Data Representativity for Machine Learning and AI Systems

LINE H. CLEMMENSEN*, Technical University of Denmark, Denmark
RUNE D. KJÆRSGAARD, Technical University of Denmark, Denmark

Data representativity is crucial when drawing inference from data through machine learning models. Scholars have increased focus on unraveling the bias and fairness in the models, also in relation to inherent biases in the input data. However, limited work exists on the representativity of samples (datasets) for appropriate inference in AI systems. This paper analyzes data representativity in scientific literature related to AI and sampling, and gives a brief overview of statistical sampling methodology from disciplines like sampling of physical materials, experimental design, survey analysis, and observational studies. Different notions of a 'representative sample' exist in past and present literature. In particular, the contrast between the notion of a representative sample in the sense of coverage of the input space, versus a representative sample as a miniature of the target population is of relevance when building AI systems. Using empirical demonstrations on US Census data, we demonstrate that the first is useful for providing equality and demographic parity, and is more robust to distribution shifts, whereas the latter notion is useful in situations where the purpose is to make historical inference or draw inference about the underlying population in general, or make better predictions for the majority in the underlying population. We propose a framework of questions for creating and documenting data, with data representativity in mind, as an addition to existing datasheets for datasets. Finally, we will also like to call for caution of implicit, in addition to explicit, use of a notion of data representativeness without specific clarification.

Additional Key Words and Phrases: data representativity, machine learning, sampling strategies, artificial intelligence, distribution shift

1 INTRODUCTION/MOTIVATION

In 1979-1980 Kruskal and Mosteller wrote four papers on the term 'representative sampling' with the motivation to unravel its ambiguities and imprecision [28–31]. In addition, they called for caution as well as additions of more specific expressions when referring to a representative sample. As they noted: "The reason for so much effort on one term is that the idea of representativeness is closely related to basic notions of statistical inference". As we draw conclusions from data or make predictions in artificial intelligence (AI) systems trained on data, it is important to understand what these data represent, and which inferences we can make. AI systems or machine learning (ML) models for decision making are widely used in industry and research, but care is not always put to the origin of the data, on which the systems are trained. This is for example seen in big data, where more data are considered better, and data often originate from a historical collection performed for e.g. control purposes or from scraping available internet sources rather than having been collected for the purpose, which it is later used for [4, 6, 23, 32]. Other examples are more general for ML/AI and include representation bias stemming from the way we define and sample from a population, evaluation bias stemming from benchmarks datasets with inherent biases, population bias when attribute distributions are different in the dataset and the target population, and sampling bias stemming from non-random sampling of subgroups [36, 40, 48].

Amongst other, Kruskal and Mosteller found that 'representative sample' was used as an assertive to underline a point without any scientific reasoning. Historically, the ImageNet competition has had the same kind of unconscious tale to it, where scientists believed good results on the ImageNet dataset would mean good results for other image recognition tasks as well [12, 36]. Torralba and Efros empirically illustrated in their paper 'Unbiased Look at Dataset Bias' from 2011 that such generalizations are not necessarily given, and described their findings as "...if we add training data that does not match the biases of the test data this will result in a less effective classifier" [52].

Recently, focus has been put on the lack of transparency around dataset design and collection procedures as well as efforts to unbias existing datasets like e.g. the ImageNet [36, 56]. We will investigate the notions of a representative sample in some of these recent initiatives within the AI community as well as related communities applying AI techniques to their domain. We have found sampling theories from the disciplines of analysis of physical material, design of experiments, as well as surveys in social sciences useful in terms of analyzing current practices and relating these to the ongoing work within AI, where the historical emphasis on data representativity has been smaller.

The rest of the paper is organized as follows. First, going through literature about representative sampling, we will outline the general notions of a 'representative sample' (Section 2) and analyze today's notions of a representative sample in ML and AI literature, together with examples in adjacent application domains, e.g. applications of AI in healthcare (Section 3). Then we move on to make a brief overview of the statistical methodologies used today for representative sampling (Section 4). Throughout these investigations, we find opposing opinions of sampling for coverage of the input space vs probability sampling mimicking population distributions. Therefore, we make empirical investigations to demonstrate the qualities these opposing data representations hold (Section 5), and finally suggest a framework for addressing data representativity in datasheets (Section 6). We round of with a discussion (Section 7).

2 NOTIONS OF A 'REPRESENTATIVE SAMPLE'

Kruskal and Mosteller identified six notions/usages of a 'representative sample' in their first surveys from 1979 [28, 29]: An assertive acclaim, absence of selective forces, a miniature of the population, an observation 'typical' or 'ideal' of the (sub)population, coverage of a population by the sample, and a reference to a sampling method later on specified in details. The sixth is a special notion in scientific writing, whereas the first five were found in both non-scientific as well as scientific writing. We will use this framing here, and link more recent literature to these.

The assertive claim (the Emperor's new clothes) has been described in the introduction, and is dangerous both as a conscious acclaim and a subconscious notion when it comes without specification. It is recommended to avoid unjustified and unspecified use.

The miniature (the model train set) population has strong ties to the theory of sampling of physical material also related to chemical or biological analysis [20, 43]. One of the guiding principles in the theory of sampling is to have as homogeneous a population (lot) as possible, and pre-mixing of the material before sampling makes a small sample resemble the lot and minimizes sampling errors. The pre-mixing of a lot is also closely related to the notion of probability sampling. In other fields, it is common to subdivide the space into smaller groups, until each group exhibits homogeneity, and then randomly sample a miniature or a sample representative of that group. However, this is difficult if the population values/distributions are unknown. Recently, Yang et al (2020) [56] proposed a framework to balance the demographics of ImageNet, but they also stated that this is only possible for one attribute at a time, as sub-categories will have too few samples if balancing across multiple attributes (e.g. race and gender). In consequence, the miniature analogy in itself breaks down, as we cannot account for all factors in the miniature, in particular not as the miniature decrease in size. Kruskal and Mosteller wrote: "We do not feel it wrong to use 'representative sample' for the miniature, but rather that it understates the attractive properties of such a sample... a miniature is usually constructed purposefully rather than through a process of probability sampling." [29].

