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

Student Name: Nikhil Woodruff
Supervisor Name: Professor Iain Stewart
Submitted as part of the degree of BSc Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—Microsimulation over survey datasets remains the dominant method 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 state-of-the-art methods including percentile adjustment. I find that the proposed approach is able to 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).

Index Terms—Machine learning, household surveys, microsimulation, survey error, survey weighting

1 Introduction

Governments allocate trillions of dollars every year through mechanisms, such as tax and benefit benefit policy. Decision-making around the design of policies capturing and directing these resources is largely informed by simulated experiments (*microsimulation*) that predict the impact of these policies, or their counterfactuals, on the general population by emulating tax and benefit legislation for respondents in large household surveys.

1.1 Household surveys

Statistics agencies around the world collect data on the characteristics of the population through household surveys. As well as serving as an input to government policy evaluation, these data provide detailed insights into the current properties of the UK household sector, informing government publications on poverty, the disposable income distribution and inequality. However, there is evidence to suggest that these surveys are at least partially inaccurate, and this inaccuracy may be more consequential for certain uses of the survey data.

In the United Kingdom, the most comprehensive household survey is the Family Resources Survey (FRS). This survey is collected annually by the Department for Work and Pensions, and includes approximately 20,000 households in a given 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 representative of.

However, despite a weight generation procedure that aims to ensure accurate representation of the UK household sector, prior research has found that the survey does not accurately predict household statistics well at the very-low or very-high ends of the income distribution: benefit aggregates derived from the FRS are around 20% lower than the in administrative benefit databases, and top income percentiles are lower than than in HMRC's administrative tax database.

There are two ways that a survey can diverge from the reality of the target population: sampling error and measurement error. In the microdata, measurement error arises where the values of individual observations are incorrect (not necessarily through deliberate deceit from survey respondents- question design and integrity checks can influence the accuracy of responses). Sampling error arises where the sample of respondents is not representative of the target population, primarily manifesting as an incorrect weighting of the household records. Both of these sources of error interact with each other: correcting for a missing household subsector can either be achieved by increasing the weight of an existing record or by transforming the values of another record.

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 particular 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.

Many existing methods for mitigating these types of inaccuracy make heavy use of 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. 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, income percentiles are a small fraction, and other targets could plausibly be thrown off by this adjustment.

This project aims to evaluate a novel use of a combination of machine learning-based techniques to correct for household survey inaccuracy from both sampling and measurement error, and to compare its performance against other state-of-the-art methods. This approach involves using random forest model inference to synthesise new records, and then using gradient descent to minimise a constructed loss function that measures the deviation of the household survey from a large set of official statistics.

2 Related Work

This section presents a survey of existing work on the problems that this project addresses. It should be between 2 to 3 pages in length. The rest of this section shows the formats of subsections as well as some general formatting information for tables, figures, references and equations.

Over the last few decades, there has been extensive research into the accuracy of household surveys for estimating socioeconomic and policy-related indicators, as well as methods of improving survey accuracy. Most of these studies have focussed on one particular mechanism by which surveys introduce inaccuracy (for example, by omitting top incomes or under-sampling low incomes), and examine a method of improving surveys which tackles this particular flaw. This literature survey aims to provide a comprehensive overview of the state of the art in improving survey accuracy, while also examining how and if these individual advancements complement each other.

2.1 Current approaches in economic surveys

It is well known that household surveys produce inconsistent results to other data sources, such as administrative databases. Given the nature of how surveys are conducted (households must first consent to an interview, and secondarily must answer truthfully to questions asked), this inaccuracy can be introduced either by sampling error or measurement error (likely both, to some extent). Over the last few decades, household surveys have become the dominant tool in measuring and projecting economic impacts of policy changes, and as such, there has been a great deal of research into improving survey accuracy.

