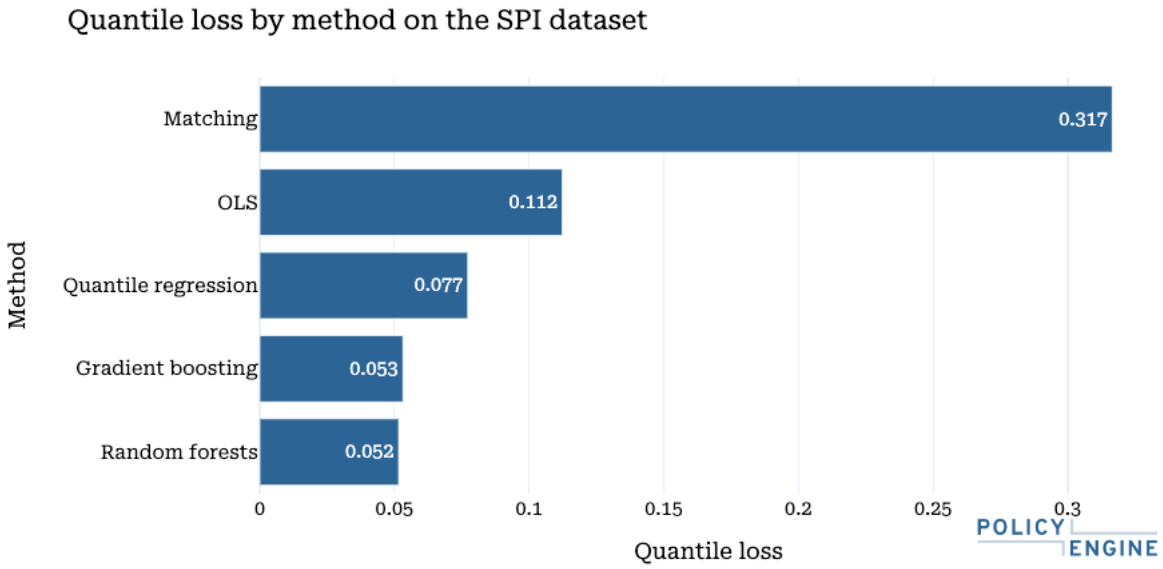

Figure 1: quantile loss on models trained to predict SPI respondent characteristics.

Figure 2 shows more validation for this conclusion: even within individual percentiles of the distribution, matching is significantly worse at predicting those percentiles than random forests.

[39] Source code available (the SPI cannot be publicly shared but can be used if provided by the reviewer in CSV format) at https://colab.research.google.com/drive/1E8F7S1Uvfw_3PmpS226Sl1LWV5NBi0CE#scrollTo=F-t95pZHz6Tr