Absence of selective forces (justice balancing the scales) has the sense that the sample is random as no forces are in play to select or de-select any specific types of observations in the target population; implying the purpose is to make inference about the target population, not the sample. This notion ties to experimental modeling and coverage as follows. In the design of experiments literature, controllable factors and uncontrollable factors are distinguished [39]. The controllable ones are indeed controlled to design as small an experiment as possible, yet with a suitable amount of observations and an appropriate coverage of the input space in order to make inference and optimize the response/output as a function of the controllable factors. Too many controlled factors make it hard to access all cross populations, and in addition there is no way of exhausting all possibilities. Selective factors can also be uncontrollable or in worst case go unnoticed. Examples of these are time-drifts in a production or non-response in surveys. These can pose problems to the statistical inference drawn from data. If observable, we can manage through our sampling design or sometimes even through post processing of data. However, unobserved or even unnoticed factors impose serious risks of bias and confounding In surveys, non-response is considered a substantial source of error caused by selection, one that is not directly related with the sampling. Selective forces can also influence survey responders through e.g. an interviewer effect. Errors stemming from such selective forces can lead to potential biases, and several corrective efforts are usually applied to adjust for these [18]. Selective sampling can also be performed on purpose, in survey sampling such examples are: quota sampling, purposive sampling, and referral sampling. These sampling designs are non-random and generalizations are therefore challenged, but sometimes samples of interest are so few, or participation recruitment so difficult, that convenience sampling designs can come in handy [18].

Typical/ideal (Superman/Superwoman or the average man/woman) refers to typical or ideal exemplars which represent a population or subgroups of a population. This is not necessarily in a statistical sense, but may mean close to the average. An example is that in [33], where cluster centers from Gaussian mixture models are sampled as representative observations of a larger dataset. In addition, a ML method like archetypal analysis [9] carries some of this notion: Archetypes in the data are identified, and all other observations are described through linear combinations of these archetypes.

Coverage (Noah's Ark) seeks to include the heterogeneity of the population in the sample. A strong requirement for coverage would be that the sample should contain at least one observation from each relevant partition of the population. In contrast to the miniature, coverage does not require proportions within partitions to match those of the population. Harry V. Roberts suggested in 1971 sampling following the coverage notion in order to select a committee and avoid conscious and unconscious biases from appointing authorities [44]. Along these lines, coverage is more about producing 'representativeness' than about obtaining a 'representative sample' in the statistical sense.

Reference to sampling, later on specified. With this notion, the term 'representative sample' in itself becomes a 'vague term', and the exact meaning is specified in the context. Kruskal and Mosteller recommend this use of the term representative sample, bearing in mind that it needs always a specification. In their mind, the specification refers to the sampling method with which the data have been obtained.

3 LITERATURE EXAMPLES FOR ML AND AI

This section describes recent examples of the notions of a 'representative sample' in the ML and AI literature. We start with references containing more general discussions of representativeness and then move on to concrete use-case examples. We will see that the notions identified by Mosteller and Kruskal in 1979 are also pertinent in recent scientific works.

One of the recent proposals to address the lack of transparency around dataset collection and design in ML/AI is that of Datasheets for Datasets [16]. One of the questions Gebru et al proposes concerns data representativity, and says: "Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable)." Here, we note a notion of representative sample meaning coverage. The description furthermore concerns some of the historical issues as the earliest endeavors by Anders Kiær (Director of Statistics Norway during 1877-1913) to go from full census to a representative sample, namely, how do we measure the representativeness? [31] Coverage may or may not be what we go for, but if we go for it, how do we measure coverage, in particular considering joint distributions from several attributes? For example if mean values or min/max of each attribute match between sample and population, this does not imply that the distributions of each attribute match between sample and target population. This only becomes more complex if we consider the joint distributions of the attributes. Second, we should note that if we strictly go for coverage, then distributions between sample and population most likely do not match, and e.g. variance or mean estimates based on the sample will differ. On the other hand, coverage has an intuitive attraction when it comes to inclusion and equality. We will demonstrate these aspects empirically in Section 5.

Another question Gebru et al proposes to answer in a datasheet refers to the method of sampling: "If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?" Underlining the historical recommendations for a specification of sampling method when referring to a 'representative sample'. We will add, that any dataset is a sample of a larger set or population, and thus this question should always be sought answered. In fact, this question may be more important to answer than that above, as the answers above heavily depend on the answers to this question. Hopefully, answers are also well aligned with the first question in the motivation part of Gebru et al's datasheet, namely "For what purpose was the dataset created?" In fact, for some purposes, small sets of data not generally representative of the entire population in question can be good enough. Subsets of data may show that some characteristic thought to be absent or rare is in fact more frequent, or vice versa, that something thought of as universal is in fact missing to at least some degree. These subsets may be representative of only a part of the underlying population and thus form basis to dismiss one of the mentioned hypotheses, but not to draw any further inference about the entire population, see also [30] for examples. With open source datasets, we should be careful, as the purpose or the hypothesis means we have collected specific data to enlighten us, and this data may not be useful to draw inference in for other hypotheses or purposes.

In Kelly et al's 2019 opinion paper 'Key challenges for delivering clinical impact with artificial intelligence' [25], they mention representative sample as follows: "The curation of independent local test sets by each healthcare provider could be used to fairly compare the performance of the various available algorithms in a representative sample of their population." This notion of a representative sample speaks to some absence of selective forces in that it is believed each healthcare provider is best of providing its own sample, representative for their population, thus arguing for local models specific for a geographic area with specific demographics. Furthermore, distribution shifts are mentioned as a challenge for the AI models in healthcare, not only across healthcare providers, but also across time. This methodological discussion of whether a population should be seen as fixed or whether it itself is taken from an underlying stochastic process has ties all the way back to discussions from the 1903 ISI Berlin meeting (World Statistical Congress) [31].

Now we turn to a concrete use-case example from the paper 'Prediction of suicide attempts in a prospective cohort study with a nationally representative sample of the US population' by Machado et al (2021) [11]. Their notions of representativity are illustrated in the following citations: "...surveyed a representative sample of the adult population of the United States, oversampling black people, Hispanic individuals, and young adults aged 18– 24 years. ... response rate of 81% ... Weighted data were adjusted to be representative of the civilian population of ... The cumulative response rate... 70.2% ... data were weighted to reflect design characteristics of the NESARC and account for oversampling." Indeed the term is at first vague in the title, the population is then described as well as the survey sampling method, with further details in the additional data reference [41]. There seems to be some notion of a miniature in particular in terms of reweighing characteristics to match populations of interest. A certain notion of coverage and absence of selective forces can also be seen in terms of age and race, for which specific sampling strategies (oversampling) have been taken. Most importantly, details of the sampling/data collection have been specified.

