1

Why Machines Matter for Survey and Social Science Researchers: Exploring Applications of Machine Learning Methods for Design, Data Collection, and Analysis

Trent D. Buskirk[1] *and Antje Kirchner*[2]

[1] *Applied Statistics and Operations Research Department, College of Business, Bowling Green State University, Bowling Green, OH, USA*
[2] *RTI International, Research Triangle Park, NC, USA*

1.1 Introduction

The earliest hard drives on personal computers had the capacity to store roughly 5 MB – now typical personal computers can store thousands of times more (around 500 GB). Data and computing are not limited to supercomputing centers, mainframe or personal computers, but have become more mobile and virtual. In fact, the very definition of "computer" is evolving to encompass more and more aspects of our lives from transportation (with computers in our cars and cars that drive themselves) to everyday living with televisions, refrigerators, doorbells, thermostats, and many other devices that are Internet-enabled and "smart." The rise of the Internet of things, smart devices, and personal computing power relegated to mobile environments and devices, like our smartphone, has certainly created an unprecedented opportunity for survey and social researchers and government officials to track, measure, and better understand public opinion, social phenomena, and the world around us.

A review of the current social science and survey research literature reveals that these fields are indeed at a crossroads transitioning from methods of active observation and data collection to a new landscape where researchers are exploring, considering, and using some of these new data sources for measuring public opinion and social phenomena. Connelly et al. (2016) comment that "whilst there may be a 'Big Data revolution' underway, it is not the size or quantity of these data that is revolutionary. The revolution centers on the increased availability of new types of data which have not previously been available for social research." In our

Big Data Meets Survey Science: A Collection of Innovative Methods, First Edition. Edited by Craig A. Hill, Paul P. Biemer, Trent D. Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg.

view, advances in this new landscape are taking place in seven major dimensions including

(1) reimagining traditional survey research by leveraging new machine learning methods (MLMs) that improve efficiencies of traditional survey data collection, processing, and analysis;
(2) augmenting traditional survey data with nonsurvey data (administrative, social media, or other Big Data sources) to improve estimates of public opinion and official statistics;
(3) enhancing official statistics or estimates of public opinion derived from Big Data or other nonsurvey data;
(4) comparing estimates of public opinion and official statistics derived from survey data sources to those generated from Big Data or other nonsurvey data exclusively;
(5) exploring new methods for enhancing survey and nonsurvey data collection and gathering, processing, and analysis;
(6) adapting and modifying current methods for use with new data sources and developing new techniques suitable for design and model-based inference with these data sources;
(7) contributing survey data, methods, and techniques to the Big Data ecosystem.

Weaved within this new landscape is the perspective of assimilating these new data sources into the process as substitutes, augments, or auxiliary to existing survey data. Computational social science continues to evolve as a social sciences subfield that uses computational methods, models, and advanced information technology to understand social phenomena. This subfield is particularly suited to take advantage of alternative data sources including social media and other sources of Big Data like digital footprint data generated as part of social online activities. Survey researchers are exploring the potential of these alternate data sources, especially social media data. Most recently, Burke et al. (2018) explored how social media data create opportunities for not only sampling and recruiting specific populations but also for understanding a growing proportion of the population who are active on social media sites by mining the data on such sites.

No matter the Big Data source, we need data science methods and approaches that are well suited to deal with these types of data. Although some have made the case that administrative data be considered as Big Data (Connelly et al. 2016), the general consensus in both the data science and survey research communities is that they are not. However, with the increased collection of paradata and the increased use of sensors and other similar peripherals used in data collection, one could argue that survey data are getting bigger – not in the number of cases, but in the number of variables that are available per case. Put another way, some

survey datasets are getting bigger because they are "wider" rather than "longer." So we could also make the case that surveys themselves are creating bigger data and could benefit from such applying these types of methods.

This chapter explores how techniques from data science, known collectively as MLMs, can be leveraged in each phase of the survey research process. Success in the new landscape will require adapting some of the analytic and data collection approaches traditionally used by survey and social scientists to handle this more data rich reality. In his recent MIT Press' Essential Knowledge book entitled *Machine Learning*, computer engineering professor Ethem Alpaydin notes, "Machine learning will help us make sense of an increasingly complex world. Already we are exposed to more data than what our sensors can cope with or our brains can process." Although the use of MLMs seems fairly new in survey research, machines and technology have long been part of the survey and social sciences' DNA. At the 1957 AAPOR conference Frederick Stephan pointed out that "Computers will tax the ingenuity, judgment and skill of technically proficient people to (a) put the job on the machine and (b) put the results in form for comprehension of human beings; and determine the courses of action we might take based on what the machines have told us" in the context of data collection. This adage still holds today, but we are moving from machines to MLMs.

This chapter provides a brief overview of MLMs and a deeper exploration of how these data science techniques are applied to the social sciences from the perspective of the survey research process including sample design and constructing sampling frames; questionnaire design and evaluation; survey recruitment and data collection; survey data coding and processing; sample weighting and survey adjustment; and data analysis and estimation. We present a wide array of applications and examples from current studies that illustrate how these methods can be used to augment, support, reimagine, improve, and in some places replace the current methodologies. Machine learning is likely not to replace all human aspects of the survey research process, but these methods can offer new ways to approach traditional problems and can provide more efficiency, reduced errors, improved measurement, and more cost-effective processing. We also discuss how errors in the machine learning process can impact errors we traditionally manage within the survey research process.

1.2 Overview of Machine Learning Methods and Their Evaluation

MLMs are generally flexible, nonparametric methods for making predictions or classifications from data. These methods are typically described by the algorithm that details how the predictions are made using the raw data and can allow for a

larger number of predictors, referred to as high-dimensional data. These methods can often automatically detect nonlinearities in the relationships between independent and dependent variables and can identify interactions automatically. Generally, MLMs can be divided into two broad categories: supervised and unsupervised machine learning techniques. The goal of supervised learning is to optimally predict a dependent variable (also referred to as output, target, class, or label), based on a range of independent variables (also referred to as inputs, features, or attributes). When a supervised method is applied to predict continuous outcomes, it is generally referred to as a regression problem and a classification problem when used to predict levels of a categorical variable. Ordinary least squares regression is a classic example of supervised machine learning. Such a technique relies on a single (continuous) dependent variable and seeks to determine the best linear fit between this outcome and multiple independent variables. MLMs can also be used to group cases based on a collection of variables known for all the cases. In these situations, there is no single outcome of interest and MLMs that are used in this context are referred to as unsupervised learners. One of the most common unsupervised methods with which social scientists and market researchers might have some familiarity is hierarchical cluster analysis (HCA) – also known as segmentation. In this case, the main interest is not on modeling an outcome based on multiple independent variables, as in regression, but rather on understanding if there are combinations of variables (e.g. demographics) that can segment or group sets of customers, respondents, or members of a group, class, or city. The final output of this approach is the actual grouping of the cases within a dataset – where the grouping is determined by the collection of variables available for the analysis.

Unlike many traditional modeling techniques such as ordinary least squares regression, MLMs require a specification of hyperparameters or tuning parameters before a final model and predictions can be obtained. These parameters are often estimated from the data prior to estimating the final model. It could be useful to think of these as settings or "knobs" on the "machine" prior to hitting the start button to generate the predictions. One of the simplest examples of a tuning parameter comes from k-means clustering. Prior to running a k-means clustering algorithm, the method needs to know how many clusters it should produce in the end (i.e. K). The main point is that these tuning parameters are needed before computing final models and predictions. Many machine learning algorithms have only one such hyperparameter (e.g. K-means clustering, least absolute shrinkage and selection operator (LASSO), tree-based models), while others require more than one (e.g. random forests, neural networks).

In contrast to many statistical modeling techniques that have a form specified in advance, MLMs are algorithmic and focus on using data at hand to describe the data-generating mechanism. In applying these more empirical methods in survey

research, one must understand the distinction between models created and used for explanation vs. prediction. Breiman (2001) refers to these two end goals as the two statistical modeling cultures, and Shmueli (2010) refers to them as two modeling paths. The first consists of traditional methods or explanatory models that focus on explanation, while the second consists of predictive models that focus on prediction of continuous outcomes or classification for categorical outcomes. Although machine learning or algorithmic methods can be used to refine explanatory models, their most common application lies in the development of prediction or classification models. The goals and methods for constructing and evaluating models from these two paths overlap to some degree, but in many applications, specific differences can occur that should be understood to maximize their utility in both research and practice.

Explanatory models are commonly used in research and practice to facilitate statistical inferences rather than to make predictions, per se. The underlying shape (e.g. linear, polynomial terms, and nonlinear terms) and content of these models is often informed by the underlying theory, experience of the researcher, or prior literature. Well-constructed explanatory models are then used to investigate hypotheses related to the underlying theory as well as to explore relationships among the predictor variables and the outcome of interest. These models are constructed to maximize explanatory power (e.g. percentage of observed variance explained) and are evaluated by goodness of fit measures and effect sizes. On the other hand, MLMs are often constructed solely to predict or classify continuous or categorical outcomes, respectively, for new cases not yet observed. In contrast to many explanatory models, the actual functional form of the predictive model is often not specified in advance and is often not of primary interest since the main objective in evaluating these models revolves around the overall prediction accuracy (Shmueli 2010) and not on the individual evaluation of the predictors themselves.

Compared to traditional statistical methods, MLMs are more prone to overfitting the data, that is, to detecting patterns that might not generalize to other data. Model development in machine learning hence usually relies on so-called *cross-validation* as one method to curb the risk of overfitting. Cross-validation can be implemented in different ways but the general idea is to use a subsample of the data, referred to as a *training* or *estimation sample*, to develop a predictive model. The remaining sample, not included in the training subsample, is referred to as a *test* or *holdout sample* and is used to evaluate the accuracy of the predictive model developed using the training sample. Some machine learning techniques use a third subsample for tuning, that is, the *validation sample*, to find those tuning parameters that yield the most optimal prediction. In these cases, once a model has been constructed using the training sample and refined using the validation sample, its overall performance is then evaluated using the test sample.

For supervised learners, these three samples contain both the predictor variables (or features) and the outcome (or target) of interest.

The predictive accuracy for machine learning algorithms applied to continuous outcomes is usually quantified using a root mean squared error statistic that compares the observed value of the outcome to predicted value. In classification problems, the predictive accuracy can be estimated using a host of statistics including sensitivity, specificity, and overall accuracy. Generally, the computation of these and related measures of accuracy are based on a *confusion matrix*, which is simply a cross-tabulated table with the rows denoting the actual value of the target variable for every sample or case in your test set and the columns representing the values of the predicted level of the target variable for every sample or case in your test set. Buskirk et al. (2018b) describe confusion matrices in more detail along with a host of accuracy metrics that can be derived from the confusion matrix and used to evaluate classification models, in particular. Kuhn and Johnson (2013) and James et al. (2013) also provide a deeper discussion of using such metrics for evaluating models within a cross-validation framework.

1.3 Creating Sample Designs and Constructing Sampling Frames Using Machine Learning Methods

1.3.1 Sample Design Creation

Many early applications of MLMs for sample design development focused on using these algorithms to create partitions of known population units into optimal groupings to facilitate stratified sampling or some other type of sampling designs. In particular, these approaches apply various types of unsupervised machine learning algorithms such as *k*-means clustering or Gaussian mixture models (GMMs) to an array of frame data and other auxiliary information to segment population units into groups for use as primary sampling units or strata for sampling designs. The population might also refer to a survey panel and auxiliary information might refer to typical frame-related variables as well as survey variables collected during prior phases. In either case, the goal of these approaches is to optimize segmentation of population units into groups that are efficient for sampling rather than on optimizing model fit. Evaluation of the final groupings obtained from the MLMs considers implications for both sampling errors and operational issues.

Although the literature illustrating applications of these types of unsupervised machine learning algorithms for developing sampling designs is small, it is growing. Burgette et al. (2019) compared three different unsupervised learning methods to create sampling strata to understand the range of care delivery structures and

processes being deployed to influence the total costs of caring for patients over time. More specifically, their team compared *k*-means clustering to a mixture of normals (also referred to as GMMs) and mixture of regression models for clustering annual total cost of care statistics over a three-year period for some 40 physician organizations participating in the Integrated Healthcare Association's value-based pay-for-performance program. Of primary interest was the total cost of care and its pattern over this period, so developing a design that groups physician organizations by total cost trajectories and magnitude was crucial. The cluster groupings from each of the three methods considered formed the basis of sampling strata. Burgette et al. (2019) evaluated the results from each method and decided on the results produced using the mixture of normals for their final sampling strata.

The work of Burgette and colleagues highlighted some key, important differences in the assumptions and mechanics of the three clustering methods. Most notably, the *k*-means clustering algorithm is concerned with differences about the cluster means and assumes that the within-cluster variances are equal, and the algorithm seeks to group units so as to minimize this within-cluster variance across the *k*-clusters. A common assumption in many models such as ANOVA is that the assumption of equal stratum variances may be too restrictive and not reasonable. The optimal number of clusters generated from the *k*-means algorithm taken as strata will be too heterogeneous relative to other stratifications that focus on grouping sampling units to reduce variation within each stratum without the restriction or assumption that this reduced variance is the same for every stratum. Burgette et al. (2019) note that the mixture of normal models does not have the assumption of equal variance within each of the clusters and allows this variance to vary within the resulting clusters. The mixture of normal models also allows researchers to incorporate a unit of size as an additional input into the modeling process separate from the variables upon which clustering is based.