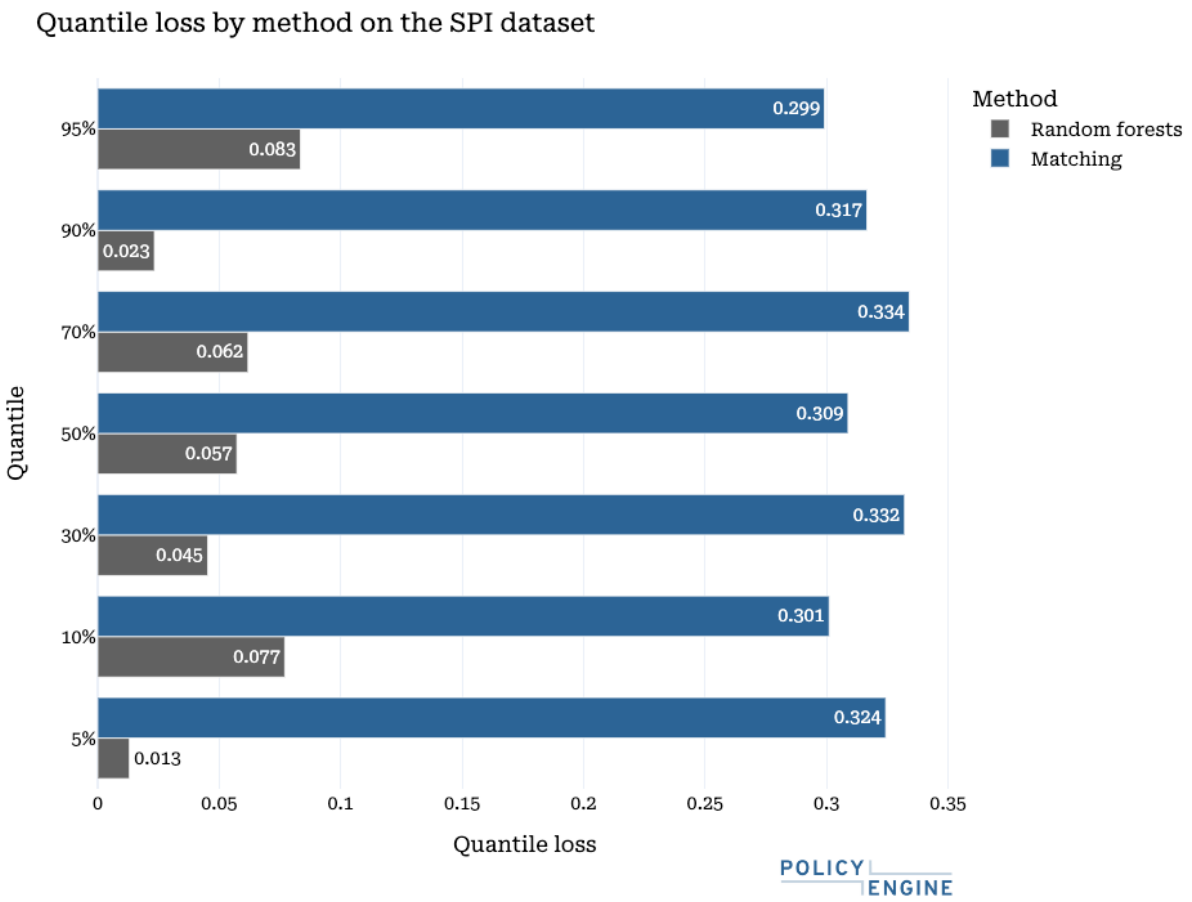

Figure 2: quantile loss by method and quantile for matching against random forests

The results of these experiments validate the use of random forest models within this paper as the imputation model by which we can extend the moldability of the survey by demonstrating that the higher performance found by Ghenis carries forward to the data context that this paper is concerned with (UK taxpayers).

This leaves the question of from which distribution we should synthesise new records. Clearly, this should not be the Family Resources Survey given this would be self-defeating. A data source more likely to produce success would be HMRC's Survey of Personal Incomes: a 1% (anonymised to meet disclosure rules) sample from HMRC's administrative tax records. The SPI does not suffer the same under-counting of incomes (in fact, its aggregates are the basis by which we know the FRS is inaccurate), and the inability of individuals to refuse to participate in the survey means that it is likely to be more representative of the population than the FRS. Therefore, since SPI records can reproduce SPI aggregates, SPI-like records sampled from the SPI in FRS format should be able to move FRS aggregates in the direction of SPI aggregates when given enough weight.

The structural form of the model to use is distinct from how it mechanically integrates with the rest of the survey enhancement pipeline. The exact method we use is set out in the following steps:

1. Identify common variables between the SPI and the FRS.
2. Partition these into two sets: the predictor set, and the imputation set.
3. Train a random forest model to predict imputation set variables given the predictor set variables.
4. Duplicate each household record once, assigning zero weight to the copy.
5. For each household record in the FRS, use the random forest model to predict the imputation set variables.
6. Override the imputation set variables in the copied household record with the predicted values.

Note that it is important that we adjust for the different sampling frames of the FRS and SPI: the SPI only includes individuals who are likely to pay Income Tax, whereas the FRS includes people with earnings too low to be taxed. We can adjust for this by adding in an approximation for the missing (approximately ten million individuals) population to the SPI as a new record with zero income- if we did not do this, the bottom of the income distribution in the SPI might be above zero, which is not true of the actual income distribution.