A similar example concerns health insurance in France by Vimont et al (2021) with the title ’Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France’ [55]. Vimont et al reference the data in [53], and we note in Tuppin et al, that random sampling is an essential part of the argumentation for representativity: "A 1/97th random sample of ..., representative of the national population of health insurance beneficiaries, was composed in 2005...". The sample is further processed and described in the following citations from [55]: "...a representative sample in terms of age, gender and location ... is available..." and "All individuals with or without claims during ... were included... Exclusion criteria were based on health status ... Pregnancy-related care ... psychiatric disorders ... living in extra-continental France ... living abroad were also excluded." There seems to be a notion of miniature for specific attributes, and of selection criteria necessary because of the lack of certain covariates able to explain specific financial costs for subgroups of the target population. Summing up, a sample which is deemed representative in Tuppin et al is modified in Vimont et al for their analysis. It is reasonable that the data sampling is aligned with the purpose, and we will simply for completeness add that the modified data sample now represents a different population, not representative of the entire nationwide claims database.

In contrast, we encountered an example with a benchmark dataset from ICCV 2019 with a very subtle notion of representativity [54]. The notion is that the samples are real-world data, implicitly indicating that the real-world scans of objects are more representative of problems expected to occur in vision tasks than computer generated object scans. For such implicit use, we will refer back to datasheets for datasets and recommend more explicit descriptions of data sampling and its purpose [16], as well as the notion of a representative sample as an acclaim (here by implicit indication), which calls for a specification. In a similar benchmark data publication with focus on real-world images [34], we additionally see a notion of coverage: "... objectives for underwater image collection: ... a diversity of underwater scenes, different characteristics of quality degradation, and a broad range of image content should be covered."

We also came across a notion of non-representativity in the ’Understanding the Demographics of Twitter Users’ by Mislove et al (2011), where they conclude that Twitter users are not representative of the US population based on argumentation of non-matching demographic distributions for geography, gender, and race/ethnicity [37]. This notion is related to that of a miniature, and we note that a dismissal of the representativity is in essence easier than proving it holds. However, even a dismissal of a sample as representative is limited to our understanding of the population. An understanding which for example is limited as explained by Taleb’s Black Swan theory [50, 51] about human’s rationalization of rare and unpredictable events. Ruths and Pfeffer later on proposed eight steps to reduce biases and flaws in social media data [46], parts of these relate to the data collection and its documentation (similar to datasheets for datasets), and another part relates to correction for biases by population matching (miniature notion) or robustness testing across time and different samples.

On top of this, we found a new use and perhaps also notion of a representative sample, meaning a sample representative of a specific target. In online tracking, this is used to help overcome occlusions when following a target in a video [42]. This meaning is perhaps most related to that of typical exemplars, here of a specific target of interest.

4 STATISTICAL METHODOLOGY

This section gives a brief overview of common sampling strategies. It is not intended to be exhaustive, but rather give a link between the statistical methodology and the notions of a representative sample previously described. Generally, we can think of sampling either as sampling from a population or from a probability distribution. A population can for example consist of physical material like the entire harvest of maize in Kenya, or it can be the population of USA. Sampling from probability distributions are used in stochastic simulations[1] [3] with applications to for example chemical kinetics[19] and cellular systems [2]. Both of these approaches have numerous different sampling methods with varying properties [17]. A schematic overview is given in Figure 1, where we have added sampling from time dependent observations as well as randomization as special cases when creating data samples. Note, that sampling strategies can be combined. For example, for a recruitment in a randomized control trial (RCT), a sampling from a population can be performed to recruit participants, and then subsequently a randomization of the sampled observations is performed to assign observations to either control or treatment. The latter corresponds to a random permutation of the assignment order.

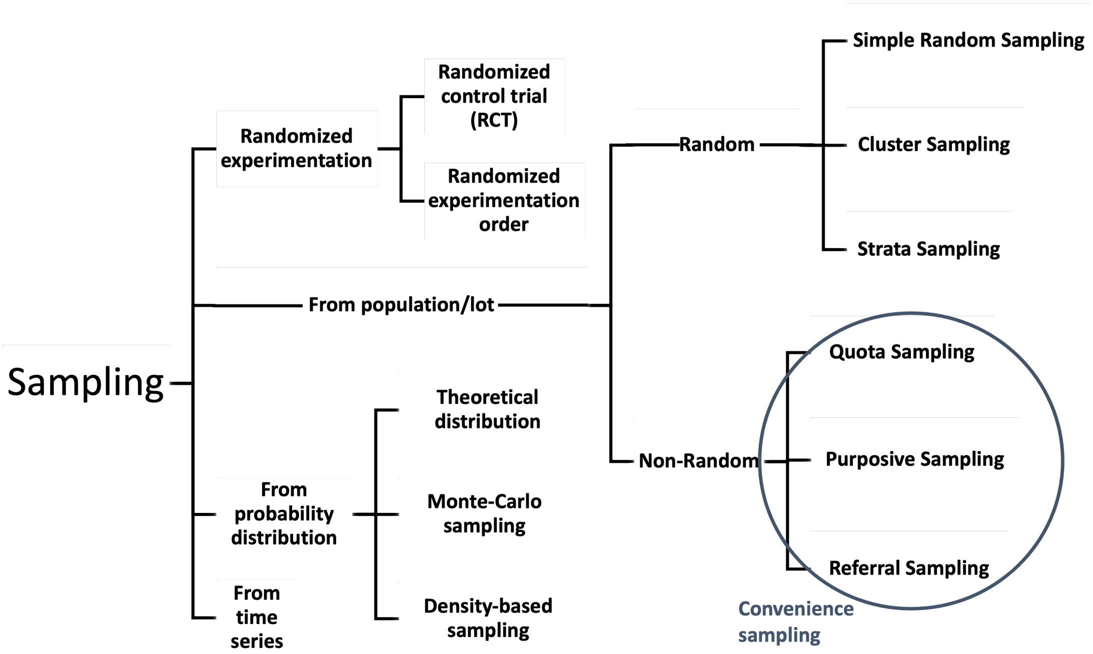


Fig. 1. Schematic overview of the sampling approaches described in this section.

Sampling from a population/set/lot includes theories from experimental disciplines (physics, chemistry etc), social sciences and survey designs, and observational studies in general [18, 39, 43, 45]. Survey sampling can be divided into random sampling and nonrandom sampling techniques. In Section 2, under (absence) of selective forces, we mentioned *non-random sampling* methods like quota sampling (sampling, by convenience, to fulfill quotas within pre-identified groups in a population), purposive sampling (seeking out samples/individuals meeting specific criteria to participate in a study), and referral sampling (exists in different variations, where the common factor is that individuals refer other

[1] A related method to stochastic simulation, is that of deterministic simulation, but here there is no random (stochastic) input component, but rather the same (deterministic) boundary settings in the model will always give the same results. Deterministic simulation is often used in fluid dynamics and finite elements modeling.

individuals to the study). As described earlier, the non-random sampling strategies may challenge the generalization, but for specific use-cases may be exactly what is needed to uncover e.g. hidden populations [22].