Despite the limitations of *k*-means clustering, it is widely used because of its speed and ease of implementation and interpretation. Buskirk et al. (2018a) also explored the utility of the *k*-means clustering algorithm applied to telephone exchange data. The goal was to create a stratified sampling design to be used to select households from 13 council districts in an urban county in the northeastern United States. Using the 13 district geographical boundaries and information about assigned landline telephones, the team created auxiliary information consisting of the percentage of each of the 13 local districts covered by each landline telephone exchange in this county. The *k*-means algorithm grouped telephone exchanges based on these coverage statistics, resulting in eight clusters. Because the final goal was to create nonoverlapping sampling strata for an random digit-dial (RDD) sample that could be used to create estimates of sufficient size from all 13 council districts, the eight clusters were combined into a smaller number of strata. The first stratum consisted of the union of two of these eight

clusters and covered four council districts, whereas the second stratum combined the remaining six clusters to cover the remaining nine council districts. Based on cross-validation information collected during the survey, this design achieved an estimated 97% overall coverage rate with 95% from the first stratum and 98% from the second.

1.3.2 Sample Frame Construction

The applications discussed in Section 1.3.1 have assumed that a population or frame already exists and have focused on leveraging known information about these units to create sample partitions or subsets using unsupervised MLMs. In some situations, however, the frames may overcover the target population and may require modification prior to segmentation to create sampling units that have a higher incidence for the target population. In yet other situations, no known frames or systematic collection of population units are available. To this end, a growing body of literature is focused on applying MLMs to *create* sampling frames or revise existing frames to improve coverage for the intended target population. Some of these applications use existing frame data, auxiliary data, and survey data to create models predicting whether or not population units in an existing frame belong to the target population while other applications use Big Data sources, such as satellite images, to create models that will identify locations of sampling units.

Garber (2009) used classification trees to predict eligibility of units included in a master mailing list for a survey targeting farms. Reductions in data collection inefficiencies due to overcoverage resulted from removing those units with low likelihood of eligibility. Bosch et al. (2018) discuss how they added web address information to a business register composed of a list of enterprises using a focused web scraping approach. Because the crawler might have misidentified the uniform record locator (URL), the team used MLMs to predict whether the URL was appropriate for each enterprise in the database. In this application, MLMs were not applied directly to identify population units or to segment them, but rather as part of the process for editing information on the frame obtained via web scraping methods. Chew et al. (2018) combine the approaches of frame refinement with frame creation by applying a two-category classification task to predict whether a satellite image scene is residential or nonresidential within a gridded population sampling framework.

Gridded samples use a geographic information system to partition areas of interest into logistically manageable grid cells for sampling (Thomson et al. 2017; Amer 2015). These designs typically involve several stages with one of the key stages selecting a series of primary grid units (PGUs) followed by a subselection of many secondary grid units (SGUs) within the PGUs. In many applications of this approach in developing countries, these SGUs are sparsely populated and

consequently may have no eligible household units within them. To improve sampling efficiency, a layer of coding is often deployed within the selected PGUs to eliminate SGUs that contain no eligible population units prior to the phase of sampling that selects a subset of the SGUs. Human coders have traditionally been used to perform these important frame revision tasks but costs for this labor continue to rise and press against budget constraints. Chew et al. (2018) explored how a collection of machine learning algorithms compare to human coders for accurately determining whether SGUs, within selected PGUs from two different developing counties, contain residential units making them eligible for inclusion in further sampling stages of the gridded population design. The team compared MLMs (such as random forests, decision trees, and support vector machines, among others) that used predictors from an administrative database along with other algorithms such as convolutional neural networks (CNNs) (as described in Eck 2018, for example) that used satellite images to predict whether SGUs should be classified as residential (and thus eligible) or nonresidential. They found that although administrative data helped predict whether SGUs were residential, neural network models using aerial images performed much better, nearly as well as human coders for one country, and outperformed human coders in another. This performance is encouraging, but the authors note that the processing and availability of aerial images may not be ubiquitous for all regions or countries.

Whereas the work of Chew et al. (2018) focused on object recognition tasks, with the primary objective of classifying the object in the image (e.g. residential or nonresidential), other research has focused on applying similar machine learning algorithms for object detection, with the primary task of counting the number and/or locations of specific images (e.g. houses). Eck et al. (2018) applied CNN models to aerial images of land plots from across a Midwestern state in the United States to generate a sampling frame of windmills. They were interested not only in classifying a sample of land plots into those with and without windmills but also in determining the location of the windmills within those land plots. Based on processing images from 10 counties within this state, the algorithm exhibited about 95% sensitivity missing only 95 of the 1913 windmills within these 10 counties. The specificity was also high at nearly 98% – however, because many land plots have no windmills, even this 2% misclassification results in nearly 8000 land plots incorrectly identified as having a windmill. However, human coders might be able to examine all images predicted to be windmills and provide final classifications for these images reducing the problem, in this case, from looking at about 9800 images instead of 300 000. So the machine learning algorithm can provide a focused set of land plots that need final adjudication by human coders, thus reducing the total costs in constructing a final sampling frame.

Identifying whether a unit belongs in the sampling frame is an important aspect of frame refinement or digitization of eligibility. Such an application

though assumes that sampling units are already identified, either from the beginning, or via some earlier stage of sampling. Another application of machine learning involves the actual identification of sampling units from a collection of aerial images. In these applications, it is not sufficient to know only that an image contains sampling units because some images may contain more than one such population unit and ultimately, the population will be sampled at the unit level. Eck et al. (2018) applied CNNs to a series of aerial images to determine the location of windmills in counties across Iowa in the United States. They found that the neural networks model was adequate for correctly identifying land plots that contained a windmill, but at the cost of a large number of false positives. Eck et al. (2018) also sought to use these models to generate estimates of the total number of windmills from each county in Iowa. Such an estimate could be used by designs as a measure of size to facilitate selection of counties proportional to the number of windmills or stratification of counties based on the number of windmills. Because of the high number of false positives from the CNN model applied to the land plots, the estimated total number of windmills varied widely across the counties with all estimates being too high by between 95 and just over 1300 windmills. Eck and colleagues refined their approach in 2019 to include a stratification step prior to training the CNN models that ensured more variability in images without windmills that found that this step greatly reduced the number of false positives. They also examined the relationship between the source of the images and various tuning parameters and the coverage properties of the resulting sampling frame. They found that coverage rates were consistent across a range of tuning parameters and that the number of images used for training had relatively little impact on the accuracy measures. Including stratification in the selection of training data was important for reducing the number of false negatives (i.e. increased sensitivity) but had relatively little impact on the false negative rate. They posit that refinements in the stratification approach may be needed to improve the false negative rates moving forward.

1.3.3 Considerations and Implications for Applying Machine Learning Methods for Creating Sampling Frames and Designs

The synthesis of this small collection of studies suggests that the application of MLMs for sampling frame and design development is not automatic and algorithmic, and data error implications should be considered and balanced carefully. We discuss each of these broad considerations in turn here.

1.3.3.1 Considerations About Algorithmic Optimization

The results deemed optimal for grouping population units into groups based on an MLM may not be optimal for sampling or data collection. For example, some

results from unsupervised learning models may produce groups with a single or too few population units to support the sampling design or estimation of sampling variance. Generally, for stratification, one wants a sample of at least two population units selected per group and in some cases, larger samples may be needed for other analytic objectives including comparisons across groups. In these cases, we need to ensure that the final segmentation has sufficient numbers of population units per unit. As Burgette et al. (2019, p. 6) point out "we are less concerned with model fit and more concerned with the sample characteristics that result from using the clusters [segments] as sample strata."

The MLMs are using an objective for optimization that is related to what researchers desire – a natural grouping of population units into segments. However, to be useful for sampling, these segments may need to have a minimum size or the number of such segments may need to be constrained to meet sampling field operations requirements or data collection cost constraints, none of which are considered in the algorithms. That is, the requirements of sample designs and survey operations are not seen by current clustering algorithms during their processing of available data to create potential sampling strata or primary sampling unit groupings. As a result, the results from the MLM serve as the basis or first step that is then refined to meet additional sampling requirements or data collection constraints.

1.3.3.2 Implications About Machine Learning Model Error

Although using MLMs has notable advantages for creating sampling designs, these methods are not without error. Misclassification errors result when MLMs fail to correctly classify units. In sample frame development, machine learning algorithms are applied to classify an image, land plot, or some other related unit as eligible for inclusion in the frame or ineligible. A false positive misclassification error results in a frame that might result in overcoverage of the actual population and lead to inefficiencies in survey data collection. A false negative misclassification error results in undercoverage and may lead to biased estimates if there is a systematic difference between population units correctly included on the frame and those excluded. To our knowledge, there is little work on the impact of this type of machine learning error on final survey estimates although Eck et al. (2019) discuss this as a future area of research. Moreover, there is even less understanding of how tuning parameter selection in these machine learning algorithms might be related to these potential survey errors and how one might incorporate this aspect of survey estimation into the classification algorithm.

By default, most machine learning algorithms optimize classifications to balance false positives with false negatives; however, one can use additional cost functions within these algorithms to penalize false positives at a higher rate than false negatives or vice versa. In frame development, one might argue that false negatives

(leading to undercoverage) may be the more critical error. However, if there are no systematic differences on the survey outcome between correctly included and incorrectly excluded cases, then there might be more tolerance for this type of error. On the other hand, one could argue that if data collection costs are expensive (as in in-person or on-the-ground data collection or collection via expensive sensor equipment), then false positives might be the more critical error. If the false positive rate is high, for example then an eligibility screener may be necessary as an additional survey process step to ensure that only eligible population units be included in final analyses. In any case, it seems important to be able to quantify the impact of these errors before using MLMs for sampling frame development so that either the final results or choices of which of many MLMs lead to the best possible sample frame for use in the final sampling design can be assessed accordingly.

1.3.3.3 Data Type Considerations and Implications About Data Errors

Most of the unsupervised MLMs create population groupings using a collection of continuous covariates. Cluster solutions are often sensitive to variable scaling so if the continuous variables have different scales, transforming the variables so they are all on the same scale is often recommended (Hastie, Tibshirani, and Friedman 2001). However, not all frame or auxiliary variables that are available for use in such segmentation are continuous. Scaling binary or nominal variables and treating them as continuous variables does not make much sense and could impact interpretability of results. Sampling designs seeking to leverage auxiliary information that is a mix of continuous, ordinal, and nominal variables may need to be modified by selecting a different proximity measure or using a method that allows for a mix of variable types such as hierarchical clustering with an appropriate distance measure. Boriah et al. (2008) compared 14 different similarity measures for categorical data for use with a *k*-nearest neighbors algorithm for outlier detection. Their experimental results suggest that there is no one best performing similarity measure, and the authors urge researchers to carefully consider how a particular similarity measure handles the different characteristics of the categorical variables in a given dataset *prior to* selecting the measure. Rather than focusing on different proximity measures for clustering data based on categorical variables, researchers have also explored using different statistics for creating the clusters including *k*-modes (Huang 1998) or fuzzy *k*-modes (Huang and Ng 1999; Kim, Lee, and Lee 2004). These methods use a single distance function appropriate for categorical variables. For datasets with categorical and continuous variables, Huang (1998) proposes a *k*-prototypes clustering algorithm based on a combination of two distance functions applied to categorical and continuous variables, respectively. When cases are to be clustered using only categorical variables or a mix of categorical and continuous variables, as often occurs in survey research applications,

the results of the clustering will likely be more reproducible and interpretable and more likely to represent the underlying constructs, if any, within the collection of variables. Sampling designs seeking to give variables unequal influence on the creation of population segments may also benefit from using an additional weighting variable in the unsupervised MLM where the weights specify each variable's relative influence on the overall segmentation.

Regardless of the type of data used in the machine learning algorithms, the impact of error within the variables themselves on the overall segmentation results is not well understood. Errors in the measurement of auxiliary, frame, or survey variables that are used in unsupervised machine learning algorithms to create population segments are of particular interest since we readily experience and quantify this type of error within the total survey error framework. Pankowska et al. (2018) explored the impact of survey measurement error on the correct classification of respondents into known groups using both GMMs as well as density-based spatial clustering of applications with noise (DBSCAN). Their simulation experiment varied the type of measurement error, the number of variables considered in the clustering that had error, the error rate, and the magnitude of the error. Pankowska and colleagues found that GMM is less sensitive to measurement errors compared to DBSCAN. Measurement error, regardless of method, has a very strong biasing effect for correctly recovering the true underlying grouping structure if the measurement error is systematic, rather than random, and all variables have high levels of measurement error. This work provides insights into the interplay between measurement error in surveys and the results of the clustering algorithms and suggests that cluster algorithms are not equally sensitive or robust to these types of survey errors. The measurement error found in survey variables that might be used to create sampling designs may have a compounding effect under certain conditions as suggested by the work of Pankowska et al. (2018). More work is needed to understand the degree of this compounding and whether other popular clustering methods may be robust against it. This example is just one of many where survey methodologists and statisticians might be able to improve these types of algorithms. The implications of revised algorithms that can incorporate error properties of data within them have a great potential for not only survey data but also for other Big Data sources, especially for Big Data sources with error properties similar to measurement error.

1.4 Questionnaire Design and Evaluation Using Machine Learning Methods

Questionnaire design, pretesting, and evaluating survey questions, improving instrument design, and incorporating alternative measures alongside survey

data are other components in the survey lifecycle that can benefit from recent developments in MLMs.

1.4.1 Question Wording

Although applications demonstrating the potential of machine learning for questionnaire design (or developing question wording in the more traditional sense) have been developed successfully, users may not know that many of these tools rely on MLMs.

