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 under-report 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

---

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

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.

2.1.5 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.

2.3 Main Text

The font used for the main text should be Palatino and the font size should be 9.5. The first line of all paragraphs should be indented by 0.5cm, except for the first paragraph of each section, subsection, subsubsection etc. (the paragraph immediately after the header) where no indentation is needed.

2.4 Figures and Tables

In general, figures and tables should not appear before they are cited. Place figure captions below the figures; place table titles above the tables. If your figure has two parts, for example, include the labels "(a)" and "(b)" as part of the artwork. Please verify that figures and tables you mention in the text actually exist. Make sure that all tables and figures are numbered as shown in Table 1 and Figure 1 below.
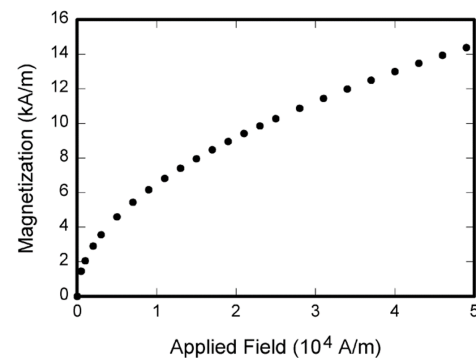

Fig. 1. Magnetization as a function of applied field. There is a period after the figure number, followed by two spaces. It is good practice to explain the significance of the figure in the caption

3 METHODOLOGY

This section presents the solutions to the problems in detail. The design and implementation details should all be placed in this section. You may create a number of sub-sections, each focusing on one issue.

This section should be between 4 to 6 pages in length. *Statements that serve as captions for the entire table do not need footnote letters. E.g. Mx = maxwell, G = gauss, Oe = oersted; Wb = weber, V = volt, s = second, T = tesla, m = meter, A = ampere, J = joule, kg = kilogram, H = henry.*

4 RESULTS

This section presents the results of the solutions. It should include information on experimental settings. The results should demonstrate the claimed benefits/disadvantages of the proposed solutions.

This section should be around 2 pages in length.

TABLE 1
Units for Magnetic Properties

Symbol	Quantity	Conversion from Gaussian and CGS EMU to SI [a]
Φ	magnetic flux	$1 \text{ Mx} \rightarrow 10^{-8} \text{ Wb} = 10^{-8} \text{ V·s}$
B	magnetic flux density, magnetic induction	$1 \text{ G} \rightarrow 10^{-4} \text{ T} = 10^{-4} \text{ Wb/m}^2$
H	magnetic field strength	$1 \text{ Oe} \rightarrow 10^3/(4\pi) \text{ A/m}$
m	magnetic moment	$1 \text{ erg/G} = 1 \text{ emu}$ $\rightarrow 10^{-3} \text{ A·m}^2 = 10^{-3} \text{ J/T}$
M	magnetization	$1 \text{ erg/(G·cm}^3) = 1 \text{ emu/cm}^3$ $\rightarrow 10^3 \text{ A/m}$
$4\pi M$	magnetization	$1 \text{ G} \rightarrow 10^3/(4\pi) \text{ A/m}$
σ	specific magnetization	$1 \text{ erg/(G·g)} = 1 \text{ emu/g} \rightarrow 1 \text{ A·m}^2/\text{kg}$
j	magnetic dipole moment	$1 \text{ erg/G} = 1 \text{ emu}$ $\rightarrow 4\pi \times 10^{-10} \text{ Wb·m}$
J	magnetic polarization	$1 \text{ erg/(G·cm}^3) = 1 \text{ emu/cm}^3$ $\rightarrow 4\pi \times 10^{-4} \text{ T}$
χ, κ	susceptibility	$1 \rightarrow 4\pi$
χ_ρ	mass susceptibility	$1 \text{ cm}^3/\text{g} \rightarrow 4\pi \times 10^{-3} \text{ m}^3/\text{kg}$
μ	permeability	$1 \rightarrow 4\pi \times 10^{-7} \text{ H/m}$ $= 4\pi \times 10^{-7} \text{ Wb/(A·m)}$
μ_r	relative permeability	$\mu \rightarrow \mu_r$
w, W	energy density	$1 \text{ erg/cm}^3 \rightarrow 10^{-1} \text{ J/m}^3$
N, D	demagnetizing factor	$1 \rightarrow 1/(4\pi)$

TABLE 2
Summary of Page Lengths for Sections

Section	Number of Pages
I. Introduction	2
II. Related Work	2-3
III. Methodology	4-6
IV. Results	2
V. Evaluation	1-2
VI. Conclusion	1

5 EVALUATION

This section should be between 1 to 2 pages in length.

6 CONCLUSION

This section summarises the main points of this paper. Do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. This section should be no more than 1 page in length.

The page lengths given for each section are indicative and will vary from project to project but should not exceed the upper limit. A summary is shown in Table 2.

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