Random sampling methods include simple random sampling (a random selection of observations in the target population), stratified sampling (simple random sampling within mutually exclusive groups of the target population/strata), and cluster sampling (random sampling of clusters/strata in the population and inclusion of all samples for the selected clusters) [22]. Ghojogh et al show that strata sampling always has lower variance than that of simple random sampling, in particular when strata have very different characteristics [17]. The theory of sampling developed to investigate physical materials is mainly concerned with random sampling, but the physical material is pre-treated to obtain as homogeneous a lot as possible, corresponding to eliminating varying strata [20, 43]. If it is not possible to homogenize the material, it is recommended to perform composite sampling, meaning sampling by systematic or stratified sample selection [43]. Another option is to sample enough random samples to obtain a convergence in the measure of interest [5].

For the survey sampling, we should pay attention to the following sampling biases: Coverage bias (some groups are systematically excluded), non-response bias (non-response that is not happening at random), sample selection bias (exclusion based on conscious or subconscious choices made by the surveyor), and sample attrition bias (caused by groups systematically unavailable for follow up studies). Furthermore, it is paramount to pay attention to errors which are not related to the sampling method, but to any other source of error in the sampling procedure. These are sometimes referred to as non-sampling errors, and includes error sources like interviewer effects for non-response as well as response observations in the surveys, and if care is not taken can be substantially larger than the sampling errors. Strategies to account for non-response exist like those for matching and reweighing samples to reflect the target population [18].

Sampling with a notion of coverage in mind often means combining non-random and random sampling methods, whereas sampling with a miniature in mind often means using random probability sampling, for example strata sampling.

Randomized controlled trials and experimentation are truly random samples, whereas *observational studies* need to be carefully designed to tackle their inherent haphazardness [45]. In the latter case, matching is performed to make treatment and control groups comparable, but unlike for experimentation, there is no basis for assuming that this extends to unmeasured factors [39, 45]. Experimental studies are often used to make causal inferences, a basis which dates back to R.A. Fisher (1935) [14]. However, causal relations can also be established through observational studies, like for example the link between smoking and lung cancer [8].

Sampling from time series is a special case, which needs mentioning as it comes with the concern of auto-correlation, which can challenge assumptions of independence between observations. There exist vast amounts of literature specifically on data with time dependency; one of the fields is that of process surveillance [58]. Here, sampling is often performed at a certain frequency, and the methods are for example useful for quality control in production settings [38]. It is our believe that these methods deserve more attention as means for continuous evaluation and monitoring of AI systems as well. We will get back to this in our discussion, in Section 7.

Sampling from distributions can either be performed from simpler parameterized distributions using the inverse cumulative distribution function, see e.g. [47], or if the distribution of interest is more complex we can use Markov Chain Monte Carlo (MCMC) or related methods like Metropolis-Hastings or Gibbs sampling [35]. Sampling from distributions, and not least joint distributions, gives the possibility of matching distributions between sample and population rather than matching simpler characteristics, like e.g. averages. In high dimensions, these methods do suffer computationally, however. As a non-parametric alternative it is possible to sample from densities. Density-based sampling approaches are for example useful under the coverage notion of representative sampling, where density estimates can be used to asses population imbalances and use this information for sampling to cover the heterogeneity of the population in the sample [26].

New sampling proposals for specific use cases like e.g. balancing imbalanced classes [49, 57], online sampling [7, 42] or reinforcement learning [13, 15] emerge constantly. We do not intend to cover all of these here, but simply mention that there is a large and growing body of literature to assist in obtaining the best possible sampling strategy for a given purpose.

5 DEMONSTRATIONS USING DATA

This section demonstrates contrasting perspectives on representativity by empirically comparing performance and fairness metrics for samples created with the notions of coverage and miniature, respectively. The samples are created from a US census data collection [10] through density based sampling to achieve coverage or through simple random sampling to obtain a miniature, respectively. Subsequently, we examine and compare the sampling strategies with respect to their robustness to distribution shifts.

5.1 Data

A dataset which is popular in the fairness community is the UCI Adult dataset from the 1994 Current Population Survey organized by the US Census Bureau [27]. This data has been used in hundreds of research papers, but its external validity has been questioned, and a collection of new datasets from US Census Bureau data have been proposed [10]. More specifically, these datasets are extracted from the American Community Survey Public Use Microdata Sample (ACS PUMS). They contain data on attributes like age, income, education, sex, ancestry and employment. The responses to the survey are controlled by privacy rules seeking to prevent re-identification of responders. Detailed documentation on the records can be found on the US Census Bureau websites. One of the proposed datasets is a replacement for the original UCI Adult dataset containing an income prediction task for a feature subset of the 2018 ACS PUMS data spanning all US states in addition to Puerto Rico. To generate the dataset the ACS PUMS data are filtered to only include individuals over the age of 16 with at least one working hour per week and an income of at least 100 USD in the past year. This leaves a total of 1,664,500 individuals in the data. Like the original UCI Adult dataset, this new dataset has a predefined income threshold (50,000 USD) used to binarize the targets into a classification task. The income threshold has been criticized to limit the external validity of the dataset [10]. We create a modified version of the income dataset and omit the income threshold to form a regression task with the continuous income as target. An overview of the dataset can be seen in Table 1. We binarize the nominal features COW (class of worker), MAR (marital status), POBP (place of birth) and RELP (relationship). COW is binarized into government / non-goverment worker, MAR is binarized into married / not married. POBP is binarized into US-born / non-US-born and RELP is binarized into reference person / non-reference person. Finally we transform the income target using the natural logarithm to obtain homoscedasticity for the residuals in our regression model.

5.2 Methodology

We compare linear regression models fitted to the log transformed income using all features in Table 1 for California (n=195,665). We reserve 20% of the California data for testing, and use the remaining 80% as training data, which we denote the full census training data. We compare models trained using the full census training data to models

Table 1. Overview of the features in our modified US Census income data. Features COW, MAR, POBP, RELP are modified from the original ACS PUMS data by binarizing into respectively goverment / non-goverment worker (COW), married / not married (MAR), US-born / non-US-born (POBP) and reference person / non-reference person (RELP). See ACS PUMS dictionary documentation for full feature descriptions including original category codes.