The free online application Survey Quality Predictor (SQP) *2.0* (http://sqp.upf.edu/) is one such example. SQP allows researchers to assess the quality of their questions, including suggestions for improvements, either prior to data collection or if this assessment is done after data collection, to allow researchers to quantify (and correct for) potentially biasing impact of measurement error (Oberski and DeCastellarnau 2019). For this assessment to work, researchers first have to code their question along several different dimensions including topic, wording, response scale, and administration mode (using the coding system developed by Saris and Gallhofer 2007, 2014). The second step is that the system then predicts the quality of said question, or more specifically, two indicators of measurement error: the reliability and the validity of the survey question.[1] This prediction task is where machine learning comes in: while SQP 1.0 relied on linear regression modeling, SQP 2.0 uses random forests consisting of 1500 individual regression trees to account for possible multicollinearity of question characteristics, nonlinear relationships, and interaction effects (Saris and Gallhofer 2007, 2014; Oberski, Gruner, and Saris 2011; Oberski and DeCastellarnau 2019).[2] These random forests explain a much higher portion of the variance, namely 65% of the variance in reliability and 84% of validity across the questions in the test sample, compared to the linear regression used in the first version of SQP (reliability: 47%; validity: 61%) (Saris and Gallhofer 2014; Oberski 2016). SQP also provides researchers with an importance measure, that is prioritization, as to which question characteristic is the most influential and how changing a particular question feature, say a 5-point response scale to an 11-point response scale would alter the predicted reliability and validity. Another tool with a slightly different approach is QUAID (question-understanding-aid) (Graesser et al. 2000; Graesser et al. 2006).

1 Reliability in the context of SQP refers to the "theoretical correlation between the true value and the observed answer," whereas validity refers to the "hypothetical spurious correlation between two otherwise unrelated questions" asked using the same method (Oberski and DeCastellarnau 2019, p. 8)).

2 The predictions of a question's quality in SQP 2.0 and higher are based on data from an extensive meta-analysis consisting of 98 survey experiments conducted in 22 countries between 1979 and 2006, with about 3500 questions and about 60 question characteristics (Oberski and DeCastellarnau 2019).

QUAID uses computational linguistics (i.e. lexicons and syntactic parsers) to assist researchers and practitioners in improving the wording, syntax, and semantics of questions.

Other applications in this domain investigate the use of machine learning in respondent and interviewer instructions, or scripted probes. Belli et al. (2016) used classification trees to uncover respondent and interviewer behavioral sequences that predicted respondents engaging in different retrieval strategies in calendar interviews which, as the authors showed, ultimately affect data quality. The challenges with these data are the dependencies within each sequence of behaviors and the (non)occurrence of behaviors at each stage that prohibits the use of more traditional approaches. These results are relevant for questionnaire design because they can directly impact placement and scripting of interviewer probes and interviewer training in how to use these probes. Machine learning can also provide insights as to how to redesign questionnaires based on (real-time) respondent behaviors, for example, to minimize item-level nonresponse, or to prevent break-offs or straight lining. If a participant is estimated to have a high likelihood of breaking off from a web survey while they are in the process of taking it, adaptations can be made to perhaps mitigate the risk of a partial complete due to break-off. One of the main issues is the speed by which models can estimate these propensities as well as being able to make adaptations in real time for online surveys based on these predictions. The other issue to tackle is at what points within the survey is break-off propensity estimated and how frequently an adaptation is made. For now these are open questions, but research into some of these areas is under way. For example, Mittereder and West (2018) used a dynamic Cox survival model based on page invariant respondent characteristics as well as page variant response behavior during the first portion of an annual survey to predict break-off from that point. A separate warning screen was displayed for those respondents with high predicted likelihood of break-off with a message encouraging respondents to continue because their answers were crucial to the goals of the study. The team determined the place where most break-offs occurred and generated a propensity model using survey data from prior years. Using Markov chains (MCs) and recurrent neural networks (RNN) to account for the sequential data, Eck et al. (2015) predict whether a respondent will break off from an online survey, based on paradata collected about respondents interactions with the survey. The authors showed that, compared to MCs, the RNNs achieve a better performance and have a higher precision in that they almost always predict break-offs correctly. More recently, Eck and Soh (2017a) extended these models to predict straight lining in survey grids before it occurs again using paradata. The authors showed that grids placed later in the survey tend to be predicted more accurately, i.e. when there were more opportunities for the algorithm to learn the behavior under investigation, and that the models perform better with more

instances of "bad" behaviors. Using the knowledge from these studies in real time would allow questionnaire designers to incorporate motivation probes or other adaptive survey interventions to retain respondent motivation and prevent straight lining or break-offs (e.g. Eck and Soh 2017b; Mittereder and West 2018; Mittereder 2019).

1.4.2 Evaluation and Testing

In addition to the tools and applications targeted to improve question wording, researchers have used machine learning in question evaluation and testing. The insights from these studies, e.g. which questions contribute to interviewer misreading and which operations in reporting cause respondent reporting errors, can be used to redesign and improve existing questionnaires, serve as a tool to assess interviewer performance and to target interviewers for retraining, and potentially help identify the type of respondent to include for testing.

For example, Timbrook and Eck (2018, 2019) investigated the respondent-interviewer interaction or more specifically interviewer reading behaviors – that is, whether a question was read with or without changes and whether those changes were minor or major. The traditional approach to measure interviewer reading behaviors is behavior coding, which can be very time-consuming and expensive. Timbrook and Eck (2018, 2019) demonstrated that it is possible to partially automate the measurement of these interviewer question reading behaviors using machine learning. The authors compared string comparisons, RNNs, and human coding. Unlike string comparisons that only differentiate between reading with and without changes, RNNs like human coders should be able to reliably differentiate between reading questions with minor or with major changes. Comparing RNNs to string matching (read with change vs. without), the authors showed that exact string comparisons based on preprocessed text were comparable to the other methods. Preprocessing text, however, is resource intensive. If the text was left unprocessed, they showed that the exact string comparisons were less reliable than any other method. In contrast, RNNs can differentiate between different degrees of deviations in question reading. The authors showed that RNNs trained on unprocessed text are comparable to manual, human coding if there is a high prevalence of deviations from exact reading and that the RNNs perform much worse when there is a low prevalence of deviations. Regardless of this prevalence, RNNs performed slightly worse identifying minor changes in question wording. McCarthy and Earp (2009) retrospectively analyzed data from the 2002 Census of Agriculture using classification trees to identify respondents with higher rates of reporting errors in surveys (more specifically, when reporting land size such as total acreage and broken down by individual components depending on land use).

Data from focus groups are also useful for evaluating and pretesting questionnaires, but often the data generated from these sessions is digitized and stored as text. The potential for MLMs to aid in processing this type of textual data along with others gathered throughout the survey process are discussed in more detail in Section 1.6.1.

1.4.3 Instrumentation and Interviewer Training

Finally, other studies exemplified the use of machine learning with a focus on instrumentation, interface development, and interviewer training. Arunachalam et al. (2015) showed how MCs and artificial neural networks (ANNs) can be used to improve computer-assisted telephone interviewing for the American Time Use Survey. Using algorithms that are particularly useful for temporal pattern recognition in combination with paradata, the authors' goal was to predict a respondent's next likely activity. This next likely activity would then be displayed live for interviewers on their computer assisted telephone interviewing (CATI)-screens based on time of day and previous activity to facilitate probing and data entry reducing item nonresponse. This process should ultimately improve data quality and increase data collection efficiencies. Although both algorithms predicted the respondents' activity sequence accurately, the authors found a higher predictive accuracy for the ANNs. Machine learning has also been used to improve the survey instrument for open-ended questions, such as questions regarding occupation (see Section 1.6.1). To facilitate respondent retrieval, decrease respondent burden, and reduce coding errors, Schierholz et al. (2018) investigated computer-assisted coding. More specifically, they assessed the performance of matching algorithms in combination with gradient-boosting decision trees, suggesting a potential occupation based on a verbatim response initially provided by the respondent. Respondents then selected their occupation from this list of suggestions (including an option "different occupation"). The authors showed that the algorithm detected possible categories for 90% of all respondents, of which 80% selected a job title and 15% selected "different occupation" thereby significantly reducing the resources needed for postinterview coding. Other applications of machine learning algorithms, such as regularization networks, test-time feature acquisition, or natural language processing can be used to reduce respondent burden and data collection cost by informing adaptive questionnaire designs (e.g. for nonresponse conversion) in which individuals receive a tailored number or order of questions or question modules, tailored instructions, or particular interventions in real time, depending on responses to earlier questions and paradata (e.g. for surveys more generally, Early 2017; Morrison et al. 2017; Kelly and Doriot 2017; for vignette surveys or conjoint analysis in marketing, Abernethy et al. 2007; for intelligent, dialogue-based

or conversational, tutoring systems, or knowledge assessments, Niraula and Rus 2014).

1.4.4 Alternative Data Sources

Other areas of application of MLMs for questionnaire design include the collection or extraction and processing of data from alternative (Big) Data sources. For example, collecting data from images (e.g. expenditure data from grocery or medical receipts, Jäckle et al. 2019), or websites and apps such as Flickr, Facebook, Instagram or Snapchat (Agarwal et al. 2011), or sensors (e.g. smartphone sensors capturing geo-location and app use, fitness trackers, or eye trackers) may allow researchers to simplify the survey questionnaire and reduce the data collection burden for respondents by dropping some questions entirely. Processing these Big Data, is however, often impossible with standard techniques and requires the use of MLMs to extract features. Among these are deep learning for image processing (e.g. student transcripts,[3] photos of meals or receipts to keep food logs in surveys about food or health,[4] or aerial images to assess neighborhood safety) (Krizhevsky, Sutskever, and Hinton 2012); natural language processing (e.g. to understand spoken meal descriptions (Korpusik et al. 2016); code student transcripts (Shorey et al. 2018)), or the use of Naïve Bayesian classifiers or density-based spatial clustering algorithms (e.g. applied to high-dimensional sensor data from smartphones to optimize the content, frequency, and timing of intervention notifications (e.g. Morrison et al. 2017), to detect home location (e.g. Vanhoof et al. 2018), or to investigate the relationship between location data and an individual's behavior such as exercise frequency (e.g. Eckman et al. 2019)).

1.5 Survey Recruitment and Data Collection Using Machine Learning Methods

Machine learning can also be useful for survey recruitment to monitor and manage data collection and to tailor data collection protocols in terms of expected response rates, nonresponse or measurement bias, or cost considerations.

---

3 While Shorey et al. (2018) focus primarily on the application of deep learning techniques enhanced with structured prior knowledge in the Minerva™ Knowledge Graph to improve coding, imputation, and quality control of survey data in postsecondary education, they also briefly discuss data extraction when processing transcripts, e.g. using optical character recognition software to extract information from pdfs.

4 Calorie Mama AI (https://www.caloriemama.ai/) is one example of an app-based food AI API to identify and track food items based on deep learning and image classification.

1.5.1 Monitoring and Interviewer Falsification

Opportunities for interviewer monitoring and the detection of fabricated interviews during survey recruitment and data collection using machine learning instead of principal component analysis, multiple correspondence analysis, or logistic regression (e.g. Blasius and Thiessen 2015; De Haas and Winker 2016) are now being explored. Using survey and paradata, for example, Birnbaum (2012) investigated different scenarios[5] using supervised (logistic regression and random forest) and unsupervised techniques (e.g. outlier detection via local correlation integral algorithms) to detect fabricated interviews and describes the anomalies found in datasets. Although the logistic regression performed comparably to the random forest at the individual level, the random forest outperformed logistic regression (92% accuracy) at the interviewer level (97% accuracy). The unsupervised algorithms also performed significantly above chance. Sun et al. (2019) took a different approach to detect interviewer falsification and performance issues. The authors relied on audio-recordings of survey interviews and speech recognition to identify fabrication (via the number of speakers) and performance issues (via interviewer reading behaviors; see also Timbrook and Eck (2018, 2019)). They could accurately detect the number of speakers and question misreading in their lab study, but the authors also showed that detecting the number of speakers was highly dependent on the quality of the audio recording.

1.5.2 Responsive and Adaptive Designs

In 2012, the US Census Bureau set out to develop a new model-based hard-to-count score by organizing the Census Return Rate Challenge that researchers and practitioners could use to tailor data collection protocols by "stratify[ing] and target[ing] geographic areas according to propensity to self-respond in sample surveys and censuses" (Erdman and Bates 2017, p. 144). The challenge asked teams to predict the 2010 Census mail-return rates using data from the census planning database (PDB). Interestingly, the three winning models used some form of ensemble methods (more specifically, gradient boosting or random forests), and the winning model included over 300 predictor variables. Using MLMs to prioritize variables, the Census Bureau then proceeded to derive their new low response score (using ordinary least squares including 25 predictors identified as most important in the winning models). These data are now available for researchers and practitioners to use at the block group and the tract level of the PDB to adjust and tailor their data collection protocols. MLMs

5 Assuming a training dataset in which it is known which interviews are fabricated and a scenario in which this is unknown, and whether or not researchers want to investigate fabrication at the question(s), the individual or the interviewer-level.

have also been applied in federal settings to prioritize resources. In particular, Earp and McCarthy (2009) used classification trees to determine characteristics of units most likely to respond both with and without incentives for the Agricultural Resource Management Survey (ARMS). From these models, the team was able to create an efficient allocation plan that provided incentives to those units for which they would likely be most effective.

Chew and Biemer (2018) presented another approach to improve the efficiency of adaptive and responsive survey designs by dynamically adjusting designs during data collection using multi-armed bandits (MABs). Stochastic MABs rely on reinforcement learning at different steps of the data collection process (e.g. using sample batches) to balance the tradeoff between "exploiting" the currently most successful intervention (e.g. with respect to response rates) while also "exploring" alternative interventions. Instead of randomly allocating the entire sample into a fixed control and treatment group prior to data collection and then assessing the effectiveness of interventions after data collection ends, the MAB algorithm iteratively assesses treatment effectiveness of each condition during data collection starting with a smaller sample replicate and iteratively assigns remaining replicates. This iterative approach allows more sampled cases to be allocated to the superior treatment earlier compared to a traditional experimental approach (however, potentially at the cost of unequal variances). Chew and Biemer (2018) compared the performance of a traditional experiment with that of a batched MAB simulating unit response and showed a modest improvement in response rates with MAB that tended to increase with larger effect sizes. The authors also showed that MABs never perform worse than the experiment: even if there is no superior treatment, MABs will always default to continue with random assignment and in that situation can be considered a special case of a traditional experiment.

