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.

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.

2.1 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.

2.2 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.

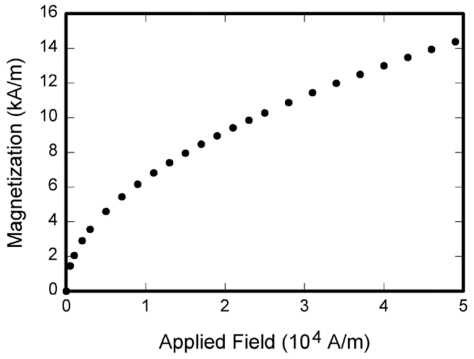

Fig. 1. Magnetization as a function of applied field. There is a period after the figure number, followed by two spaces. It is good practice to explain the significance of the figure in the caption

2.3 References

The list of cited references should appear at the end of the report, listed by order of appearance in the paper. The required style is to note citations in individual brackets, followed by a comma, e.g. "[1], [5]" (as opposed to the more common "[1, 5]" form.) Citation ranges should be formatted as follows: [1], [2], [3], [4] (as opposed to [1]-[4]). When citing a section in a book, please give the relevant page numbers [2]. In sentences, refer simply to the reference number, as in [3]. Do not use "Ref. [3]" or "reference [3]" At the beginning of a sentence use the author names instead of "Reference [3]," e.g., "Smith and Smith [3] show ... .".

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.

TABLE 1
Units for Magnetic Properties

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

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.

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.

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

REFERENCES

[1] J.S. Bridle, *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition*, Neurocomputing—Algorithms, Architectures and Applications, F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989. (Book style with paper title and editor)

[2] W.-K. Chen, *Linear Networks and Systems*, Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style)

[3] H. Poor, *A Hypertext History of Multiuser Dimensions*, MUD History, http://www.ccs.neu.edu/home/pb/mud-history.html. 1986. (URL link *include year)

[4] K. Elissa, *An Overview of Decision Theory*, unpublished. (Unpublished manuscript)