Feature Type	Feature Name	Description	Data Type	Categories	Min/Max
Input	AGEP	Age	Continuous	-	17 - 96
Input	COW	Class of worker	Nominal	2	-
Input	SCHL	Educational attainment	Ordinal	24	-
Input	MAR	Marital status	Nominal	2	-
Input	POBP	Place of birth	Nominal	2	-
Input	RELP	Relationship	Nominal	2	-
Input	WKHP	Hours worked per week	Continuous	-	1 - 99
Input	SEX	Sex	Nominal	2	-
Input	RAC1P	Race	Nominal	9	-
Target	PINCP	Total income	Continuous	-	104 - 1,423,000

trained on samples of the training data following either the miniature or coverage notion of representativity. We compare performances on a range of metrics including overall performance using the mean squared errors (MSE) on in-distribution (the California test data) and out-of-distribution data (the remaining US states) as well as performance on fairness criteria related to demographic parity and equalized odds. For classification problems demographic parity seeks equality of positive rates for subsets of a protected attribute while equalized odds seeks equality of true positive rates (TPR) and false positive rates (FPR) for subsets of protected attributes. Equalized odds can be relaxed to only require non-discrimination within the advantaged outcome, which is known as equal opportunity [21]. These fairness criteria are well defined for classification problems but have only been studied sparsely for regression problems [1]. Formally a predictor satisfies demographic parity if the predictions are independent of the protected attribute [1]. We form a demographic parity criterion by measuring the extent to which the means of the prediction distributions for individuals within subsets of a protected attribute differ. Under perfect demographic parity this criterion will be zero. We formulate the parity criterion as:

$$P = \left| \frac{1}{|A_0|} \sum_{i \in A_0} \hat{y}_i - \frac{1}{|A_1|} \sum_{i \in A_1} \hat{y}_i \right|, \tag{1}$$

where P is the parity criterion, $\hat{y}_i$ is the model prediction for the i^{th} observation, A_0 is a population subset of individuals where the protected attribute A is false, while A_1 is a subset where the protected attribute is true, and $|A|$ is the number of observations in subset A. In essence, this criterion measures the degree to which predictions are equal across subsets of a protected attribute. As an alternative, the equalized odds definition of fairness requires non-discrimination in errors across subsets of a protected attribute. We form the following equality criterion under this notion by measuring the extent to which the MSE differs between subsets of a protected attribute:

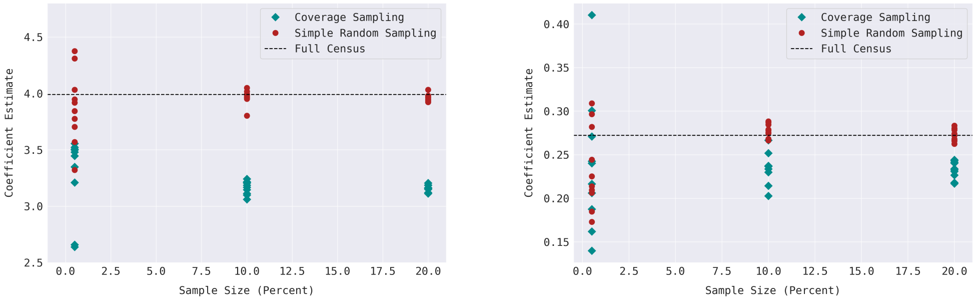
$$E = \left| \frac{1}{|A_0|} \sum_{i \in A_0} (y_i - \hat{y}_i)^2 - \frac{1}{|A_1|} \sum_{i \in A_1} (y_i - \hat{y}_i)^2 \right|, \tag{2}$$

where E is the equality criterion and y_i is the target for the i^{th} observation.

We generate the coverage sample using a density based weighted sampling strategy proposed in [26], where they measure density around observations as the mean distance to the 100 nearest neighbors. These density estimates are then scaled to sum to one, and subsequently used as sampling probabilities in a weighted random sampling scheme with replacement. This approach causes observations in low-density regions to be sampled with high probability and conversely observations from high-density regions to be sampled with low probability. In doing so, the coverage sampling approach seeks to equally cover the input space regardless of the demographic proportions in the population. To generate the miniature sample we perform simple random sampling (SRS), which samples without replacement and gives each observation the same probability of being sampled. We assume that the low number of features in the income dataset coupled with a sufficiently large sample size will allow the SRS to accurately mimic the population distribution and constitute a miniature sample. We empirically evaluate this by investigating regression coefficient estimates for the two sampling approaches.

5.3 Results

We plot the estimates of two regression coefficients from models trained on an increasing number of observations sampled with the two methods (density based coverage and simple random sampling), as well as the estimate based on the full census training data in Figure 2. As the sample size increases, the coefficient estimates for the SRS converge to those from the full census training data, while the coefficient estimates for the coverage sample generally do not. Thus we argue that for a sufficient sample size, the SRS acts like a miniature sample mimicking the population distribution in the US Census income data (at least with respect to income and the linear effects of the features included here).



(a) Hours worked per week feature.

(b) Marital status feature.

Fig. 2. Parameter estimates of regression coefficients from 10 iterations of coverage and SRS from the full census training data. The full census training data regression coefficients are plotted as dotted lines. The SRS parameter estimates converge to the full census regression coefficients for a sufficiently large sample size, while the coverage parameter estimates do not. For specific features like sex and age, the parameter estimates for both sampling approaches converge to the full census regression coefficients (see Appendix A).

MSE on the test data can be seen in Table 2. It is evident that the model trained on the full census training data has the lowest MSE and that this performance is followed by the miniature sample, while the coverage sample model has the highest MSE. Table 2 also illustrates how the samples score on the fairness criteria for parity and equality between white and non-white individuals from the protected attribute RAC1P. None of the models demonstrate true parity or equality, but on average the coverage sample has the best performance on the parity and equality criteria.

Table 2. Performance metrics from 10 iterations of training and testing models on the California data. For each iteration 80% of the California data is randomly chosen for training a full census model and the remaining 20% is used for testing. For each iteration a miniature and coverage sample is drawn from the full census training data and tested on the test data. The miniature and coverage sample size is 20% of the full census training data. The mean MSE and standard deviations (SD) in addition to parity and equality scores for white / non-white individuals is shown. The parity and equality scores are measured according to Equations 1 and 2.