Another approach to tailor interventions, or rather to optimize the frequency, timing, and content of notifications to deliver cognitive-behavioral interventions (e.g. to use a stress management app) was presented by Morrison et al. (2017).[6] Comparing the performance of intelligent, sensor-driven notifications (accounting for contextual information based on a Naïve Bayesian classifier), to daily and occasional notifications with predefined time frames, the authors showed there is little difference in the percentage of notifications viewed (intelligent 30%, daily 28%, and occasional 30%). However, they also showed that the optimized timing of the intelligent intervention increased the percentage of individuals participating in the intervention suggested by the notification (intelligent 25%, daily 19%, and occa-

6 Just-in-Time Adaptive Interventions (JITAIs) are a class of mobile health interventions for tailored, real-time behavior change support that is system-triggered, e.g. via the use of sensor information (e.g. Hardeman et al. 2019; Spruijt-Metz and Nilsen 2014; Nahum-Shani et al. 2018; Schembre et al. 2018).

sional 19%). These findings are particularly relevant for survey researchers aiming to optimize and tailor the timing and frequency of contact attempts to increase unit response and for researchers using ecological momentary assessments (e.g. in combination with global positioning system (GPS) and accelerometer data) to trigger measurements based on activities (e.g. Stone and Shiffman 1994; Stone et al. 2007).

MLMs can be useful for recruitment in longitudinal settings (panels and repeated cross-sections with a similar data-generating process), for example, to predict panel attrition to ultimately tailor data collection protocols in subsequent waves (e.g. Liu and Wang 2018; Kolb, Kern, and Weiß 2018; Mulder and Kieruj 2018) or mode preferences (Megra et al. 2018). Kern et al. (2019) compared five different tree-based algorithms to a standard logistic regression model predicting panel attrition. The authors showed that each of the different tree-based algorithms had certain advantages that "can be utilized in different contexts" (Kern et al. 2019, p. 87). While the ensemble methods, that is random forests and extreme gradient boosting, performed particularly well for predicting nonresponse, other tree-based methods, such as model-based recursive partitioning or conditional inference trees, provided insights into distinct attrition patterns for certain subgroups. Liu and Wang (2018) took a similar approach but compared three different MLMs – random forests, support vector machines, and LASSO – to a traditional logistic regression model to predict the likelihood to attrite in a follow-up survey using prior wave information. Since these MLMs require (almost) no model specification regardless of whether the true relationships are relevant, linear or not, monotonic or not, or a product of self-selection, they have the potential to increase prediction accuracy and decrease bias by more accurately detecting the true complexity of the underlying model. The MLMs that Liu and Wang (2018) explored did not outperform logistic regression in terms of accuracy, sensitivity, or specificity, but these models provided valuable insights into the relative importance and prioritization of individual variables for the prediction of attrition.

The use of MLMs to derive propensity models and to adapt or tailor data collection designs is not limited to panel studies. Buskirk et al. (2013) compared the use of logistic regression and random forest models to derive response propensity models for a large periodic, national, cross-sectional survey. The models were estimated using responses from one cross-sectional sample along with available household characteristics as well as aggregated census block-group information. These models were then applied to an entirely new sample selected in a subsequent quarter of the year using the same sampling design to estimate temporal and external validity of the response propensity models. Buskirk et al. (2013) found that the random forest propensity model had slightly lower temporal validity, as measured by model accuracy and area under the curve statistics, when

compared to a logistic regression model constructed using principal components. Earp et al. (2014) applied an ensemble of trees to predict nonresponse using census data and response status from multiple years of an establishment survey. They used this model to predict response of establishments for the next survey and found a weak, but significant, relationship between actual and estimated survey response. These studies highlight the importance of considering temporal and external validity measures of performance for models developed and applied in a responsive/adaptive survey design context, especially if we suspect that respondents from future waves of data collection might have different response patterns or if the relationship between predictors and response changes over time.

As these examples illustrate, many types of variables are included in the propensity models deployed within a responsive or tailored survey design and range from only frame or auxiliary variables to a combination of these variables and survey data. A growing body of work is exploring the application of MLMs to paradata to create variables for inclusion in these propensity models, such as information regarding respondent's prior survey experience or unstructured text from interviewer comments. Lugtig and Blom (2018) showed that paradata such as break-offs in the prior survey wave, the time between the survey invitation and completion, whether or not respondents needed reminders, and interview duration were among the best predictors of panel attrition and that these predictors were largely consistent across waves. More importantly, using classification and regression trees (CARTs), the authors demonstrated that complex interactions between these paradata indicators contributed to the prediction of panel attrition. MLMs can also be used to extract indicators from paradata including unstructured interviewer comments. Guzman (2018) used two natural language processing methods – sentiment analysis and topic modeling – to extract information from interviewer comments to then investigate the relationship between these open-ended interviewer comments and panel attrition (see also Ward and Reimer 2019 for a sentiment analysis of interviewer comments and panel attrition). The author showed that both negative sentiments and the topics covered in the interviewer notes were significant predictors of survey completion. Ward and Reimer (2019) noted, however, that while the sentiments found in interviewer comments were predictive, they also tended to correlate with other paradata such as number of contact attempts. Other challenges when working with interviewer comments in particular pertained to the fact that interviewer comments tended to be in shorthand, include abbreviations, and incomplete sentences (Guzman 2018; Ward and Reimer 2019).

These examples illustrate how propensity scores have been developed with MLMs to improve unit-level responses in an adaptive or responsive survey design context. Applications of MLMs for estimating propensity scores used for calibrating and adjusting samples are further explored in Section 1.7.

1.6 Survey Data Coding and Processing Using Machine Learning Methods

One of the most labor-intensive aspects of the survey research process apart from interviewing is the coding and processing of survey data. We consider processing and coding to encompass data validation, editing, imputation, and data linkage. Research applying MLMs to each of these aspects of coding and processing is beginning to emerge, and we explore a great number of these in turn in this section.

1.6.1 Coding Unstructured Text

For the processing of unstructured text and open-ended responses, survey researchers often spend a lot of time and resources to manually code this type of response or discard them altogether (e.g. Conrad et al. 2016; see Couper and Zhang 2016 for an exploration how to incorporate technology in web surveys to assist respondents). Text mining or natural language processing can help to at least partially automate some of this process, and unsupervised algorithms can help researchers to infer topics from the data rather than applying a predefined set of categories (for earlier approaches to automation see, e.g. Creecy et al. 1992 or the literature review for coding and classifying medical records by Stanfill et al. 2010). Schonlau and Couper (2016), for example demonstrated that gradient boosting can categorize 47–58% of the open-ended responses to single categories in surveys with an accuracy of 80% or higher, thereby significantly increasing efficiencies particularly in large surveys. Given these results, the authors suggested using hybrid approaches, that is, automated categorizations, where possible, complemented with manual categorizations where necessary. An extension of this approach using support vector machines (e.g. Schonlau, Gweon, and Wenemark 2018), L1-regularized logistic regression, or RNN (e.g. Card and Smith) is more flexible in that it allows for the coding of multi-categorical data. The study by Schierholz et al. (2018; see Section 1.4.3) also supported the arguments for a hybrid approach, however, in the context of predictive occupation coders suggesting candidate job categories to respondents during the survey interview. In a follow-up study Schierholz (2018) compared the performance of 10 algorithms – among those multinomial logistic regression, gradient boosting with decision trees, adapted nearest-neighbor matching, memory-based reasoning, and an algorithm based on string similarities and Bayesian ideas. The author showed that depending on the goal of the prediction task (e.g. fully automated coding vs. computer-assisted coding), different algorithms tended to perform better, although their approach, based on string similarities and Bayes, tended to outperform the other algorithms.

Shorey et al. (2018) used a knowledge graph developed for natural language processing tasks to incorporate prior domain knowledge (e.g. about the US education system) to prepare student transcripts for statistical analyses. The advantage of combining knowledge graphs with neural networks to (partially) automate coding courses from transcripts is improved information extraction and classification, that is, improved modeling power and decreased computational cost. More specifically, the authors showed that their system can make successful recommendations to human coders 94% of the time (i.e. the correct course code is suggested in the top five most likely predictions), thereby increasing intercoder reliability between 3% and 5% points (e.g. from 84% to 87% for the general course category) and coding efficiency.

Korpusik et al. (2016) investigated coding of text and spoken meal descriptions (for the latter using a deep neural network acoustic model for the speech recognition) comparing different algorithms to assign semantic labels to these spoken descriptions (e.g. *k*-means clustering) and then in a second step comparing different algorithms (e.g. Naïve Bayes, logistic regression, and random forests) to associate food-properties with the given label. They showed that adding word vector features significantly improved model performance in both stages.

Instead of using supervised learners that require a labeled dataset and prespecified categories, a series of more recent studies investigated the use of topic modeling to create clusters of categorizations that are not predetermined by researchers (e.g. Can, Engel, and Keck 2018; McCreanor, Wronski, and Chen 2018; Ye, Medway, and Kelley 2018). Areas of application for topic modeling are manifold:

- Topic modeling can be used to analyze open-ended responses. For example, Can et al. (2018) discussed how open-ended questions on data privacy in surveys compare to publicly available comments in newspaper articles on the same topic. The authors also compared topic modeling with the more traditional content analysis for each data source and showed that different data sources and different modeling approaches may lead to different substantive conclusions. Roberts et al. (2014) used semiautomated structural topic modeling to analyze open-ended responses. Structural topic models have the advantage of incorporating contextual information about the document (e.g. covariates such as, gender, or treatment assignment), allowing topic prevalence and topic content to vary, and thereby allowing the estimation of treatment effects.
- Topic modeling can also be used to analyze larger bodies of unstructured text. One example for the analyses of unstructured text involved the use of two different topic modeling approaches to identify research on emerging and published topics and trends in survey statistics and methodology by mining 10 years of abstracts from relevant conferences and journals (Thaung and Kolenikov 2018).

- Topic modeling can also assist researchers to categorize question characteristics. Sweitzer and Shulman (2018), for example, used topic modeling to code question topics in their analysis of the relationship between language difficulty and data quality in public opinion questions.

1.6.2 Data Validation and Editing

Data validation and data editing, that is, the examination and correction of data for errors, has also benefited from recent developments in machine learning. For surveys that are repeated periodically, instead of having subject-matter experts specify the conditions that data edits need to satisfy, tree-based methods can be used to automatically identify logical statements about the edits to be performed using the historic, clean data as a training dataset (e.g. Petrakos et al. 2004; Ruiz 2018). TREEVAL is one example of an automated data editing tool and validation strategy that derives the optimal functional form of the edits and of statistical criteria based on the training data to edit the raw incoming data (Petrakos et al. 2004). An extension of this approach would be to use unsupervised machine learning algorithms to "identify expected patterns, systematic mistakes (unexpected patterns) or to observe cases that do not fall into any patterns" both during and after data collection (Ruiz 2018, p. 10).

1.6.3 Imputation

Another area of application for MLMs is imputation at the unit or item level (e.g. Lu, Li, and Pan 2007; Borgoni and Berrington 2013; Beretta and Santaniello 2016; Loh et al. 2019). More traditional imputation methods addressing item nonresponse often face challenges. These approaches (i) often assume a (too) simplistic missing data pattern and implement either univariate or simple multivariate imputation (such as a monotone missing data patterns instead of more complex missing data patterns), (ii) can be computationally inefficient, (iii) can be impossible to estimate if there are too many auxiliary variables, or a high rate of missing data in these auxiliary variables. To address these challenges, researchers recently started investigating alternative approaches including tree-based techniques. Borgoni and Berrington (2013), for example, demonstrated the flexibility of an iterative sequential tree-based imputation assuming a complex missing data pattern and showed that the results compare to existing techniques. Because these tree-based models are nonparametric (an advantage), they are less sensitive to model misspecification and outliers while being computationally efficient even with many variables to be imputed (see also Fischer and Mittereder 2018). Loh et al. (2019) exposed CART and forest to a series of robustness analyses and showed that tree- and forest-based methods are always applicable, fast, and

efficient even under what they call "real data" situations where other methods, such as likelihood-based approaches fail. At the unit level, Drechsler and Reiter (2011) provided an example of using tree- and forest-based methods as well as support vector machines to create synthetic data in instances in which intense redaction was needed (see also Caiola and Reiter 2010 to create partially synthetic categorical data). The authors showed that tree-based methods provide reliable estimates and minimize disclosure risks.

1.6.4 Record Linkage and Duplicate Detection

Record linkage – e.g., linking records of individuals from administrative records and survey data, traditionally either using rule-based or probabilistic approaches – can also benefit from recent developments in machine learning (Winkler 2006). Supervised machine learning algorithms, such as random forests, develop a matching strategy based on a set of training data that can then be applied out of sample (e.g. Christen 2012; Ventura, Nugent, and Fuchs 2015). The advantage of using machine learning algorithms over more traditional approaches is that (i) they allow the researcher to start with a range of predictor variables and then assess variable importance to tailor the linkage strategy and determine whether a field is relevant for linkage; and (ii) they allow predictor variables to be continuous contrary to some of the more traditional approaches. Morris (2017) investigates the use of logistic regression, classification trees, and random forests to investigate the quality of administrative records composite matching a person to their (US) census day address using administrative records, characteristics of the person, and the housing unit. Their approach ultimately allowed researchers to address the question as to whether or not administrative data can be used as a proxy for nonrespondents in the 2010, and potentially the 2020, US Census and thus contribute to cost savings if a case that was successfully matched could be excluded from fieldwork (this is also an example of responsive designs, see Section 1.5.2). The author showed how the choice of a particular classification technique can imply a tradeoff between simplicity and predictive accuracy.

Unsupervised techniques have the advantage in that they do not require a labeled training dataset and can be used to assess proximity between cases from different datasets. Supervised and unsupervised MLMs have also been used in sample matching (see Section 1.7.2). Related to record linkage is the detection of duplicate records. Elmagarmid et al. (2007) presented a comprehensive review of existing traditional and machine learning algorithms used to detect nonidentical duplicate entries in datasets (for record linkage and deduplication, see Christen 2012).