2.1.1 Under-coverage of high incomes

The Department for Work and Pensions is required by law to report on poverty and inequality metrics every year, and in meeting this requirement, it publishes a household-level dataset of disposable incomes, termed the Households Below Average Income (HBAI) dataset. [1] Since 1992, it has applied an adjustment to the disposable incomes of a subset of the dataset in order to make the coverage of top incomes more comparable with that of HMRC's Survey of Personal Incomes (SPI) dataset - this adjustment termed the 'SPI adjustment'. In [2], the authors examine the methodology of this adjustment, as well as its performance against its original goals.

The authors document[1] the steps of the SPI adjustment, which involve first identifying a set of 'rich' households. The definition of rich applies a condition that a household's income must be above a certain threshold, where separate thresholds are used for pensioner and non-pensioner households. The target used to set thresholds is generally to ensure that around 0.5% of records are altered, varying by year. The HBAI 'rich' households are then modified by replacing gross incomes (an income measure which the SPI also contains) with the average values for records in the same group in the SPI. Finally, the survey weights are recalculated: in the original survey, weights are solved by matching administrative statistics on population sizes; under the SPI adjustment, population sizes of the 'rich' groups are included in the set of statistical targets to hit. The authors find that the SPI adjustment has been successful in improving the coverage of top incomes in the HBAI dataset, but raise a number of issues:

2.1.1.1 Income decomposition: The SPI adjustment is applied to a singular income variable, but the FRS contains a number of components. Modifying gross income, but not modifying employment income, savings income, etc. breaks the link between these variables, which prevents researchers from conducting decomposition analyses.

2.1.1.2 Stratification: There is no obvious justification for separate thresholds for pensioners and non-pensioners (and further, between households in Great Britain and Northern Ireland). The authors suggest these stratification choices were made in order to minimise methodological changes over time, for example as the survey expanded to Northern Ireland.

2.1.1.3 SPI lag: The Survey of Personal Incomes is not routinely available at the same time as the Family Resources Survey (from which the HBAI microdata is derived). Therefore the SPI adjustment is applied to the HBAI dataset using a lagged SPI dataset, which may introduce additional inaccuracy.

2.1.2 Adjustments using administrative tax data

For the 2019 edition of the Households Below Average Income series, the ONS published details of the methodology used to tune the dataset with the SPI in [3]. They respond to some of the concerns raised by [2]:

2.1.2.1 Pensioner stratification: The authors show that high-income pensioners and non-pensioners are both under-represented in their respective populations but comparing the ratios of incomes at different quantiles, finding that a common threshold for both groups would fail to ensure that pensioners (who have lower income, on average) are sufficiently affected by the SPI adjustment.

2.1.2.2 Choice of income threshold: The authors discuss possible justifications for a particular income threshold, mostly based on the quantile at which divergence between the FRS and SPI 'became an issue'. However, the choice to use a binary variable (rather than, for example, phasing in an SPI adjustment) here is arbitrary, and the authors do not address the reasons why this choice was made.

1. Previously, the DWP had not published its research underlying the methodology of the SPI adjustment

2.1.2.3 SPI lag: The authors acknowledge the issue of using SPI projections, rather than actual outturn data, and examine the size of this effect. They find that revising recent survey releases with the actual SPI data later released changed the Gini coefficient of income inequality estimates by around 0.2 percentage points. This is considered to be small and therefore recommend against the need for the ONS to re-publish statistics when current SPI data becomes available.

2.1.3 Capital income imputation

The issue of income decomposition remained largely untackled until [4], in which the authors attempt to improve the reporting of a specific component of gross income which is more severely under-reported in the FRS than others: capital income. They first establish that income under-reporting is mostly due to 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 in order to match the totals in SPI data.

The authors find 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.