Training Data	Mean MSE	Mean Parity	Mean Equality	MSE SD	Parity SD	Equality SD
Full Census	**0.8032**	0.2056	0.0445	0.0041	0.0064	0.0157
Miniature Sample	0.8034	0.2055	0.0450	0.0042	0.0094	0.0155
Coverage Sample	0.8355	**0.1899**	**0.0371**	0.0054	0.0151	0.0132

5.4 Out-of-distribution results

Model robustness towards distribution shifts and consequently out-of-distribution performance has experienced increased focus from scholars. Kaushal et al (2020) [24] have examined the geographical distribution of US cohorts used to train machine learning models, and uncovered a systemic bias in the patient cohorts used to train models for clinical applications. They find that 71% of the analyzed studies used cohorts from at least 1 of 3 states, namely California, Massachusetts and New York, while 34 states did not contribute to any cohorts. California cohorts appeared in 39% of all analyzed studies. Models trained on cohorts from specific geographical locations, which are then applied to draw inference on data from different locations, are prone to suffer in performance, and fairness, if geographical distributional shits are present. This is indeed relevant for the US Census data, which has been shown to exhibit significant geographical variation [10]. To investigate the role of data representativity for drawing inference under distribution shifts, we compare mean squared errors on the training state (California) as well as on the remaining 49 states and Puerto Rico for the different sampling strategies. We choose sample sizes for both the coverage and miniature samples of 20% (n=31,306) of the full California (training) data. We perform each sampling approach 10 times and compare linear regression models trained on either the full California (training) data or on coverage and miniature samples drawn from the California training data.

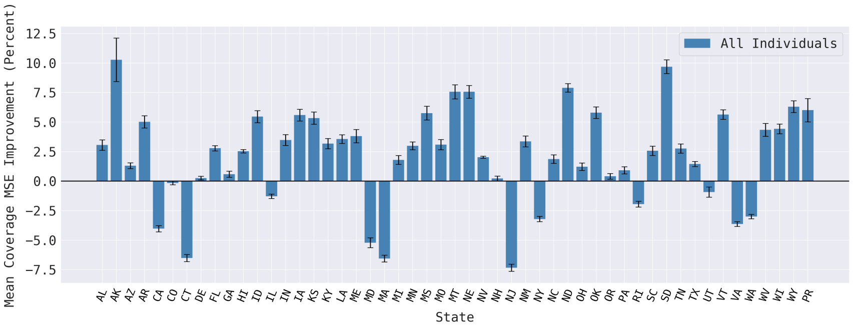


Fig. 3. Mean performance improvement/deterioration when using the coverage sample compared to using the full census California training data over 10 iterations of sampling and training linear regression models. The coverage sample is 20% of the full census training data. Error bars indicate the standard deviation. Positive values express that the mean MSE of the models trained on the coverage samples is improved over the mean MSE of the full census training data models. Negative values indicate that the sampling deteriorates the performance compared to using the full census training data. The mean performance improvement across all states is 2.1%.

Figure 3 compares models trained on coverage samples to models trained on the full census California training data. The coverage sample deteriorates the performance significantly when testing the models on the in-distribution data (CA) as well as a number of states, most notably CT (Connecticut), MD (Maryland), MA (Massachusetts), NJ (New Jersey), NY (New York), VA (Virginia) and WA (Washington), which are similar to CA. However, the coverage sample improves the average MSE across all states by 2.1%. By equally representing the input space of the training data, the coverage sample experiences higher robustness to the interstate geographical distributional shifts in the US Census data and achieves a better predictive performance on states that are dissimilar to the training state. This is most notable on the states AK (Alaska), AR (Arkansas), ID (Idaho), IA (Iowa), KS (Kansas), MS (Mississippi), MT (Montana), NE (Nebraska), ND (North Dakota), OK (Oklahoma), SD (South Dakota), VT (Vermont), WY (Wyoming) and PR (Puerto Rico).

Figure 4 compares the coverage sample performance to the miniature sample performance on individuals either born in the US or not. US-born individuals are the majority in all states of the US Census income data, while non-US-born individuals are the minority (respectively 84.6% US-born and 15.4% non-US-born individuals across the entire income dataset). The coverage sample improves the out-of-distribution performance over the miniature sample on average across all states by 2.2%, but the improvement is larger on the underrepresented non-US-born individuals (3.9%) than the overrepresented US-born individuals (1.8%).

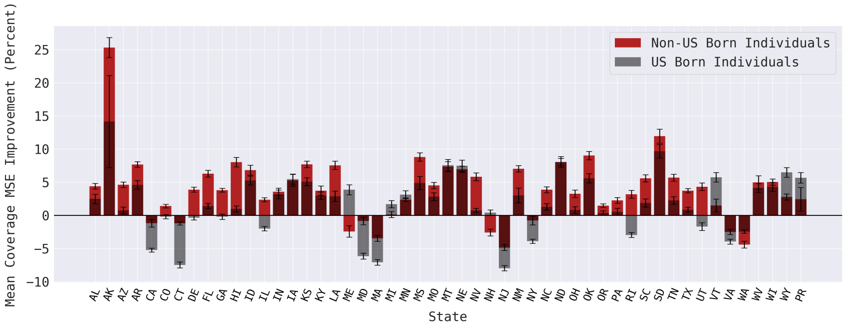


Fig. 4. Mean performance improvement/deterioration for US born and non-US born individuals over 10 iterations of training linear regression models from coverage and miniature samples of the California training data. The sizes of both the miniature and coverage samples are 20 % of the full census training data. Error bars indicate the standard deviation. Positive values express that the mean MSE of the models trained on the coverage samples is improved over the mean MSE of the models trained on the miniature samples. Negative values indicate that the coverage sampling deteriorates the performance compared to using the miniature sampling.

5.5 Summing up experiments on data

While the coverage sample has merits such as robustness to distribution shifts and improved performance for underrepresented parts of the input space, the coverage notion fails to accurately represent the distribution of the underlying population and consequently incurs a loss in predictive power on the majority of said population, measured by the MSE. On the contrary the miniature sample accurately represents the demographic distribution of the underlying population and as such is particularly appropriate for historical or in-distribution inference on the majority. This is evident for model performances on in-distribution data from California, where the full census training data and miniature sample achieve better predictive performance than the coverage sample, and similar regression coefficient estimates, and thus a similar interpretation of relations in the dataset.

6 FRAMEWORK FOR DATA REPRESENTATIVITY

This section presents our proposed framework of questions for assessing data representativity when creating and documenting data. The framework naturally fits into both datasheets for datasets [16] as well as shorter, more general data descriptions, and our aim here is to make it as concise and manageable as possible. With this in mind, and based on our literature study and empirical investigations, we propose answering and adhering to the following questions and guidelines:

6.1 Purpose:

What is the purpose of collecting/creating the data, and what/who is the target population? In addition, when building AI systems; what is the intended aim of the AI system along side its intended use?

6.2 Sampling methodology:

What is the sampling method and procedure used to create the data? The methodology should be specified to a degree that makes reproducibility possible. If a code base is used to create the data, we recommend making it open source.

6.3 Evaluation:

Are the collected data representative of the target population or 'good enough' for the aim? We recommend making this evaluation in accordance with the purpose, and not as a general statement of representativity. In addition, known limitations of the representativity, in terms of coverage as well as distributional match to target population, are always desirable to document for datasets to assess possible limitations, and not least because open source datasets may be used for purposes not originally anticipated.