1.7 Sample Weighting and Survey Adjustments Using Machine Learning Methods

Probability sample survey data is often adjusted to account for the probabilities of inclusion as well as to mitigate potential biases due to various sources of nonsampling error including coverage and response. Similar adjustments are also made to nonprobability samples to account for nonsampling errors, most notably for coverage and self-selection (Valliant, Dever, and Kreuter 2013). Although raking and poststratification (Kalton and Flores-Cervantes 2003; Valliant et al. 2013) are two of the most widely applied techniques to correct for coverage related issues for both probability and nonprobability samples, to our knowledge there has been little to no work applying MLMs to these calibration techniques. On the other hand, a steady stream of research has explored various MLMs for propensity score modeling and sample matching – two of the more common methods for adjusting for nonresponse and self-selection (Valliant and Dever 2011; Elliott and Valliant 2017; Mercer et al. 2017). We discuss recent developments in applying MLMs to sample matching and propensity score estimation as it relates to sample weighting and adjustment. Propensity score models constructed to identify respondents or used for sample management or recruitment were discussed in Section 1.5.

1.7.1 Propensity Score Estimation

Propensity score adjustments trace their origins to observational studies that are used to compare treatment affects when treatments are not assigned at random (Rosenbaum and Rubin 1983). This method was since adapted to adjust for nonresponse in probability sample settings (Brick 2013; Valliant et al. 2013; Baker et al. 2013) and for self-selection in nonprobability sample settings (Valliant et al. 2013; Valliant and Dever 2011; Lee 2006). For probability samples, the propensity scores represent the inverse probability of response modeled as a function of available covariates. In the nonprobability survey setting, propensity score estimation is also used to derive so-called quasi-randomization weights that are estimated from the model using nonprobability data as well as data from a probability-based reference sample (Valliant and Dever 2011; Elliott and Valliant 2017). In either setting, logistic regression was by far the most common approach for estimating propensity models (Lee 2006; Valliant et al. 2013). However, as discussed in Mendez et al. (2008), limitations about specifying the parametric model form, convergence, and possible under specification associated with logistic regression models create opportunities for considering other modeling approaches.

Valliant et al. (2013) discussed how tree-based MLMs may be used for modeling response adjustments and how CARTs have more flexibility over chi-square automatic interaction detector (CHAID) models (Kass 1980) in that they can create predictions using a combination of continuous and categorical variables. As a case in point, Lee et al. (2010) explored the use of supervised machine learning algorithms for estimating propensity scores in an epidemiological observational study. More specifically, they compared propensity scores derived from main effects logistic regression models and a host of MLMs including CART, pruned CART and several ensemble methods including bagged CART, boosted CART, and random forests using various metrics of model performance. Generally, they reported adequate performance across the various models, but note that under moderate nonadditivity and moderate nonlinearity, the logistic regression models fared worse than models developed from ensemble methods such as boosted CART models.

Survey researchers are making similar comparisons about various MLMs used to derive propensity score models used for nonresponse adjustment. Much of the current work on propensity score estimation within the survey context has started to include many supervised machine learning algorithms, especially tree-based methods. In probability settings, the applications of MLMs to propensity scores explored both direct adjustments and adjustments made through propensity score stratification (Bethlehem, Cobben, and Schouten 2011; Brick 2013; Valliant et al. 2013). Schouten and de Nooij (2005) presented a classification tree method to construct weighting strata that simultaneously account for the relation between response behavior, survey questions, and covariates. Phipps and Toth (2012) applied regression trees to data from the Occupational Employment Statistics Survey to estimate response propensities for sampled establishments.

Lohr et al. (2015) also explored propensity estimation with an expanded collection of methods including CART (Breiman et al. 1984), conditional trees (Strobl et al. 2007), random forests (Breiman 2001), and trees with random effects (Sela and Simonoff 2012). Lohr et al. (2015) also evaluated mean squared errors for survey-related estimates based on propensity weights derived using several different approaches including direct adjustment, propensity stratification adjustment, and terminal nodes for the CART and conditional tree models as suggested by Toth and Phipps (2014). Propensity models for each of the MLMs included both a regression approach (e.g. treating binary response as a continuous 0/1 variable) as well as classification (e.g. treating the outcome as binary/categorical). The team tested the performance of these MLMs and adjustment approaches under 32 different sample scenarios formed by varying the response rates, nonresponse mechanisms, number of primary sampling units, and the degree of latent nonresponse within primary sampling units (PSUs). Within each particular MLM, Lohr and colleagues also varied particular tuning parameters such as minimum node size and other

process settings such as the use of pruning or not and incorporating weights or not, among others. This study represents one of the most comprehensive comparisons of propensity methods, theoretical response mechanisms, and MLMs to date. In the end, Lohr et al. (2015) found that unadjusted settings, which included no pruning, no use of weights and other adjustment factors resulted in better performance. They also reported that regression versions of trees and forests performed better than classification versions and found that conditional tree models outperformed CART models in almost every scenario they considered. The team noted that of all the methods, random forests fared the best under direct propensity adjustment but did not fare as well under propensity stratification.

The results reported by Lohr et al. (2015) regarding random forests were fairly consistent with those reported by Buskirk and Kolenikov (2015) who also compared random forest models to logistic regression models for deriving propensity adjustments to sample surveys using both the direct and stratification approaches. However, the results of Buskirk and Kolenikov for random forest models using propensity stratification were far less favorable because the resulting survey estimates had much larger variances compared to direct propensity estimation or to logistic regression models with either direct or stratification adjustments. The team noted that large ranges in estimated propensities resulted in respondents being assigned to different propensity strata across bootstrap resamples that were used in evaluating the variances. Lohr et al. (2015) did not evaluate mean squared errors using replicate variance but instead relied on repeated subsamples taken from a larger population. Differences in these two studies also illuminated an important point when comparing various applications of MLMs to derive and apply propensity adjustments. Namely, to examine differences in results, it is important to consider both the type of adjustment method (e.g. direct vs. stratification) and the type of MLM as well as the tuning parameters and other related settings along with the method of evaluation. Although most MLMs are robust to a wide range of tuning parameters, differences in them can impact the results even when all other settings, data sources, covariates, methods, and adjustment techniques are equal.

Within the nonprobablity framework, estimation of propensity models to create so-called pseudo-weights under the quasi-randomization framework (Valliant and Dever 2011; Elliott and Valliant 2017) have been generated in large part using logistic regression models but a growing body of research is also exploring a wide array of MLMs. The choice of how weights are computed from derived propensities can be an important factor in comparing various MLMs and approaches. Nearly all of the studies we reviewed used the Hajek weights, which equal the quotient of the estimated probability for being in the probability sample divided by the estimated probability the case was from the nonprobability sample as explained by Schonlau and Couper (2017) and Elliott and Valliant (2017). Mercer (2018) used

Bayesian Adaptive Regression Trees to create propensity-based pseudo-weights and doubly robust estimates for several nonprobability surveys. Mercer et al. (2018) expanded the use of tree-based approaches to an ensemble approach by applying random forest models based on two different sets of covariates to derive propensity models for creating pseudo-weights for an array of nonprobability samples. They also apply a calibration step to created pseudo-weights to balance the final weighted nonprobability sample to known population benchmarks as a way to smooth out or minimize the impact of extreme weights. They found that the combination of propensity models along with a final raking step performed better than raking alone, but also that propensity models alone fared worse than just raking. They also note that beyond the method of adjustment what mattered most was the choice of covariates citing that the expanded covariate set consisting of both demographic and politically related variables reduced biases more than just using demographics alone. This result reiterates the strong ignorability assumption that Schonlau and Couper (2017) reference when emphasizing that the success of the propensity scoring adjustments depends strongly on whether the available auxiliary variables adequately capture the difference between the nonprobability sample and the reference sample for all outcomes of interest.

Garcia and Rueba (2018) expanded further still the MLMs applied to deriving pseudo-weights based on propensity models estimated using several decision tree methods, *k*-nearest neighbors, Naïve Bayes, random forests, and gradient-boosting machines under various scenarios that varied the number of covariates and sample sizes for the nonprobability sample. They reported that nearest-neighbor methods provided good performance as long as the number of covariates was small and Naïve Bayes created unstable estimates for most scenarios. They reported that beyond these results, the performance of MLMs varied according to the sample size of the nonprobability sample as well as the type of underlying mechanism governing self-selection.

Although many of these studies used tree-based methods for estimating response propensities, either as adjustment factors in probability samples or as the basis of pseudo-weights in the nonprobability setting, the approach used to estimate these propensities as well as methods to bin them into strata is important. Lohr et al. (2015) as well as Buskirk and Kolenikov (2015) noted that regression approaches to these decision tree-based models proved better for estimating propensities – that is decision trees were applied to binary 0/1 outcome treated as a continuous variable. Settings in the MLMs such as minimum node size might impact the overall range of estimated propensities, making some of these estimated to be 0 and others 1, which will create irregularities in the weighting adjustments. Buskirk and Kolenikov (2015) explicitly transformed the propensities derived from forest models slightly to prevent these irregularities.

Provost and Domingos (2003) encouraged the use of probability estimating tree approaches to creating propensities from these types of MLMs by applying a Laplace correction to the estimated propensities that is a function of the sample size and number of classes underling the prediction problem (e.g. two classes, in the case of response). Their early work on applying this adjustment approach indicated increased accuracy and stability in the resulting estimated propensity scores. Of course, when propensity stratification is used, there is less concern about the irregularities imposed by estimated propensities that are 0. In most of the survey research literature a general rule of 5–10 propensity strata is used (Cochran 1964). However, as Buskirk and Kolenikov (2015) point out, the range of propensities estimated from MLMs is often wider than that observed for logistic regression models, largely because these MLMs often include more complex model structures incorporating important interactions and higher-order effects that might not be included in parametric models based on logistic regression. For propensity stratification to be effective, the values of the propensity scores should be similar within strata (see Lee 2006) and so for the case of MLMs, the common practice of using five groups may be insufficient for the range of estimated propensities often seen with MLMs.

1.7.2 Sample Matching

Sample matching is another technique currently being explored for mitigating the effects of nonresponse in probability surveys and self-selection in nonprobability surveys (Baker et al. 2013; Bethlehem 2015; Elliott and Valliant 2017). Simply put, sample matching in the nonprobability sample setting matches a subset of members from the nonprobability sample to members of a smaller, reference probability sample using a set of common covariates (Elliott and Valliant 2017; Bethlehem 2015). Mercer et al. (2017) drew parallels between sample matching for observational studies and how this approach can be adapted to the nonprobability framework. Essentially, sample matching can be done at the population level, as in quota sampling, or at the individual or case level. MLMs have been almost exclusively applied to facilitating matches at the individual level. Cases are identified as matches by using some type of proximity measure computed based on a set of covariates measured from both the nonprobablity sample and the reference sample. Proximity measures can be computed using MLMs based on any number and combination of categorical and continuous variables offering an advantage over more traditional metrics such as the simple matching coefficient or Euclidean distance which suffer from the so-called “curse of dimensionality” and generally require all variables to be of the same type. These proximity measures can also be used to facilitate record linkage and more applications of MLMs in this context are described in the Section 1.6.4.

Regardless of the number of variables used in sample matching, Mercer et al. (2017) noted that the utility of the method hinges on identifying variables related to the survey outcomes of interest as well as those that ensure at least conditional exchangeability within the nonprobability sample. For a given nonprobability survey, these variables may not be available on the reference survey but could be added rather easily if a small probability-based reference survey was available as was the case in the applications of sample matching employed by Buskirk and Dutwin (2016). Valliant et al. (2013), Bethlehem (2016), and Mercer et al. (2017) noted that the reference survey must be a good representation of the target population for sample matching to be effective in reducing biases. Generally, there is a tradeoff in the flexibility afforded by smaller target samples and the representation offered by larger government surveys, for example (Elliott and Valliant 2017). Mercer et al. (2018) applied a clever method of creating a synthetic reference sample by first imputing variables from the Current Population Survey and the General Social Survey to cases in the American Community Survey using a hot deck procedure powered by a random forest model. An expanded so-called synthetic sample resulted that expanded the possible variables in the American Community Sample available for use in the matching process. They created matches between this reference sample and various nonprobability samples using proximity measures derived from a random forest that classified cases in a combined dataset as being from the reference or nonprobability datasets. Mercer et al. (2018) noted that estimates produced with sample matching using both demographic and political variables resulted in the lowest average bias across the full set of benchmarks that were considered. They also noted that using raking alone resulted in just slightly more biased estimates, overall. Matching also tended to improve as the ratio of the sample size of the nonprobability sample to the reference dataset increased.

Buskirk and Dutwin (2016) also applied random forest models to match nonprobability sample cases to a small probability-based reference survey. However, in their approach, they used an unsupervised version of random forest methods to create a forest model describing the structure present in the probability reference set using two different sets of covariates measured from both samples. Proximity measures between cases in the probability reference sample and nonprobability sample were then estimated using this random forest model. Similar to the work of Mercer et al. (2018), Buskirk and Dutwin explored a simple set of covariates as well as an expanded set of covariates and varied the ratio of the nonprobability sample to the probability sample. They found that the matched samples generated using proximities computed using random forest models were generally consistent with those computed using the simple matching coefficient but in some scenarios were more stable across multiple replicate samples.

To date, we know of no study directly comparing the use of supervised vs. unsupervised applications of random forest models or other MLMs to estimate

proximity measures that can be used in the matching process. One advantage of using the unsupervised approach is inherent in the desired outcome to generate proximity measures within a full range so the matching process can better discern the best matching case. With supervised methods, the algorithms are optimized to create final nodes that are pure in terms of sample parity – putting cases from the probability samples in the same nodes and the same for nonprobability cases. Of course, there are exceptions, and not every node is completely pure allowing for some variability in the proximity measures derived from these supervised tree-based models – which base proximity calculations on a function of the proportion of times two different cases in the combined file fall in the same final node. With the unsupervised approach, there is no label of sample parity that allows the algorithm to naturally find patterns in the datasets based on the variables used in the matching process. One might argue that using the reference sample only in this step is important, especially if it is a better representation of the target population so that these patterns are not spurious or driven by self-selection or other issues possibly present in the nonprobability samples. Then using this model as the basis of measuring proximity between probability and nonprobability is more of a function of the relationships present among the matching variables and as such should lead to a wider range of proximity measures. Buskirk and Dutwin's work illustrated differences in the range of proximity measures created from using unsupervised random forests and those computed using the simple matching coefficient from the same covariate sets and found that the measures from the simple matching coefficient were very negatively skewed and concentrated very close to 1, while those from the random forest model were more evenly spread between 0 and 1.