2.1.4 Under-coverage of very low incomes

In [5], the authors examine the other end of the income spectrum, finding that very low-income households tend to spend much more than moderately low-income households in the Living Cost and Food Survey (a household survey with similar administration to the FRS). The authors report a variety of evidence that income at the low end is misreported in the survey:

2.1.4.1 Missing benefit spending: By comparing total reported receipt of benefits by recipients with aggregate spending figures published by the DWP and HMRC, the authors find that the FRS and LCFS consistently underreport benefit income by around 5%, and that this figure has become worse over the last decade, rising from 2.5% in 2000.

2.1.4.2 Sub-minimum wage reporting: In the LCFS, individuals report both hours worked and annual earnings, enabling researchers to calculate the implied hourly wage. For 10.5% of individuals in 2009, this was below the legal minimum wage. Although this does not guarantee a breach of employment law,[2] the proportion is substantial and implies that either earnings are under-reported or hours worked are over-reported. The authors use a model of consumption smoothing to determine whether the overly high spending (compared to income) for low-income households can be explained by lifetime consumption smoothing, but find that this is not the case.

2. Employers can count some in-kind benefits as payment towards the minimum wage, and there are other legal exceptions.

2.1.5 Linking data directly to administrative data

All of the previously covered research into survey inaccuracy has identified a common question: how much of the survey error is due to non-response bias, and how much is due to measurement error? In [6], the authors attempt to quantify the measurement error of the FRS by linking individual households with data from the DWP's administrative records, using non-public identifiers. 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 adults recipients for whom no link could be found, and the percentage of adults represented only by administrative data.

The authors find 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.

2.1.6 Linear programming

Linear programming, a mathematical technique for solving linearly constrained optimisation problems, is commonly used to determine survey weight values, where the criteria are defined maximum deviations from top-level demographic statistics. In [7], 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. In both of [8] and [9], 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.

2.2 Applicable machine learning techniques

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,

the percentile adjustment methodology in [3], which explicitly partitions households into 'rich' and 'non-rich' using arguably arbitrary definitions). In other domains, for example image classificaion, a move away from prescriptive methods towards loss function minimisation has seen substantially improved accuracy and robustness. [10]

2.2.1 Gradient descent

Gradient descent is a technique for finding parameters which minimise a loss function, by iteratively updating the parameters in the direction of the steepest negative gradient. [11] 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. [12] 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. [13]

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, [14] AdaGrad, [15] and RMSProp.

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.

2.2.2 Random forest models

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. [16] 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 [17] or transformers [18]), 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 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. [19]

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 [20] the authors found that capital gains variables are 'unnaturally' distributed in order to respond to incentives set by particular tax law parameters.

3 Methodology

3.1 Loss

3.1.1 Concept

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:

$$L(S) = \sum_{c \in C} L_c(S) \tag{1}$$

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:

$$L_c(S) = w * (\frac{\sum_i^N X_i \cdot W_i}{y} - 1)^2 \tag{2}$$

where X_i is the value (of a particular variable) of the ith household record, W_i is the weight of the ith 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.

3.1.2 Normalisation

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:

$$L(S) = \frac{\sum_{c \in C} L_c(S)}{L(S_0)} \tag{3}$$

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[3].

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 ESA†
    j) Income-based JSA†
    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 * contain statistics that are also used to determine the original FRS survey weights. Categories marked with † 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 categories is approximately defined by:

$$\begin{aligned} L_{Programs}(S) = {} & 45 \times 10^9 L_{UniversalCredit}(S) \\ & + 11 \times 10^9 L_{ChildBenefit}(S) \\ & + \dots \\ & + 200 \times 10^9 L_{IncomeTax}(S) \end{aligned} \tag{4}$$

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

3.2 Gradient descent

A given survey S is itself a set of variables $X_{i,j}$ (where i is the household record and j is the variable), as well as household weights W_i.[4] 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 $L(S) = L(X, W)$. Our loss minimisation task therefore becomes finding the solution to Equation 5.

$$\frac{\partial L(X, W)}{\partial W} = 0 \tag{5}$$

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 differentation packages such as PyTorch. [21] Under the gradient descent algorithm [17], the weights are iteratively updated in the direction of the steepest negative gradient, until the loss function is minimised.