At the end of this process, the new FRS dataset is identical to the original dataset because the new additions are zero-weighted. However, the optimiser now has a far expanded parameter space to work with and can therefore (potentially) make more progress towards the target.

However, even if we can accurately capture the relationship between variables in a more accurate dataset, this doesn't guarantee that applying the model to the FRS will produce the same record distributions as in the SPI, because the SPI and FRS have different sampling frames: imputing income variables might correct for the FRS undercounting high incomes *among people who would have high incomes if HMRC were asking*, but the FRS also will have less of these people than would exist in the SPI (due to sampling bias).

Another pitfall of the imputation method is that it might not produce the most efficiently loss-potential-reducing records (from the perspective of the space they take up) or might even fail to produce the specific records that exist but not in either survey. For example, many of the individuals with the highest incomes will not appear even in the SPI. How can we ensure these records are created? One naive approach would be to simply craft them by hand, but this is both not scalable, and might introduce inaccuracy if the hand-made records have some unseen internal contradiction that would mean they cannot exist in real life. Instead, we can preserve the strength of our random forest models at capturing relationships between variables, and just adjust the distribution they predict.

Random forest regressors use the average of the results of a tree set as their outputs. The outputs from each tree incorporate all the learned information

about the output variables conditional on the input variables, making it the optimal interception point to modify the model outputs without sacrificing the model's learned relationships. We can from here change the output from an average over all trees to a given percentile of the tree results. Since heterogeneity is still important, this percentile (per household observation) can be sampled randomly from a distribution (evenly and centred at 50% by default). Parametrising this distribution as a Beta distribution allows us to control its skew by a set parameter.[40] Adjusting this skew parameter now enables us to adjust the distribution of predicted values upwards or downwards as needed.

Multivariate prediction might need to make use of separate distribution parameters for individual variables (for example, if dividend income is more likely to be under-represented by sampling bias than employment income). To allow for this while still retaining consistency between predicted variables, we can train individual variable predictive models on the sequential predictor variables. For example, if we are to predict employment income and dividend income from age, we would train one model predicting employment income given age, and another predicting dividend income given age and (previously predicted) employment income. This is a straightforward extension of the univariate case and can be implemented in the same way.

Data lag

The existence of data lag, as identified by Burkhauser et al in the SPI adjustment, in which a household survey data belongs to a previous year (often 2-3 years before the current), presents a problem for its accuracy in many household surveys and this project. If we have a survey with taxes and benefits from 2019, are optimising weights to fit statistics from 2022, the optimiser might struggle to perform well (because we are essentially asking it to sample a realistic set of 2022-like households as best it can from a set of 2019-like households). This problem is made worse the more tax and benefit policy changes between the survey year and the target year (and there have been substantial changes in tax-benefit policy, particularly over the pandemic years).

To fix this, we can use a microsimulation tax-benefit model to correct policy-influenced data in the survey. Such a model is a predictor of a set of tax and benefit-related variables (for example, Income Tax, Universal Credit, etc.) from a set of input variables (e.g., employment incomes, household structure, etc.). Therefore, we can apply this process to each household in the survey data and correct the relevant variables with those simulated by the microsimulation model, according to 2022 policy.

[40] The Beta distribution is parametrised by two parameters, but it is trivial to express this as one.

There are multiple microsimulation models capable of this for UK policy, but for this validation experiment we use the open-source model PolicyEngine-UK[41] which is therefore used in this project.

The FRS has a sample size of around 20,000 households. This can reasonably be expected to be enough for the optimiser to be able to obtain good performance, but there are specific reasons why it might not. The largest likely problem is Universal Credit (UC), the UK's central means-tested benefit. Universal Credit is currently being phased-in across the UK, as a replacement for six previous legacy' benefits, the bulk of which happened between 2019 and 2022. This means that the 2019-20 FRS has a significantly lower share of households claiming Universal Credit (but instead claiming legacy benefits) than would be realistic for 2022. The problem becomes clear if we consider the case of optimising weights to a year in which Universal Credit is fully rolled out: all the 2019 households with legacy benefits would be essentially worthless, because they would have to be zero-rated to not produce erroneous aggregates.