1.8 Survey Data Analysis and Estimation Using Machine Learning Methods

Analysis of probability-based survey data requires understanding of the sampling design structure and generally, estimates are derived that incorporate this structure and a set of final sampling weights. Two broad frameworks are used to analyze and produce estimates from survey data – the so-called design-based framework that focuses on repeated sampling to generate inference and a model-based framework that leverages a postulated statistical model for the relationships among variables to create estimates. Models can be incorporated into the design-based approach to improve the efficiency of estimates using the model-assisted approach as described by Breidt and Opsomer (2017). Although the model-based framework is not as common for generating estimates from probability samples, it is becoming popular for generating population estimates

using nonprobability samples (Elliott and Valliant 2017). Applications of MLMs for estimation in the survey context are concentrated around three main areas, the first of which leverages capabilities of MLMs to identify key correlates of a survey outcome of interest from among many possible variables measured within the survey context and to gain insights into possible complexities in relationships among these variables. The second area encapsulates adaptations of these MLMs to incorporate survey design and weighting features for use in design-based population inference. And finally, the third area leverages the predictive nature of these MLMs within a model-assisted framework to improve finite population inference for both probability and nonprobability surveys. We review some of the current advances in each of these areas, in turn.

1.8.1 Gaining Insights Among Survey Variables

Arpino et al. (2018) used survival random forests to model changes in marital status over time in Germany based on data from the German Socio-Economic Panel survey to understand determinants of marriage dissolutions in Germany over the past two decades. They used variable importance measures to identify influential variables with the largest impact on predicting dissolution of marriages and partial dependence plots to gain insights into the direction of the association between the most influential variables and marriage dissolution status. Their investigation revealed that relationship between key continuous predictors and the marriage dissolution status may not be linear and that other effects may be affecting the outcome not as a main effect but through the moderation of another variable. These insights highlight future work in confirming these more complex relationships among a key set of predictors and marriage dissolution status.

As the previous example illustrates, variable importance measures, for example, derived from random forest models can be helpful in providing focus on a much smaller set of influential covariates. However, importance measures derived using regular random forests are often biased in the presence of many variables that might be correlated or might be of different variable types (Strobl et al. 2007). In the survey setting, some level of correlation or association almost always exists between the variables for any given prediction problem, and we certainly have a mix of variable types. So to identify a smaller set of important variables within the survey context, an alternate version of variable importance is needed. The fuzzy forests method applies a recursive feature elimination process and has been noted to offer estimates of variable importance that are nearly as unbiased as those computed for the more computationally expensive conditional forests (Conn et al. 2015). Dutwin and Buskirk (2017) applied a series of fuzzy random forest models to first identify a smaller set of important predictors for predicting household Internet status, from a collection of over 500

variables identified from 15 probability-based surveys. The goal was to identify a small, manageable set of variables that could be used to create a model for predicting Internet status and then be used to create weighting adjustments for coverage in subsequent online samples that fielded this small collection of questions. A probability-based RDD survey was fielded that identified about the same number of Internet and non-Internet households that asked the three dozen questions identified as important by the fuzzy forest models. Another fuzzy random forest model applied to the RDD survey data identified a final set of about a total of about 12 demographic and nondemographic variables that were most important for predicting Internet status. A final regular random forest model predicting non-Internet status was fit using data from the RDD survey and model performance measures for this model revealed that the demographic variables were important for identifying households without Internet (e.g. high sensitivity), while the core set of nondemographic variables were most powerful for identifying those households with Internet (e.g. high specificity).

1.8.2 Adapting Machine Learning Methods to the Survey Setting

Although the number of studies making adaptations of existing MLMs to incorporate sample design and weighting information is rather limited to date, we anticipate that such studies will continue to increase in the near future and will collectively represent an important contribution of survey research, statistics, and methodology to data science. McConville et al. (2017) adapted the LASSO and adaptive LASSO approaches to the survey setting. Their work established the theory for applying these methods to probability survey data and proposed two versions of lasso survey regression weights so that this method can be used to develop model-assisted, design-based estimates of multiple survey outcomes of interested. In a similar vein, Toth and Eltinge (2011) developed a method for adapting regression trees via recursive partitioning to the probability survey context by incorporating information about the complex sample design, including design weights. Toth (2017) later developed the R-package, rpms, that makes estimation of regression trees numerically possible and accessible for probability-based survey data. Toth and Eltinge's work also unlocked the potential for using regression trees within the model-assisted approach to design-based inference. McConville and Toth (2019) explored regression tree estimators to automatically determine poststratification adjustment cells using auxiliary data available for all population units. They noted that one strength of this method over other model-assisted approaches is how naturally regression trees can incorporate both continuous and categorical auxiliary data. These poststrata have the ability to capture complex interactions between the variables, and as

shown in the simulations, they can increase the efficiency of the model-assisted estimator. Additionally, the estimator is calibrated to the population totals of each poststratum.

Although not explicitly applicable only to survey data, Zhao et al. (2016) adapted random forest models to handle missing values on covariates and for facilitating improved estimation of proximity measures that rely on all observations rather than just those that are out of bag. Although individual decision trees have the ability to handle missing data via surrogates, this advantage is lost within the general random forest approach by virtue of the mechanics of only selecting a subset of variables for node splitting at each node. Zhao et al. applied this adapted random forest method to examine impacts of smoking on body mass index (BMI) using data from the National Health and Nutrition Examination Survey. Missing values were implicitly handled within their revised approach via surrogates and thus eliminated the need to explicitly imputing missing values for any of the covariates used in the model.

1.8.3 Leveraging Machine Learning Algorithms for Finite Population Inference

Breidt and Opsomer (2017) explored how MLMs can be used as the basis of the working models in the model-assisted approach to design-based estimation for probability samples. They provided a general framework within which predictions from various MLMs can be incorporated and used to derive finite population estimates and inference. They specifically illustrated how methods like *k*-nearest neighbors, CARTs and neural networks could be used within this framework. More recently, Buelens et al. (2018) explored similar approaches for using MLMs for inference from nonprobability samples. Their work compares quasi-randomization methods to model-based methods (or super-population models, as more formally explained by Elliott and Valliant (2017)), where various MLMs are used in creating the models. More specifically, Buelens et al. (2018) compared sample mean estimation, quasi-randomization pseudo-weighting based on poststratification via a known auxiliary variable for the entire population, generalized linear models, and a host of MLMs including *k*-nearest neighbors, ANNs, regression trees and support vector machines as the basis of generating model-based estimates. They tuned each of the MLMs using a repeated split-sample scenario based on 10 bootstrap replications, and the optimal values of the respective tuning parameters were then used to form models upon which final estimates were generated. In predicting a continuous outcome, Buelens and colleagues reported generally adequate, although varied, results from the MLMs compared to using either the sample mean or the pseudo-weighted quasi-randomization based estimator. Generally, the MLMs removed more or

nearly the same amount of bias due to self-selection compared to either the sample mean or pseudo-weighted estimator, especially under moderate to severe levels of self-selection. They also identified support vector machines as a top performer compared to all other methods in almost all scenarios they examined.

Model-based estimates generally work well if the predictions made using the model are well suited for population members not included in the sample. The models are estimated using data from the sample (be it probability or nonprobability) based on a set of covariates that are available from members of the sample and population. If these covariates fail to fully represent the self-selection bias mechanism, the range of values on the outcome of interest may differ between the sample and population members not included in the sample. If so, some models will not be able to generate predicted values that extrapolate beyond the range of values observed from the population and this limitation will result in biased model-based estimates. Buelens and colleagues noted this limitation for the sample mean, pseudo-weighted estimator, *k*-nearest neighbors, and regression trees. They noted that generalized linear models, neural networks, and support vector machines (from among the MLMs they explored) are stronger choices in this situation. Reiterating points of exchangeability made by Mercer et al. (2017), the ability to have auxiliary information that can explain this self-selection and a predictive algorithm that can utilize it are very important. As this current work demonstrates, MLMs are not all equal in their ability to use such information adequately and some have limitations for such applications making method selection an important component in addition to variable selection for creating finite population inference using nonprobability samples.

1.9 Discussion and Conclusions

This collection of examples, while not exhaustive, provides a glimpse into how survey researchers and social scientists are applying an assortment of data science methods within the new landscape. These examples show that data science methods are and can add value to the survey research process. But what is not as clear, yet, is just how the social sciences can add value to the broader data science community and Big Data ecosystem. Grimmer (2015) was quick to identify strengths that social scientists and survey researchers can bring to this conversation by stating, "Data scientists have significantly more experience with large datasets but they tend to have little training in how to infer causal effects in the face of substantial selection. Social scientists must have an integral role in this collaboration; merely being able to apply statistical techniques to massive datasets is insufficient. Rather, the expertise from a field that has handled observational data for many years is required." In fact, the much-reported algorithmic bias

among data scientists is a problem almost certainly related to coverage bias – an issue survey researchers have been investigating and mitigating for decades. More generally, advances on understanding and quantifying data error sources are yet another example where survey researchers can add value to the Big Data ecosystem. But apart from methodology, we believe that survey data in and of itself can add much needed insights into this ecosystem with the value proposition being directly related to the fact that surveys are often designed to maximize information about context and provide signals related to the "why" question. In this vein, survey data offer valuable additional sources of information that can enhance prediction accuracy.

This chapter is in no way comprehensive. Our goal was to provide survey researchers and social scientists, as well as data and computer scientists with some examples and ideas that illustrate how MLMs are and can be used throughout the main phases of the survey research process. If we were to cluster these examples in terms of how MLMs have been used, we would find four major uses simply labeled as processing, preparation, prioritization, and prediction. Within each of these four main application areas we saw many cases where MLMs offered considerable benefit to the survey process. We saw examples that leveraged both human effort and MLMs to create better outcomes with respect to the survey research process. We also saw many examples where the MLMs enhanced the process by providing insights into how best to use human effort and machine effort. However, we saw very few examples where MLMs were working completely autonomously within the survey research process. Collectively, these studies seem to suggest that while MLMs offer efficiencies not realized without their use, human decision-making and engineering the application and deployment of these methods is tantamount to their success throughout the survey research process. Nonetheless, in the new landscape, the rise of MLMs has provided survey and social scientists with additional resources that – while certainly no cure-all for the challenges our fields face – are meaningful and warrant further exploration.

References

Abernethy, J., Evgeniou, T., Toubia, O., and Vert, J.-P. (2007). Eliciting consumer preferences using robust adaptive choice questionnaires. *IEEE Transactions on Knowledge and Data Engineering* 20 (2): 145–155.

Agarwal, S., Furukawa, Y., Snavely, N. et al. (2011). Building Rome in a day. *Communications of the ACM* 54 (10): 105–112. https://doi.org/10.1145/2001269.2001293.

Amer, S.R. (2015). Geo-sampling: from design to implementation. 70th Annual Conference of the American Association for Public Opinion Research, Hollywood, FL (14–17 May 2015).

Arpino, B., Le Moglie, M., and Mencarin, L. (2018). Machine-learning techniques for family demography: an application of random forests to the analysis of divorce determinants in Germany. http://hdl.handle.net/10230/34425 (accessed 5 July 2019).

Arunachalam, H., Atkin, G., Wettlaufer, D., et al. (2015). I know what you did next: predicting respondent's next activity using machine learning. Paper presented at the 70th Annual Conference of the American Association for Public Opinion Research, Hollywood, FL (17 May 2015).

Baker, R., Brick, J.M., Bates, N.A. et al. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1 (2): 90–143. https://doi.org/10.1093/jssam/smt008.

Belli, R., Miller, L.D., Al Baghal, T. et al. (2016). Using data mining to predict the occurrence of respondent retrieval strategies in calendar interviewing: the quality of retrospective reports. *Journal of Official Statistics* 32 (3): 579–600. https://doi.org/10.1515/jos-2016-0030.

Beretta, L. and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making* 16 (S3, Suppl 3): 74. https://doi.org/10.1186/s12911-016-0318-z.

Bethlehem, J. (2015). Essay: Sunday shopping-the case of three surveys. *Survey Research Methods* 9 (3): 221–230.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? *Social Science Computer Review* 34 (1): 59–77.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*, Chapter 11. Hoboken, NJ: Wiley https://doi.org/10.1002/9780470891056.

Birnbaum, B. (2012). Algorithmic approaches to detecting interviewer fabrication in surveys. Doctoral thesis. Department of Computer Science & Engineering, University of Washington, Seattle, WA.

Blasius, J. and Thiessen, V. (2015). Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research* 52: 479–493. https://doi.org/10.1016/j.ssresearch.2015.03.006.

Borgoni, R. and Berrington, A. (2013). Evaluating a sequential tree-based procedure for multivariate imputation of complex missing data structures. *Quality & Quantity* 47 (4): 991–2008. https://doi.org/10.1007/s11135-011-9638-3.

Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity measures for categorical data: a comparative evaluation. *Proceedings of the SDM, SIAM*, Atlanta, GA (April 24–26, 2008). https://doi.org/10.1137/1.9781611972788.22.

Breidt, F.J. and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science* 32 (2): 190–205. https://doi.org/10.1214/16-STS589.

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Breiman, L., Friedman, J.H., Olshen, R.A. et al. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: a critical review. *Journal of Official Statistics* 29 (3): 329–353. https://doi.org/10.2478/jos-2013-0026.