3.3 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 is capable of reaching. To circumnavigate this barrier, we must add new records and weights to the parameters optimised by the gradient descent algorithm.

Synthesing 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. In [22], Ghenis benchmarked several of these methods

3. Income Tax revenue in 2022 was around £200 billion, while Child Benefit outlays were around £10 billion.

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

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. Preserving heterogeneity in distributions is important here: microsimulation modelling's core strength comes from its ability to simulate independent outcomes across a highly diverse sample of the population.

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 are able to 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.

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.

3.3.1 Distribution adjustment

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 in reality 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 also 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 centered at 50% by default). Parametrising this distribution as a Beta distribution allows us to control its skew by a set parameter.[5] Adjusting this skew parameter now enables us to adjust the distribution of predicted values upwards or downwards as needed.

3.3.2 Multivariate prediction

Multivariate prediction might need to make use of separate distribution parameters for individual variables (e.g. 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.

3.4 Data ageing

The existence of *data lag*, 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 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, in particular 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 (e.g. Income Tax, Universal Credit, etc.) from a

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

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.

There are multiple microsimulation models capable of this for UK policy, though only one is both publicly available and open-source: PolicyEngine-UK [23], which is therefore used in this project.

3.5 Data pooling

The FRS has a sample size of around 20,000 households. This can reasonably 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 in order 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), in order to increase robustness against this issue of small sample sizes. For this project, we can combine the 2018-19, 2019-20 and 2020-21[6] datasets (with only the 2019-20 dataset given its initial weights and the others zero).

3.6 Overall method

The final, combined pipeline is as follows:

1) Combine the three consecutive FRS years into a single dataset.
2) Duplicate the dataset.
3) Train the random forest models on income data from the SPI.
4) Solve for the distribution parameters for each variable s.t. 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.[7]
6) Step through the gradient descent algorithm to minimise survey loss w.r.t. the household weights of the resultant dataset.
7) Measure the relative change in survey loss from the original dataset.

In order 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.

4 Results

Table 1 shows the change in total normalised survey loss under each survey improvement method.

TABLE 1
Loss reductions under different survey improvement methods

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 + reweighting	-88.00%

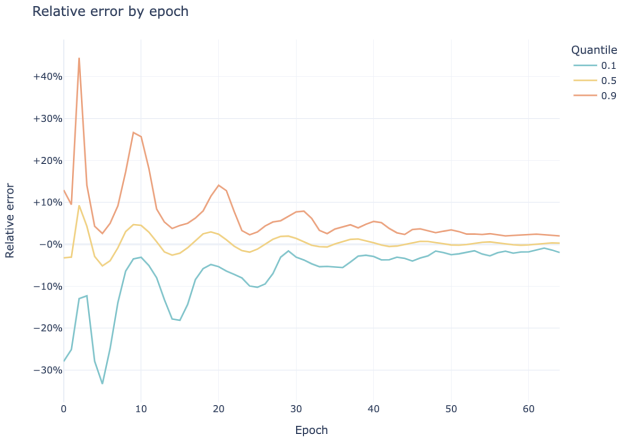

Fig. 1. Distribution of relative error rates by epoch

5 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)[8]. Only one model publish 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

6. 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.

7. 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.

8. 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.

statistics compare to external aggregates. [24] 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.

6 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.

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- and 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.