This issue can't fully be solved by data ageing: since there is still some mix of Universal Credit and legacy benefits, the microsimulation model must decide which to simulate based on which benefit the households report already receiving. We could in theory override the households' UC-legacy status to match the rollout percentage, but this might introduce inaccuracy in the household records.

Instead, a cleaner solution is simply to pool multiple years of the survey (before data ageing), to increase robustness against this issue of small sample sizes. For this project, we can combine the 2018-19, 2019-20 and 2020-21[42] datasets (with only the 2019-20 dataset given its initial weights and the others zero).

Overall method

The final, combined pipeline is as follows:

1. Combine the three consecutive FRS years into a single dataset.
2. Duplicate the dataset with zero weights in the new copy records.
3. Train the random forest models on income data from the SPI.
4. Solve for the distribution parameters for each variable such that SPI aggregates are reproduced in the model's weighted FRS predictions.
5. Impute on the FRS and replace-impute income variables from the SPI in the second half of households.[43]

[41] https://github.com/policyengine/policyengine-uk

[42] This data release exists, but we do not use it in the default case due to concerns about reliability given lower participation and telephone interviewing (due to the pandemic). Using it here, only given weight at the whim of the optimiser avoids this issue.

[43] At this point in the process, the resultant dataset contains six version of the FRS, and only the second (the 2019-20 FRS) has nonzero weights.

6. Step through the gradient descent algorithm to minimise survey loss with respect to the household weights of the resultant dataset.
7. Measure the relative change in survey loss from the original dataset.

To measure the success of the process, we can record the loss value from the 2019-20 FRS and compare against the loss for the optimised-pooled-imputed FRS, as well as any alternate versions of the FRS for comparison.

The implementation[44] for this pipeline is in Python, using the PyTorch library for the gradient descent routine and the Scikit-learn library for the synthetic data generation. For the gradient descent optimisation, I trained the weights for 256 epochs with a learning rate of 1 (given the weights are un-normalised by design, and in the order of 1,000). Random forest models were trained with 100 trees and a maximum depth of 10. For training, I used the default Google Colab environment for reproducibility.

Results

The implementation of the survey enhancement routine is publicly available as a Python package (*survey-enhance*), and with public documentation.[45] This package contains the main survey enhancement routines for reweighting and synthetic data generation and includes an example application in the UK context.

Comparison against other methods

To validate the performance of the pipeline (and its components), we need a unified metric for survey accuracy (a mapping of household surveys to real-valued-space). For this, we can use the original loss definition described in the method (for overfitting concerns, see the experiments in REF that validate against this). We can then compare the loss values for each of the following datasets:

- Original 2019-20 FRS: the original dataset, with no changes.
- Percentile-matched (all) FRS: the FRS with percentiles for all major income sources matched to the SPI.
- Percentile-matched (dividends only) FRS: the FRS with percentiles for dividend income only matched to the SPI.
- Percentile-matched (pensioner split) 2019-20 FRS: the FRS with percentiles for all major income sources matched to the SPI, but with separate distributions for pensioners and non-pensioners.
- Gradient descent-based reweighting: the FRS with weights optimised using the gradient descent algorithm.

[44] Full model code, documentation and results is available at https://github.com/policyengine/survey-enhance

[45] Documentation available at https://nikhilwoodruff.github.io/survey-enhance. The package is available on the widely used Python package repository *pypi.org*.

- Imputed and reweighted FRS: the FRS with zero-weighted, FRS-structured, SPI-sampled synthetic records added, and weights optimised using the gradient descent algorithm.

The same hyperparameters for all machine-learning-based models were used between training runs in these comparisons. The table below shows the change in total normalised survey loss under each survey improvement method.

Table 1: loss reductions under different survey improvement methods on the FRS

Adjustment	Loss change
Percentile matching (all)	+3.92%
Percentile matching (pensioner split)	+0.90%
None	0.00%
Percentile matching (dividends only)	-0.13%
Gradient descent-based reweighting	-59.13%
SPI RF imputation and reweighting	-88.00%

The results show that the gradient descent-based reweighting method is the second most effective, reducing the loss by 59%. The imputation and reweighting method is the most effective, reducing the loss by 88%.