Buelens, B., Burger, J., and van den Brakel, J.A. (2018). Comparing inference methods for non-probability samples. *International Statistical Review* 86 (2): 322–343. https://doi.org/10.1111/insr.12253.

Burgette, L.F., Escarce, J.J., Paddock, S.M. et al. (2019). Sample selection in the face of design constraints: use of clustering to define sample strata for qualitative research. *Health Services Research* 54 (2): 509–517. https://doi.org/10.1111/1475-6773.13100.

Burke, A., Edwards, B., and Yan, T. (2018). The future is now: how surveys can harness social media to address 21st-century challenges. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/167/178/BigSurv_PPT_Burke_Garcia_Edwards_Yan_FINAL_FOR_DISSEMINATION_181025.pptx (accessed 5 July 2019).

Buskirk, T.D., Bear, T., and Bareham, J. (2018a). Machine made sampling designs: applying machine learning methods for generating stratified sampling designs. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.bigsurv18.org/conf18/uploads/177/203/BuskirkBearBarehamBigSurv18ProgramPostedVersionFinal.pdf (accessed 20 June 2019).

Buskirk, T.D. and Dutwin, D. (2016). Matchmaker, data scientist or both? Using unsupervised learning methods for matching nonprobability samples to probability samples. Paper presented at the 2016 Joint Statistical Meetings, Chicago, IL (30 July–4 August 2016).

Buskirk, T.D., Kirchner, A., Eck, A. et al. (2018b). An introduction to machine learning methods for survey researchers. *Survey Practice* 11 (1): 1–10. https://doi.org/10.29115/SP-2018-0004.

Buskirk, T.D. and Kolenikov, S. (2015). Finding respondents in the forest: a comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach*. http://surveyinsights.org/?p-5108. https://doi.org/10.13094/SMIF-2015-00003 (accessed 2 January 2019).

Buskirk, T.D., West, B.T., and Burks, A.T. (2013). Respondents: who art thou? Comparing internal, temporal, and external validity of survey response propensity models based on random forests and logistic regression models. Paper presented at the 2013 Joint Statistical Meetings, Montreal, Canada (3–8 August 2013).

Caiola, G. and Reiter, J. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy* 3: 27–42.

Can, S., Engel, U., and Keck, J. (2018). Topic modeling and status classification using data from surveys and social networks. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018).

Chew, R. and Biemer, P. (2018). Machine learning in adaptive survey designs. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/165/21/BigSurv18_Bandits.pptx (accessed 5 July 2019).

Chew, R.F., Amer, S., Jones, K. et al. (2018). Residential scene classification for gridded population sampling in developing countries using deep convolutional neural networks on satellite imagery. *International Journal of Health Geographics* 17 (1): 12. https://doi.org/10.1186/s12942-018-0132-1.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg, Germany: Springer Science & Business Media https://doi.org/10.1007/978-3-642-31164-2.

Cochran, W.G. (1964). *Sampling Techniques*, 2e. New York, NY: Wiley.

Conn, D., Ngun, T., Li, G. et al. (2015). *Fuzzy Forests: Extending Random Forests for Correlated, High-Dimensional Data*. Technical Research Report; eScholarship. Los Angeles, CA: University of California http://escholarship.org/uc/item/55h4h0w7.

Connelly, R., Playford, C.J., Gayle, V. et al. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research* 59: 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015.

Conrad, F.G., Couper, M.P., and Sakshaug, J.W. (2016). Classifying open-ended reports: factors affecting the reliability of occupation codes. *Journal of Official Statistics* 32 (1): 75–92. https://doi.org/10.1515/jos-2016-0003.

Couper, M.P. and Zhang, C. (2016). Helping respondents provide good answers in web surveys. *Survey Research Methods* 10 (1): 49–64. https://doi.org/10.18148/srm/2016.v10i1.6273.

Creecy, R.H., Masand, B.M., Smith, S.J. et al. (1992). Trading MIPS and memory for knowledge engineering. *Communications of the ACM* 35 (8): 48–64. https://doi.org/10.1145/135226.135228.

De Haas, S. and Winker, P. (2016). Detecting fraudulent interviewers by improved clustering methods – the case of falsifications of answers to parts of a questionnaire. *Journal of Official Statistics* 32 (3): 643–660. https://doi.org/10.1515/jos-2016-0033.

Drechsler, J. and Reiter, J. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis* 55 (12): 3232–3243. https://doi.org/10.1016/j.csda.2011.06.006.

Dutwin, D. and Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious?: Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly* 81 (S1): 213–239. https://doi.org/10.1093/poq/nfw061.

Early, K. (2017). Dynamic question ordering: obtaining useful information while reducing user burden. Doctoral thesis. Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA.

Earp, M., Mitchell, M., McCarthy, J., and Kreuter, F. (2014). Modeling nonresponse in establishment surveys: using and ensemble tree model to create nonresponse propensity scores and detect potential bias in an agricultural survey. *Journal of Official Statistics* 30 (4): 701–719.

Earp, M. S. and McCarthy, J. S. (2009). Using respondent prediction models to improve efficiency of incentive allocation. United States Department of Agriculture, National Agricultural Statistics Service. http://purl.umn.edu/235087 (accessed 3 July 2019).

Eck, A. (2018). Neural networks for survey researchers. *Survey Practice* 11 (1): 1–11. https://doi.org/10.29115/SP-2018-0002.

Eck, A., Buskirk, T. D., Fletcher, K., et al. (2018). Machine made sampling frames: creating sampling frames of windmills and other non-traditional sampling units using machine learning with neural networks. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018).

Eck, A., Buskirk, T. D., Shao, H., et al. (2019). Taking the machines to class: exploring how to train machine learning methods to understand images for survey sampling. Paper presented at the 74th Annual American Association of Public Opinion Research Conference, Toronto, Canada (18–21 May 2017).

Eck, A. and Soh, L.-K. (2017a). Sequential prediction of respondent behaviors leading to error in web-based surveys. Paper presented at the 72nd Annual Conference of the American Association for Public Opinion Research, New Orleans, LA (18–21 May 2017).

Eck, A. and Soh, L.-K. (2017b). Can an automated agent anticipate straightlining answers? Paper presented at the 72nd Annual Conference of the American Association for Public Opinion Research, New Orleans, LA (18–21 May 2017).

Eck, A., Soh, L.-K., and McCutcheon, A. L. (2015). Predicting breakoff using sequential machine learning methods. Paper presented at the 70th Annual Conference of the American Association for Public Opinion Research, Hollywood, FL (18–21 May 2017).

Eckman, S., Chew, R., Sanders, H., et al. (2019). Using passive data to improve survey data. Paper presented at the MASS Workshop, Mannheim, Germany (4–5 March 2019).

Elliott, M. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science* 32 (2): 249–264.

Elmagarmid, A.K., Ipeirotis, P.G., and Verykios, V.S. (2007). Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering* 19 (1): 1–16. https://doi.org/10.1109/TKDE.2007.250581.

Erdman, C. and Bates, N. (2017). The Low Response Score (LRS). A metric to locate, predict, and manage hard-to-survey populations. *Public Opinion Quarterly* 81 (1): 144–156. https://doi.org/10.1093/poq/nfw040.

Fischer, M. and Mittereder, F. (2018). Can missing patterns in covariates improve imputation for missing data? Paper Presented at the BigSurv18 Conference, Barcelona, Spain (25−27 October 2018). https://www.bigsurv18.org/conf18/uploads/42/64/MF_FM_MissInd_BigSurv18.pdf (accessed 1 June 2019).

Garber, S.C. (2009). *Census Mail List Trimming Using SAS Data Mining* (RDD Report 09-02). Fairfax, VA: Department of Agriculture, National Agricultural Statistics Service.

Garcia and Rueba. (2018). Efficiency of classification algorithms as an alternative to logistic regression in propensity score adjustment for survey weighting. Paper Presented at the BigSurv18 Conference; Barcelona, Spain (25−27 October 2018). https://www.bigsurv18.org/conf18/uploads/132/117/ExpBigSurv18.pdf (accessed 1 June 2019).

Graesser, A.C., Cai, Z., Louwerse, M.M. et al. (2006). Question Understanding Aid (QUAID) a web facility that tests question comprehensibility. *Public Opinion Quarterly* 70 (1): 3–22. https://doi.org/10.1093/poq/nfj012.

Graesser, A.C., Wiemer-Hastings, K., Kreuz, R. et al. (2000). QUAID: a questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments, & Computers* 32 (2): 254–262. https://doi.org/10.3758/BF03207792.

Grimmer, J. (2015). We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics* 48 (1): 80–83. https://doi.org/10.1017/S1049096514001784.

Guzman, D. (2018). Mining interviewer observations. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/139/174/bigSurv18_daniel_guzman.pptx (accessed 5 July 2019).

Hardeman, W., Houghton, J., Lane, K. et al. (2019). A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity* 16 (1): 31. https://doi.org/10.1186/s12966-019-0792-7.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY: Springer https://doi.org/10.1007/978-0-387-21606-5.

Huang, Z. (1998). Extensions to the K-modes algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* 2 (3): 283–304. https://doi.org/10.1023/A:1009769707641.

Huang, Z. and Ng, M.K. (1999). A fuzzy K-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems* 7 (4): 446–452. https://doi.org/10.1109/91.784206.

Jäckle, A., Burton, J., Couper, M.P. et al. (2019). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases. *Survey Research Methods* 13 (1): 23–44. https://doi.org/10.18148/srm/2019.v13i1.7297.

James, G., Witten, D., Hastie, T. et al. (2013). *An Introduction to Statistical Learning*, vol. 112. New York, NY: Springer https://doi.org/10.1007/978-1-4614-7138-7.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics* 19 (2): 81–97.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C* 29 (2): 119–127.

Kelly, F. and Doriot, P. (2017). Using data mining and machine learning to shorten and improve surveys. [Article ID: 20170526-1]. *Quirk's Media*.

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods* 13 (1): 73–93. https://doi.org/10.18148/srm/2019.v13i1.7395.

Kim, D.W., Lee, K.H., and Lee, D. (2004). Fuzzy clustering of categorical data using fuzzy centroids. *Pattern Recognition Letters* 25 (11): 1263–1271. https://doi.org/10.1016/j.patrec.2004.04.004.

Kolb, J.-P., Kern, C., and Weiß, B. (2018). Using predictive modeling to identify panel dropouts. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018).

Korpusik, M., Huang, C., Price, M., et al. (2016). Distributional semantics for understanding spoken meal descriptions. *IEEE SigPort*. http://sigport.org/676 (accessed 14 April 2019).

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 1: 1097–1105.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*, vol. 26. New York, NY: Springer https://doi.org/10.1007/978-1-4614-6849-3.

Lee, B.K., Lessler, J., and Stuart, E.A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* 29 (3): 337–346. https://doi.org/10.1002/sim.3782.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer web panel surveys. *Journal of Official Statistics* 22 (2): 329–349.

Liu, M. and Wang, Y. (2018). Using machine learning models to predict follow-up survey participation in a panel study. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/93/87/BigSurv_presentation_Liu_upload.pdf (accessed 18 June 2019).

Loh, W.-Y., Eltinge, J., Cho, M.J. et al. (2019). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica* 29 (1): 431–453. https://doi.org/10.5705/ss.202017.0225.

Lohr, S., Hsu, V., and Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. In: *Proceedings of the Joint Statistical Meetings, 2015 Survey Research Methods Section: American Statistical Association*, 2071–2085. Alexandria, VA: American Statistical Association (8–13 August).

Lu, C., Li, X.-W., and Pan, H.-B. (2007). Application of SVM and fuzzy set theory for classifying with incomplete survey data. In: *2007 International Conference on Service Systems and Service Management*, Chengdu, China, 1–4. https://doi.org/10.1109/ICSSSM.2007.4280164.

Lugtig, P. and Blom, A. (2018). Advances in modelling attrition: the added value of paradata and machine learning algorithms. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/192/47/Lugtig_and_Blom_big_surv_paradata_and_attrition_machine_learning.pdf (accessed 18 June 2019).

McCarthy, J.S. and Earp, M.S. (2009). *Who makes mistakes? Using data mining techniques to analyze reporting errors in total acres operated* (RDD Research Report Number RDD-09-02). Washington, DC: U.S. Departmente of Agriculture https://pdfs.semanticscholar.org/dee9/cf6187b6a77a1e8e1f64a76a4942dc2e0f6d.pdf?_ga=2.139828851.1790488900.1562077670-74454473.1562077670.

McConville, K.S., Breidt, F.J., Lee, T.C. et al. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology* 5 (2): 131–158. https://doi.org/10.1093/jssam/smw041.

McConville, K.S. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics* 46 (2): 389–413. https://doi.org/10.1111/sjos.12356.

McCreanor, R., Wronski, L., and Chen, J. (2018). Automated topic modeling for trend analysis of open-ended response data. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.bigsurv18.org/program2018?sess=38#286 (accessed 18 June 2019).

Megra, M., Medway, R., Jackson, M., et al. (2018). Predicting response mode preferences of survey respondents: a comparison between traditional regression

and data mining methods. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018).

Mendez, G., Buskirk, T.D., Lohr, S. et al. (2008). Factors associated with persistence in science and engineering majors: an exploratory study using random forests. *Journal of Engineering Education* 97 (1): 57–70. https://doi.org/10.1002/j.2168-9830.2008.tb00954.x.

Mercer, A.W., Kreuter, F., Keeter, S., and Stuart, E.A. (2017). Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opinion Quarterly* 81 (S1): 250–271.

Mercer, A., Lau, A. and Kennedy, C. (2018) For weighting online opt-in samples, what matters most? Pew Research Center Report released on January 26, 2018. https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/ (accessed 2 June 2019).

Mercer, A., Lau, A., and Kennedy, C. (2018). For weighting online opt-in samples, what matters most? Pew Research Center. https://www.ewresearch.org/methods/2018/01/26/for‐weighting‐online‐opt‐in‐samples‐what‐matters‐most/ (accessed 30 June 2019).