References

[1] DWP, "Households below average income (hbai) statistics," *GOV.UK*, 1992. [Online]. Available: https://www.gov.uk/government/collections/households-below-average-income-hbai–2

[2] R. V. Burkhauser, N. Hérault, S. P. Jenkins, and R. Wilkins, "Survey under-coverage of top incomes and estimation of inequality: What is the role of the uk's spi adjustment?" *Fiscal Studies*, vol. 39, no. 2, pp. 213–240, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-5890.12158

[3] R. Tonkin, D. Webber, O. Beha, M. Shine, and C. Clark, "Top income adjustment in effects of taxes and benefits data: methodology," *ONS*, 2020. [Online]. Available: https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/compendium/ economicreview/february2020/ topincomeadjustmentineffectsoftaxesandbenefitsdatamethodology

[4] T. C. Ooms, "Correcting the underestimation of capital incomes in inequality indicators: with an application to the uk, 1997–2016," *Social Indicators Research*, vol. 157, no. 3, pp. 929–953, 2021. [Online]. Available: https://doi.org/10.1007/s11205-021-02644-4

[5] M. Brewer, B. Etheridge, and C. O'Dea, "Why are households that report the lowest incomes so well-off?" *The Economic Journal*, vol. 127, no. 605, pp. F24–F49, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12334

[6] S. McKay, "Evaluating approaches to Family Resources Survey data linking." [Online]. Available: https://www.gov.uk/government/publications/family-resources-survey-data-linking-wp110

[7] C. Lound and P. Broad, "Initial review of the family resources survey weighting scheme," *Office for National Statistics*, 06 2013.

[8] T. P. Center, "Tax model documentation," *TPC*, 09 2022. [Online]. Available: https://www.taxpolicycenter.org/resources/brief-description-tax-model

[9] P. S. Library, "Tax-data model documentation," *GitHub*, 2020. [Online]. Available: https://github.com/pslmodels/taxdata

[10] D. Lu, "A survey of image classification methods and techniques for improving classification performance," *International Journal of Remote Sensing*, vol. 28, pp. 823 – 870, 03 2007.

[11] P. Baldi, "Gradient descent learning algorithm overview: a general dynamical systems perspective," *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 182–195, 1995.

[12] Y. Wen, K. Luk, M. Gazeau, G. Zhang, H. Chan, and J. Ba, "An empirical study of stochastic gradient descent with structured covariance noise," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 26–28 Aug 2020, pp. 3621–3631. [Online]. Available: https://proceedings.mlr.press/v108/wen20a.html

[13] S. Khirirat, H. R. Feyzmahdavian, and M. Johansson, "Mini-batch gradient descent: Faster convergence under data sparsity," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 2017, pp. 2880–2887.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[15] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. 61, pp. 2121–2159, 2011. [Online]. Available: http://jmlr.org/papers/v12/duchi11a.html

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[17] W. Schöllhorn and J. Jäger, "A survey on various applications of artificial neural networks in selected fields of healthcare," *Neural Networks in Healthcare: Potential and Challenges*, pp. 20–58, 01 2006.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/ 3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[19] B. Jarmulska, "Random forest versus logit models: which offers better early warning of fiscal stress?" *European Central Bank Working Papers*, 05 2020. [Online]. Available: https://www.ecb.europa.eu/pub/ pdf/scpwps/ ecb.wp2408 aa6b05aed7.en.pdf?9551c7c6e8e8fdbd35e5512b5afcf097

[20] T. Dowd and R. McClelland, "The bunching of capital gains realizations," *National Tax Journal*, vol. 72, no. 2, pp. 323–358, 2019. [Online]. Available: https://doi.org/10.17310/ntj.2019.2.02

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[22] M. Ghenis, "Quantile regression, from linear models to trees to deep learning," 10 2018. [Online]. Available: https://towardsdatascience.com/quantile-regression-from-linear-models-to-trees-to-deep-learning-af3738b527c3

[23] N. Woodruff, "Policyengine-uk microsimulation model," *GitHub*, 2020. [Online]. Available: https://github.com/policyengine/policyengine-uk

[24] D. Collado, D. Popova, and M. Eshraghi, "Ukmod country report 2019-2025," *CeMPA Working Paper Series*, no. CEMPA2/22, 2022. [Online]. Available: https://www.iser.essex.ac.uk/research/publications/working-papers/cempa/cempa2-22