7 DISCUSSION

We found that the notions of what constitutes a 'representative sample' from the 1979 reviews by Mosteller and Kruskal are still pertinent. When building machine learning models and AI systems, particularly two contrasting views of representativity are of relevance: The notion of coverage vs that of a miniature. We find that the two are useful for different purposes. Coverage is useful for robustness towards distribution shifts as well as better equality and demographic parity. The miniature is useful to mimic the target population and to achieve the minimum average errors on same. However, we should keep in mind, that average errors mean that predictions are best for the majority, and not necessarily equal for population subgroups.

The notion of a 'representative sample' as an assertive acclaim without specification was mainly used in AI related literature as an implicit acclaim, without explicit mentioning of representativity, but with an indication of an inference link (generalization from data) matching that of representativity. We call for attention on such implicit use, and recommend avoiding it, thus always specifying the sampling methodology as well as purpose and target population of collected data along with an evaluation of representativity and limits of same for the given sample.

Through our investigations we found that we cannot talk about general representativeness of a sample, but need to consider data collection and representativeness in coherence with our purpose (and data analysis) whether this is a research hypothesis our an aim for our AI system.

As we reach limitations from our understanding of the target distributions and/or from a large number of attributes (and their interactions), it is practically impossible to make guarantees of representativeness. As a consequence, evaluations based on several datasets as well as ’in use’ data (for deployed ML models or AI systems) are encouraged. Furthermore, accounting for all possible distribution shifts that may happen in the future (where our AI system will be in production), is also practically impossible. As an alternative, or rather addition, we suggest to perform continuous monitoring of AI systems and their performance while they are in production. An AI system may also at first be deployed in shadow mode if risks are too high to use predictions without further (live) testing.

8 ACKNOWLEDGEMENTS

The authors would like to acknowledge colleague Murat Kulahci for insightful feedback on the manuscript. Additionally, author Rune D. Kjærsgaard is funded by a university alliance scholarship between UiB (University of Bergen) and DTU (Technical University of Denmark).

REFERENCES

[1] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.

[2] M. Ander, P. Beltrao, B. Di Ventura, J. Ferkinghoff-Borg, M. Foglierini, C. Lemerle, I. Tomás-Oliveira, and L. Serrano. 2004. Stochastic Simulation of Chemical Kinetics. *Systems Biology* 1 (2004), 129–138. Issue 1.

[3] Søren Asmussen and Peter W. Glynn. 2007. *Stochastic Simulation - Algorithms and Analysis*. Springer.

[4] Maciej Bereswicz. 2017. A Two-Step Procedure to Measure Representativeness of Internet Data Sources. *International Statistical Review* 85 (2017), 473–493. Issue 3.

[5] Megan L. Blatchford, Chris M. Mannaerts, and Yijian Zeng. 2021. Determining representative sample size for validation of continuous, large continental remote sensing data. *International Journal of Applied Earth Observations and Geoinformation* 94 (2021), 102235.

[6] Danah Boyd and Kate Crawford. 2011. Six Provocations for Big Data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society* (September 2011).

[7] Romain Camilleri, Zhihan Xiong, Maryam Fazel, Lalit Jain, and Kevin Jamieson. 2021. Selective Sampling for Online Best-arm Identification. *Part of: Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS 2021)* (2021).

[8] Jerome Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, , and Ernst L. Wynder. 2009. Smoking and lung cancer: recent evidence and a discussion of some questions. *International Journal of Epidemiology* 38 (2009), 1175–1191.

[9] Adele Cutler and Leo Breiman. 1994. Archetypal analysis. *Technometrics* 36 (1994), 338–347. Issue 4.

[10] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems* 34 (2021).

[11] Cristiane dos Santos Machado, Pedro L. Ballester, Bo Cao, Benson Mwangi, Marco Antonio Caldieraro, Flávio Kapczinski, and Ives Cavalcante Passos. 2021. Prediction of suicide attempts in a prospective cohort study with a nationally representative sample of the US population. *Psychological Medicine* (2021), 1–12.

[12] Ravit Dotan and Smitha Milli. 2020. Value-laden Disciplinary Shifts in Machine Learning. *FAT* ’20, January 27-30, 2020, Barcelona, Spain* (2020).

[13] Aleksandra Faust, Kenneth Oslund, Oscar Ramirez, Anthony Francis, Lydia Tapia, Marek Fiser, and James Davidson. 2018. PRM-RL: Long-range Robotic Navigation Tasks by Combining Reinforcement Learning and Sampling-Based Planning. *In Proceedings of: 2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018).

[14] Ronald A. Fisher. 1935. *The Design of Experiments*. Oliver and Boyd.

[15] Scott Fujimoto, David Meger, and Doina Precup. 2021. A Deep Reinforcement Learning Approach to Marginalized Importance Sampling with the Successor Representation. *In Proceedings of: The 38th International Conference on Machine Learning, PMLR* 139 (2021).

[16] Timnit Gebru, Jamie Morgenstern, Briana Vechhione, Jennifer Wrotmen Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. 2021. Datasheets for Datasets. *arXiv:1803.09010v8* (2021).

[17] Benyamin Ghojogh, Hadi Nekoei, Aydin Ghojogh, Fakhri Karray, and Mark Crowley. 2020. Sampling algorithms, from Survey Sampling to Monte Carlo Methods: Tutorial and Literature Review. *arXiv:2011.00901v1* (2020).

[18] Lior Gideon. 2012. *Handbook of Survey Methodology for the Social Sciences*. Springer.

[19] Daniel T. Gillespie. 2007. Stochastic Simulation of Chemical Kinetics. *Annual Review of Physical Chemistry* 58 (2007), 35–55.

[20] Pierre Gy. 1998. *Sampling for Analytical Purposes*. Wiley.

[21] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[22] Mary Hibberts, R. Burke Johnson, and Kenneth Hudson. 2012. *Common Survey Sampling Techniques*. Springer. 53–74 pages.

[23] Jonathan Yinhao Huang. 2021. Representativeness Is Not Representative - Addressing Major Inferential Threats in the UK Biobank and Other Big Data Repositories. *Epidemiology* 32 (2021), 189–193. Issue 2.

[24] Amit Kaushal, Russ Altman, and Curt Langlotz. 2020. Geographic distribution of US cohorts used to train deep learning algorithms. *Jama* 324, 12 (2020), 1212–1213.

[25] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* 17 (2019). Issue 195.