The percentile matching methods are less effective, with the pensioner split method being the most effective, reducing the loss by less than 1%. The percentile matching method with all income sources is the least effective, reducing the loss by 3.92%. The percentile matching method with dividends only is the least effective, reducing the loss by 0.13%. There is selection bias with the choice of parameters for the percentile matching methods: I tried many different configurations of the algorithm, in order to find at least one example which reduced survey loss. The likely cause of the inaccuracy is that employment income aggregate is already very close to the SPI aggregate, and so if the method only changes the percentiles in the top 5 percent, then it will overestimate the aggregate by doing so (when instead, we should be reducing the percentiles in the bottom 95 percent to compensate). This issue is absent when only adjusting dividend income because dividends are so poorly captured in the baseline FRS (under-reported by 80%).

Validation on specific metrics

Although a single real-valued loss result is useful for ranking survey improvement methods, it doesn't tell us much about the accuracy of the survey across all the different targets involved in the loss function. Figure 4 shows the distribution parameters (the 10th, 50th and 90th percentiles) of the relative errors against the (unweighted) set of statistical targets involved in the loss function, by training epoch. This shows that the optimisation process was broadly effective at reducing errors jointly across a diverse set of targets (and does not simply rely on hitting a particularly high-weighted target, like the employment income aggregate, at the expense of others).

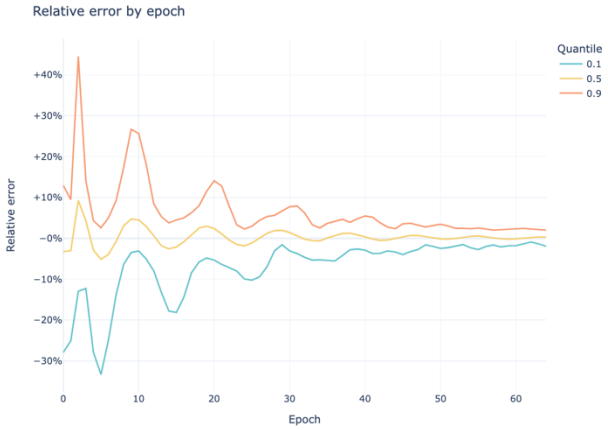

Figure 3: distribution of relative error rates by epoch

Fitting the entire set of statistical targets is challenging because it includes targets that conflict with each other, as well as in harmony. For example, employment income and Income Tax aggregates are correlated (and therefore if the survey moving further towards accurately reproducing one likely moves towards the other as well), but high-income-specific Income Tax aggregates are in conflict with the population of London (if the algorithm tries to hit the high-income-specific Income Tax aggregate by increasing the weights of high-income taxpayers, who are usually resident in London, it might quickly find that it has severely overestimated the regional population). This is a possible explanation for which the error rates in Figure 4 are wave-like: the algorithm is in a constant state of trying to hit one target, but then finding that it has hit another target instead, and then trying to hit the first target again, and so on.

Validation of imputation variable covariance

A key concern with imputing new values for household properties is whether the new values capture similar same covariation behaviour as in the training data. Figure 5 shows the correlation matrix for three datasets:

1. The original SPI training data
2. The predictions of the trained random forest model back on the training data
3. The predictions on the FRS data

Most of the correlation coefficients are similar to the original dataset, though some differ (for example, dividends).