Mittereder, F. and West, B. (2018). Can response behavior predict breakoff in web surveys? Paper presented at the General Online Research Conference, Cologne, Germany. https://www.gor.de/gor18/index.php?page=downloadPaper&filename=Mittereder-Can_Response_Behavior_Predict_Breakoff_in_Web_Surveys-135.pdf&form_id=135&form_version=final (accessed 3 July 2019).

Mittereder, F. K. (2019). Predicting and preventing breakoffs in web surveys. Doctoral dissertation. Survey Methodology, University of Michigan, Ann Arbor, MI.

Morris, D.S. (2017). A modeling approach for administrative record enumeration in the decennial census. *Public Opinion Quarterly* 81 (S1): 357–384. https://doi.org/10.1093/poq/nfw059.

Morrison, L.G., Hargood, C., Pejovic, V. et al. (2017). The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial. *PLoS One* 12 (1): e0169162. https://doi.org/10.1371/journal.pone.0169162.

Mulder, J. and Kieruj, N. (2018). Preserving our precious respondents: predicting and preventing non-response and panel attrition by analyzing and modeling longitudinal survey and paradata using data science techniques. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018).

Nahum-Shani, I., Smith, S.N., Spring, B.J. et al. (2018). Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52 (6): 446–462. https://doi.org/10.1007/s12160-016-9830-8.

Niraula, N. B. and Rus, V. (2014). A machine learning approach to pronominal anaphora resolutionin dialogue based intelligent tutoring systems. *Proceedings of the Computational Linguistics and Intelligent Text Processing. Proceedings (Part 1) of the 15th International Conference CICLing*, Kathmandu, Nepal (pp. 307–318). Berlin, Heidelberg: Springer.

Oberski, D.L. (2016). Questionnaire science. In: *The Oxford Handbook of Polling and Survey Methods* (eds. L.R. Atkeson and R.M. Alvarez), 113–140. New York, NY: Oxford University Press https://doi.org/10.1093/oxfordhb/9780190213299.013.21.

Oberski, D. L. and DeCastellarnau, A. (2019). Predicting measurement error variance in social surveys. (Unpublished Manuscript).

Oberski, D.L., Gruner, T., and Saris, W.E. (2011). The prediction of the quality of the questionsbased on the present data base of questions. In: *The Development of the Program SQP 2.0 for the Prediction of the Quality of Survey Questions* (eds. W.E. Saris, D. Oberski, M. Revilla, et al.), 71–81. Spain: Barcelona.

Pankowska, P., Oberski, D., and Pavlopoulos, D. (2018). The effect of survey measurement error on clustering algorithms. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/219/239/clustering_measurement_error.ppt (accessed 29 June 2019).

Petrakos, G., Conversano, C., Farmakis, G. et al. (2004). New ways of specifying data edits. *Journal of the Royal Statistical Society. Series A (General)* 167 (2): 249–274. https://doi.org/10.1046/j.1467-985X.2003.00745.x.

Phipps, P. and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics* 6 (2): 772–794. https://doi.org/10.1214/11-AOAS521.

Provost, F. and Domingos, P. (2003). Tree induction for probability-based ranking. *Machine Learning* 52 (3): 199–215. https://doi.org/10.1023/A:1024099825458.

Roberts, M.E., Stewart, B.M., Tingley, D. et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science* 58 (4): 1064–1082. https://doi.org/10.1111/ajps.12103.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1): 41–55. https://doi.org/10.1093/biomet/70.1.41.

Ruiz, C. (2018). Improving data validation using machine learning. Paper presented at the Conference of European Statisticians, Workshop on Statistical Data Editing, Neuchatel, Switzerland (18–20 September 2018).

Saris, W.E. and Gallhofer, I.N. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods* 1 (1): 29–43. https://doi.org/10.1002/9780470165195.ch12.

Saris, W.E. and Gallhofer, I.N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: Wiley https://doi.org/10.1002/9781118634646.

Schembre, S.M., Liao, Y., Robertson, M.C. et al. (2018). Just-in-time feedback in diet and physical activity interventions: systematic review and practical design framework. *Journal of Medical Internet Research* 20 (3): e106. https://doi.org/10.2196/jmir.8701.

Schierholz, M. (2018). A comparison of automatic algorithms for occupation coding. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/56/82/Schierholz_Barcelona_26_10_2018.pdf (accessed 5 July 2019).

Schierholz, M., Gensicke, M., Tschersich, N. et al. (2018). Occupation coding during the interview. *Journal of the Royal Statistical Society. Series A (General)* 181 (2): 379–407. https://doi.org/10.1111/rssa.12297.

Schonlau, M. and Couper, M. (2017). Options for conducting web surveys. *Statistical Science* 32 (2): 279–292. https://doi.org/10.1214/16-STS597.

Schonlau, M. and Couper, M.P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods* 10 (2): 143–152. https://doi.org/10.18148/srm/2016.v1012.6213.

Schonlau, M., Gweon, H., and Wenemark, M. (2018). Classification of open-ended questions with multiple labels. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/173/39/all_that_apply_barcelona.pdf (accessed 5 July 2019).

Schouten, B. and de Nooij, G. (2005). *Nonresponse Adjustment Using Classification Trees*. The Netherlands: CBS, Statistics Netherlands.

Sela, R.J. and Simonoff, J.S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning* 86 (2): 169–207. https://doi.org/10.1007/s10994-011-5258-3.

Shmueli, G. (2010). To explain or to predict? *Statistical Science* 25 (3): 289–310. https://doi.org/10.1214/10-STS330.

Shorey, J., Jang, H., Spagnardi, C., et al. (2018). Enriching surveys with knowledge graphs. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018).

Spruijt-Metz, D. and Nilsen, W. (2014). Dynamic models of behavior for just-in-time adaptive interventions. *IEEE Pervasive Computing* 13 (3): 13–17. https://doi.org/10.1109/MPRV.2014.46.

Stanfill, M.H., Williams, M., Fenton, S.H. et al. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association* 17 (6): 646–651. https://doi.org/10.1136/jamia.2009.001024.

Stone, A.A. and Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine* 16 (3): 199–202. https://doi.org/10.1093/abm/16.3.199.

Stone, A.A., Shiffman, S., Atienza, A.A. et al. (2007). *The Science of Real-Time Data Capture: Self-Reports in Health Research*. New York, NY: Oxford University Press.

Strobl, C., Boulesteix, A.-L., Zeileis, A. et al. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8 (25): 25. https://doi.org/10.1186/1471-2105-8-25.

Sun, H., Rivero, G., and Yan, T. (2019). Identifying interviewer falsification using speech recognition: a proof of concept study. Paper presented at the 74th Annual Conference of the American Association for Public Opinion Research, Toronto, Canada (16–19 May 2019).

Sweitzer, M.D. and Shulman, H.C. (2018). The effects of metacognition in survey research experimental, cross-sectional, and content-analytic evidence. *Public Opinion Quarterly* 82 (4): 745–768. https://doi.org/10.1093/poq/nfy034.

ten Bosch, O., Windmeijer, D., van Delden, A., et al. (2018). Web scraping meets survey design: combining forces. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.bigsurv18.org/conf18/uploads/73/61/20180820_BigSurv_WebscrapingMeetsSurveyDesign.pdf (accessed 30 June 2019).

Thaung, A. and Kolenikov, S. (2018). Detecting and comparing survey research topics in conference and journal abstracts. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018).

Thomson, D.R., Stevens, F.R., Ruktanonchai, N.W. et al. (2017). GridSample: an R package to generate household survey primary sampling units (PSUs) from gridded population data. *International Journal of Health Geographics* 16 (1): 25. https://doi.org/10.1186/s12942-017-0098-4.

Timbrook, J. and Eck, A. (2018). Comparing coding of interviewer question-asking behaviors using recurrent neural networks to human coders. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/158/193/Timbrook_Eck_BigSurv_2018.pdf (accessed 5 July 2019).

Timbrook, J. and Eck, A. (2019). Humans vs. machines: comparing coding of interviewer question-asking behaviors using recurrent neural networks to human coders. Paper presented at the Interviewers and Their Effects from a Total Survey Error Perspective Workshop, University of Nebraska-Lincoln (26–28 February 2019). https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1025&context=sociw (accessed 5 July 2019).

Toth, D. (2017). rpms: Recursive partitioning for modeling survey data. R Package Version 0.2.0. https://CRAN.R-project.org/package=rpms (accessed 20 October 2017).

Toth, D. and Eltinge, J.L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association* 106 (496): 1626–1636. https://doi.org/10.1198/jasa.2011.tm10383.

Toth, D. and Phipps, P. (2014). Regression tree models for analyzing survey response. In: *Proceedings of the Government Statistics Section*, 339–351. American Statistical Association.

Valliant, R., Dever, J., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York, NY: Springer https://doi.org/10.1007/978-1-4614-6449-5.

Valliant, R. and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research* 40 (1): 105–137. https://doi.org/10.1177/0049124110392533.

Vanhoof, M., Lee, C., and Smoreda, Z. (2018). Performance and sensitivities of home detection from mobile phone data. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/135/169/BigSurv_MainText_Performance_and_sensitivities_of_home_detection_from_mobile_phone_data.pdf (accessed 30 June 2019).

Ventura, S.L., Nugent, R., and Fuchs, E.R.H. (2015). Seeing the non-stars: (some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy* 44 (9): 1672–1701. https://doi.org/10.1016/j.respol.2014.12.010.

Ward, C. and Reimer, B. (2019). Using natural language processing to enhance prediction of panel attrition in a longitudinal survey. Paper presented at the 74th Annual Conference of the American Association for Public Opinion Research, Toronto, Canada (16–19 May 2019).

Winkler, W.E. (2006). *Overview of Record Linkage and Current Research Directions* (Technical Report Statistical Research Report Series RRS2006/02). Washington, DC: U.S. Bureau of the Census.

Ye, C., Medway, R., and Kelley, C. (2018). Natural language processing for open-ended survey questions. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018).

Zhao, P., Su, X., Ge, T. et al. (2016). Propensity score and proximity matching using random forest. *Contemporary Clinical Trials* 47: 85–92. https://doi.org/10.1016/j.cct.2015.12.012.

Further Reading

Buskirk, T.D. (2018). Surveying the forests and sampling the trees: an overview of classification and regression trees and random forests with applications in survey research. *Survey Practice* 11 (1): 1–13. https://doi.org/10.29115/SP-2018-0003.

Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 25 (2): 295–313.

Dutwin, D. and Buskirk, T. D. (2017). A deep dive into the digital divide. Paper presented at the 72nd annual American Association of Public Opinion Research Conference, New Orleans, LA (18–21 May 2017).

Ferri-García, R. and del Mar Rueda, M. (2018). Efficiency of classification algorithms as an alternative to logistic regression in propensity score adjustment for survey weighting. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/132/117/ExpBigSurv18.pdf (accessed 28 June 2019).

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27 (4): 857–874. https://doi.org/10.2307/2528823.

Gray, J. (2009). On eScience: a transformed scientific method. In: *The Fourth Paradigm* (eds. T. Hey, S. Tansley and K. Tolle), xvii–xxxi. Redmond, WA: Microsoft Research.

Han, B. and Baldwin, T. (2011). Lexical normalization of short text messages: makn sens a #twitter. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland OR (19–24 June), 368–378.

Japec, L., Kreuter, F., Berg, M. et al. (2015). Big data in survey research: Aapor task force report. *Public Opinion Quarterly* 79 (4): 839–880. https://doi.org/10.1093/poq/nfv039.

Khoshrou, A., Aguiar, A.P., and Pereira, F.L. (2016). Adaptive sampling using an unsupervised learning of GMMs applied to a fleet of AUVs with CTD measurements. In: *Robot 2015: Second Iberian Robotics Conference. Advances in Intelligent Systems and Computing* (eds. L. Reis, A. Moreira, P. Lima, et al.), 321–332. Cham: Springer https://doi.org/10.1007/978-3-319-27149-1.

Kim, J.K. and Wang, Z. (2019). Sampling techniques for big data analysis in finite population inference. *International Statistical Review* S1: S177–S191. https://doi.org/10.1111/insr.12290.

Kirchner, A. and Signorino, C.S. (2018). Using support vector machines for survey research. *Survey Practice* 11 (1): 1–14. https://doi.org/10.29115/SP-2018-0001.

Krantz, A., Korn, R., and Menninger, M. (2009). Rethinking museum visitors: using K-means cluster analysis to explore a museum's audience. *Curator* 52 (4): 363–374. https://doi.org/10.1111/j.2151-6952.2009.tb00358.x.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research* 37 (3): 319–343. https://doi.org/10.1177/0049124108329643.

Linden, A. and Yarnold, P.R. (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice* 23 (4): 703–712. https://doi.org/10.1111/jep.12744.

McCaffrey, D.F., Griffin, B.A., Almirall, D. et al. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* 32 (19): 3388–3414. https://doi.org/10.1002/sim.5753.

Mercer, A.W., Kreuter, F., Keeter, S. et al. (2017). Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opinion Quarterly* 81 (S1): 250–271. https://doi.org/10.1093/poq/nfw060.

Pirracchio, R., Petersen, M.L., and van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology* 181 (2): 108–119. https://doi.org/10.1093/aje/kwu253.

Signorino, C.S. and Kirchner, A. (2018). Using LASSO to model interactions and nonlinearities in survey data. *Survey Practice* 11 (1): 1–10. https://doi.org/10.29115/SP-2018-0005.

Timpone, R., Kroening, J., and Yang, Y. (2018). From big data to big analytics: automated analytic platforms for data exploration. Paper presented at the BigSurv18 Conference, Barcelona, Spain (25–27 October 2018). https://www.europeansurveyresearch.org/bigsurv18/uploads/307/367/From_Big_Data_to_Big_Analytics_102618_paper.pdf (accessed 30 June 2019).