[26] Rune D Kjærsgaard, Manja G Grønberg, and Line KH Clemmensen. 2021. Sampling To Improve Predictions For Underrepresented Observations In Imbalanced Data. *arXiv preprint arXiv:2111.09065* (2021).

[27] Ronny Kohavi and Barry Becker. 1996. Adult data set. *UCI machine learning repository* 5 (1996), 2093.

[28] William Kruskal and Frederick Mosteller. 1979. Representative sampling, I: Non-scientific Literature. *International Statistical Review* 47 (1979), 13–24.

[29] William Kruskal and Frederick Mosteller. 1979. Representative sampling, II: Scientific Literature, Excluding Statistics. *International Statistical Review* 47 (1979), 111–127.

[30] William Kruskal and Frederick Mosteller. 1979. Representative sampling, III: the Current Statistical Literature. *International Statistical Review* 47 (1979), 245–265.

[31] William Kruskal and Frederick Mosteller. 1980. Representative sampling, IV: the History of the Concept in Statics, 1895-1939. *International Statistical Review* 48 (1980), 169–195.

[32] Murat Kulahci, Flavia Dalia Frumosu, Abdul Rauf Khan, Georg Ørnskov Rønsch, and Max Peter Spooner. 2020. Experiences with big data: Accounts from a data scientist's perspective. *Quality Engineering* 32 (2020), 529–542. Issue 4.

[33] Herbert K. H. Lee, Matthew Taddy, and Genetha A. Gray. 2010. Selection of a Representative Sample. *Journal of Classification* 27 (2010), 41–53.

[34] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. 2020. An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Transactions on Image Processing* 29 (2020), 4376–4389. Issue 1.

[35] David J.C. MacKay. 2005. *Informaiton Theory, Inference, and Learning Algorithms* (7 ed.). Cambridge University Press.

[36] Ninahreh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54 (2021). Issue 6.

[37] A. Mislove, S. S. Lehmann, Y.-Y. Ahn, J. p. Onnela, and J. Rosenquist. 2011. Understanding the Demographics of Twitter Users. *Proceedings of: Fifth International AAAI Conference on Weblogs and Social Media* 5 (2011), 554–557. Issue 1.

[38] Douglas C. Montgomery. 2013. *Introduction to statistical quality control* (7th ed.). Wiley.

[39] Douglas C. Montgomery. 2019. *Design and Analysis of Experiments* (10th ed.). Wiley.

[40] Alexandra Olteanu, CarlosCastillo, FernandoDiaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Front. Big Data* 2 (2019). Issue 13.

[41] National Institute on Alcohol Abuse and Alcoholism. 2006. National epidemiologic survey on alcohol and related conditions (NESARC). *Alcohol Alert* 70 (2006), 1–6. Issue 1.

[42] Weihua Ou, Di Yuuan, and Yongfeng Cao. 2018. Object tracking based on online representative sample selection via noon-negative least square. *Multimed Tools Appl* 77 (2018), 10569–10587.

[43] Lars Petersen, Pentti Minkkinen, and Kim H. Esbensen. 2005. Representative sampling for reliability data analysis: Theory of Sampling. *Chemometrics and Intelligent Laboratory Systems* 77 (2005), 261–277.

[44] Harry V. Roberts. 1971. Committee Selection by Statistical Sampling. *The American Statistician* 25 (Feb 1971), 18–20. Issue 1.

[45] Paul R. Rosenbaum. 2010. *Design of Observational Studies*. Springer.

[46] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346 (2014), 1063–1064. Issue 6213.

[47] William T. Shaw. 2006. Sampling Student's T distribution-use of the inverse cumulative distribution function. *Journal of Computational Finance* 9 (2006), 37. Issue 4.

[48] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv:1901.10002v1* (2019).

[49] Seba Susan and Amitesh Kumar. 2021. The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. *Engineering Reports* 3 (2021), e12298.

[50] Nassim Nicholas Taleb. 2007. *The Black Swan: The Impact of the Highly Improbable*. Random House.

[51] Nassim Nicholas Taleb. 2020. *Statistical Consequences of Fat Tails*. STEM Academic Press.

[52] Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. *CVPR* (2011), 1521–1528.

[53] P Tuppin, J Rudant, P Constantinou, C Gastaldi-Ménager, A Rachas, L de Roquefeuil, G Maura, H Caillol, A Tajahmady, J Coste, C Gissot, A Weill, and A Fagot-Campagna. 2017. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. 65 (2017). Issue 4.

[54] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 2019. Revisiting Point Cloud Classification: A New Benchmark Dataset and Classification Model on Real-World Data. *Proceedings of: IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 1588–1597.

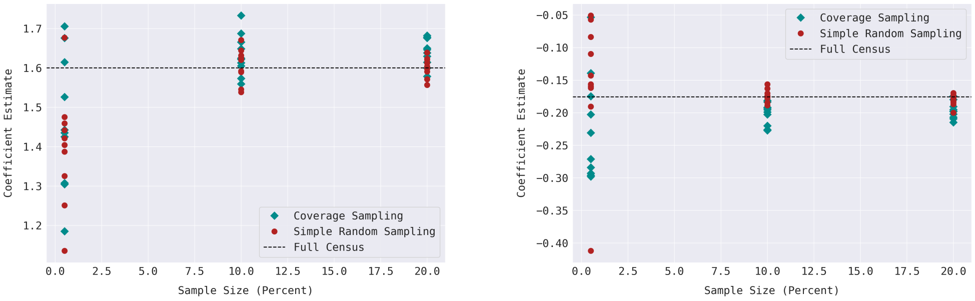
[55] Alexandre Vimont, Henri Leleu1, and Isabelle Durand-Zaleski. 2021. Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France. *The European Journal of Health Economics* (2021).

[56] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. *FAT* '20, january 27-30* (2020).

[57] Ming Zheng, Tong Li, Xiaoyao Yu, Chuanming Chen, Ding Zhoou, Changlong Lv, and Weiyi Yang. 2021. UFFDFR: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced classification. *Information Sciences* 576 (2021), 658–680.

[58] Inez M. Zwetsloot and William H. Woodall. 2021. A Review of Some Sampling and Aggregation Strategies for Basic Statistical Process Monitoring. *Journal of Quality Technology* 53 (2021), 1–16. Issue 1.

A ADDITIONAL PARAMETER ESTIMATES



(a) Age feature.

(b) Sex feature.

Fig. 5. Parameter estimates of regression coefficients from 10 iterations of coverage and SRS from the full census California training data. The full census training data regression coefficients are plotted as a dotted line. The SRS parameter estimates converge to the full census regression coefficients for a sufficiently large sample size. The coverage parameter estimates generally do not converge to the full census training data coefficients, but for specific features like age and sex shown here, they do.