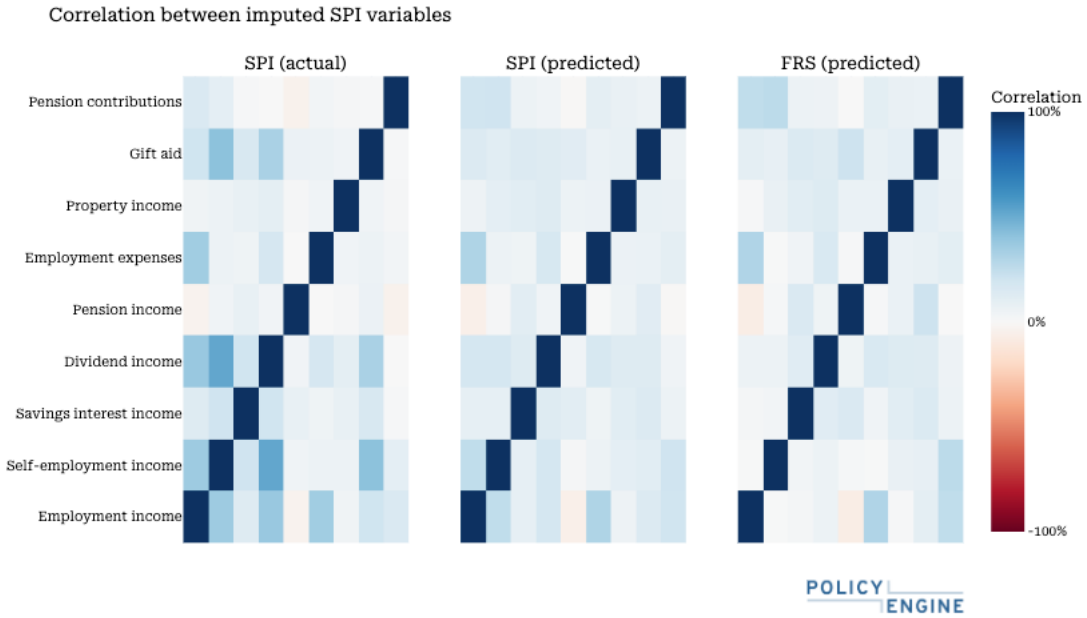

Figure 4: correlation matrices under training and testing data

Reform analysis comparison

Baseline-only inference (how well a microsimulation model can describe the current household sector) is important, but the primary purpose of microsimulation models is to estimate the impact of a hypothetical reform to policy rules. To validate not just the accuracy of the survey improvement method on the current baseline, but in reforms, we can measure how the optimised FRS compares to a source of truth' for reform impacts. For example, the Dividend Allowance is a tax relief that exempts the first £2,000 of dividend income from Income Tax. Table 2 compares microsimulation findings for the net cost of the tax relief[46] using the original and enhanced FRS datasets against the estimates against HM Treasury's internal modelling (which uses the Survey of Personal Incomes, HMRC's administrative dataset).

Table 2: comparison of reform impact estimates

Source	Estimate
HM Treasury (internal)	£720m
Original FRS	£411m
Enhanced FRS	£680m

[46] Defined as the difference in tax revenues compared to a scenario where it did not exist.

While there is no definite ȍource of truth' for hypothetical reforms, the SPI dataset is as close as we can get. This indicates the strength of the enhanced FRS dataset: suggesting that on this reform, it has combined the budgetary accuracy of the SPI (which is incapable of household microsimulation) with the applicability of the FRS.

General evaluation

Measuring the resultant accuracy against the actual data and models used by the established research groups inside and outside government is largely impossible, because all but one of the groups who carry out research using microsimulation do not publish standard model outputs or any validation against external statistics (this includes modelling groups inside and outside government).[47] Only one model publishes validation statistics: UKMOD, managed by ISER at the University of Essex.

UKMOD publishes a comprehensive set of statistics regularly, detailing how the model's tax-benefit aggregate statistics compare to external aggregates.[48] A comparison of these results against the equivalent outputs from PolicyEngine-UK with the data enhancement methodology in this project shows the optimised FRS in this project outperforming UKMOD. Most aggregates in UKMOD's validation set have a relative error in the region of 10 to 30 percent against administrative truth; compared to 0 to 5 percent under the optimised datasets here.

The optimisation process is constrained to non-negative weights, which is a reasonable assumption for a survey. However, experimentation with the optimisation process shows, as expected, that the optimiser can achieve a lower survey loss with this constraint removed. This raises an interesting and unorthodox question: do negative weights have a meaningful interpretation in the context of a survey? We might conclude that the intuitive meaning of the survey weight is not that a household record represents a negative number of households (which does not seem reasonable), but instead that the optimiser is trying to exploit the negativity to *construct a new household record*: from the linear combination of the negative-weight household and some other positive-weighted household in the survey. Without the ability to alter individual household records, reweighting is the only way it can achieve this. Future research exploring this idea, as well as testing if this explanation can be intuitively demonstrated in the data (for example, does this manifest as a simple household acting as a 'missing half', or is the actual function less intuitive?) could substantially increase the performance of the optimised weights and the survey as a whole.

[47] The microsimulation models of note which cannot be used for comparison due to this are: the IFS' TAXBEN, the IPPR model' at PERU, Manchester Metropolitan University, the DWP's PSM, HMRC's IGOTM.

[48] https://www.iser.essex.ac.uk/research/publications/working-papers/cempa/cempa2-22

Other methods exist for generating synthetic data. For example, generative adversarial networks (GANs) have been shown to be highly effective at capturing and reproducing patterns arising from complex compositions of low-level data features. However, this approach (in this case, using conditional GANs) would likely have been less effective, for two reasons. Primarily, random forest models are inherently more robust to particular distributional features such as 'bunching' (recall the capital gains example: tax policies often cause spikes at exact values of income amounts). Neural networks might approximate this, but with less precision than a decision tree. Secondly, modifiability: the random forest model architecture enables the distribution parameter adjustment 'trick', but no such ability exists for GANs without significant modification, or retraining (which is costly).

Explainability is also key for microsimulation modelling. Given that outputs from the model are the result of aggregating tens of thousands of individual household outcomes, being able to interrogate characteristics of individual household records, *and understand why they are a given value*, is essential for the model's results (which can often be counterintuitive) to retain legitimacy. With the random forest model distribution output, it is relatively straightforward to evaluate particular synthetic households. Figure 6 shows an interactive application developed for this purpose in this project, which allows the user to select input values for a given household and simulate the random forest model's distribution outputs for the relevant income variables (reflecting the SPI income imputation stage of the enhancement process).

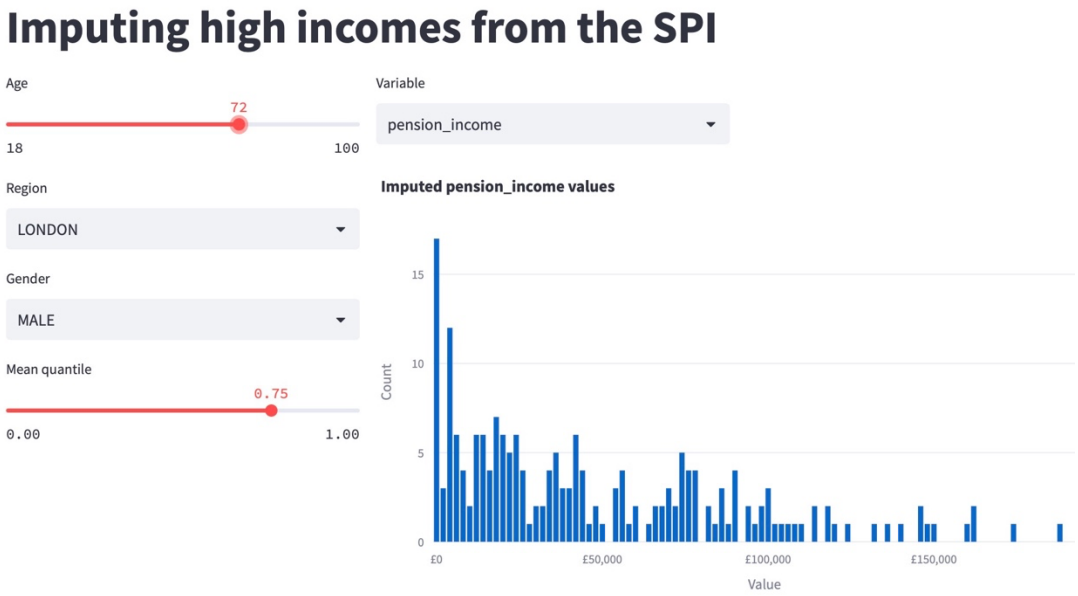

Figure 5: the synthetic data explorer

When directly targeting a set of statistical targets, there is a risk of overfitting to just those targets. For example, if we only fit the weights to tax statistics without also targeting demographic totals, we might find that the household

population is significantly overestimated. This specific issue is solved by adding demographic targets to the loss function, but there will always be out-of-sample statistical targets: cross-tabulations by ethnicity, regional program statistics, etc. To test the severity of this issue, we can exclude a random percentage of the targets from the training loss function, and test how performance on those targets differs from the training targets. Cross-validation can also increase the robustness of this analysis, by rotating the selection used for validation to ensure no individual sample is under-represented in the validation set. Table 3 shows the results of this analysis, using 5-fold cross-validation.

Table 3: cross-validation results

Statistical target set	Loss relative to original FRS
Training targets	-85%
Validation targets	-42%
Combined targets	-76%

This is also significantly better performance than the original FRS, which overfits to demographic targets *by design*: FRS weights being optimised to hit population totals alone causes them to underfit financial targets in the way that they do.

One finding when carrying out this research was of the importance of not segmenting co-dependent targets between training and validation. For example, it was found that if the population targets for all regions except one (e.g. the South West) were in the training set and the remaining one was in the validation set, then the model was able to effectively exploit this by treating the population of the South West as a ṣpare household region' whose population did not matter and whose inhabitants could be used to plug any gaps in other targets as required.

Conclusion

The methods outlined in this project are independent of country profile or policy, requiring only a survey of reasonable accuracy and a set of target statistics of higher accuracy than the survey. The evaluation focussed on the UK, but the very same approaches could be applied to e.g., the United States or other jurisdictions, which often suffer from similar issues around under-reporting and sampling bias.

While the code to reproduce this enhanced FRS is publicly available, the actual dataset is limited to UK academic researchers and non-profit

organisations due to UK data licensing conditions. This limits the value that it can provide, increasing barriers to entry for potential new research and application. Countries like the United States do not impose this limitation (for example, the Current Population Survey, an analogue to the UK's FRS, is publicly available). This project used synthetic data techniques to *extend*, rather than generate from scratch: future research investigating the possibility to create a new, calibrated FRS dataset with all the properties of the enhanced FRS would be a highly valuable contribution to the field, by enabling the resultant dataset to be made completely public.

The results of the experiments on the FRS suggest improvements can be made to the accuracy of tax-benefit microsimulation modelling by adjusting survey weights and values over the original outputs of the statistical agencies that manage them, at least in the context of the UK's DWP and ONS. Whether this logic extends to the data produced in other countries might require similar experiments on the relevant survey microdata. However, given the findings of the wide literature on household survey inaccuracy that the same issues of measurement and sampling bias manifest across countries in a similar way, it seems plausible that the approach in this paper could show similar results.

Generating synthetic household data for this project is a complex task, primarily because the FRS is a relational dataset: households can have variable numbers of members, and members have unique relationships with each other. Graph neural networks might be a promising approach to this problem, as they are able to model complex relationships between data points. However, the current state of the art in this area has not yet reached the same level of maturity as more established methods like regular ANNs, GANs, or random forests as in this implementation.

The positive results achieved in the UK context suggest a strong improvement in the accuracy of the FRS, and by extension the microsimulation model that uses the data. By combining the best of both worldś in granularity (from the FRS) and accuracy (from the SPI and other data sources), the pipeline effectively enables much more accurate (and consistent) microsimulation modelling of the UK tax-benefit system. Improving the accuracy of microsimulation modelling brings potentially sizeable improvements in the abilities of policymakers to understand the likely impacts of reforms to tax-benefit policy, and ultimately improve the effectiveness of such reforms in achieving their stated aims.

This project benefited from discussions and presentations made to other industry stakeholders across government, non-profit and for-profit sectors. I presented the methodology and its applications to the HM Treasury internal distributional analysis team, who provided highly useful feedback and shed light on possible future applications. Other organisations and individuals also provided useful feedback and ideas for applications outside the scope of this project: for example, how using higher quality microdata for the random forest model imputations might affect the resultant optimised survey data quality.

Acknowledgements

The authors would like to thank: my supervisor at the University of Durham, Professor Iain Stewart, for his guidance and support throughout this project; the many individuals and organisations who provided feedback and thoughts on the methodology and ideas behind it, including the HM Treasury distributional analysis team; the Joseph Rowntree Foundation modelling team; Matteo Richiardi at ISER and the IPPR model maintainers.