Too good to be true?

Machine learning in the editing process

Eva Charlotte Berner Solveig Bjørkholt

April 29, 2023

This version: April 29, 2023

Measurement error in survey data poses a critical risk to inferences. As such, many survey organizations work hard to minimize the survey error. Processes of quality assurance do, however, often involve an editing process that is partly manual, rendering the process time consuming and costly. On the flip side, the process also produces high-quality training data which machine learning models can utilize to make predictions about survey responses. In this paper, we use the Research and Development survey from Statistics Norway to show how machine learning can improve efficiency, transparency and transferability in the editing process, freeing up time that can be used to also improve overall quality of the survey data.[1]

Keywords— Survey error, Measurement error, Validation, Imputation, Editing, Machine learning, NSO

[1] Github: https://github.com/statisticsnorway/fou-editering/tree/main

1 Introduction

Survey data is prone to measurement error. A survey that fails to capture the true value of a variable has by definition induced error somewhere in the process (Asher 1974). These errors can have important consequences for research and policy, but finding and correcting them can be severely time-consuming. Because of this, finding efficient methods to ensure survey data quality is vital.

Machine learning methods can be used to improve upon efficiency, and they have become increasingly available and powerful. Some studies have shown how machine learning models can be used for data cleaning tasks such as outlier detection and text classification (Dai, Yoshigoe and Parsley 2018; Hoque et al. 2018; Kolluri, Razia and Nayak 2020). These studies are particularly relevant for large survey organizations that mandate and regularly work on data collection and distribution. Many of these are public organizations, where machine learning has been slow to enter the work processes although the potential is significant (OECD 2019; Wirtz, Weyerer and Geyer 2019).

To illustrate how machine learning can aid in the quality assurance of survey data, we case-point the survey on Research and Development administered by Statistics Norway. As machine learning requires well-crafted training data, institutionalized surveys that run over several years are particularly well suited for this method. Many organizations provide valuable time-series data collected from surveys. Eurostat offers the Harmonized European Time Use Surveys, ILO has the Labour Force Survey, and then there are national and international research programs such as World Values Survey. National statistics offices (NSO) such as Statistics Norway also frequently use surveys to collect data and publish national statistics. The benefits of using machine learning in these cases are two-fold. While the repeated and institutionalized processes of quality assurance within the organization produce good training data, this repeated process also makes efficient quality assurance particularly useful.

In this paper, we proceed by introducing the well-known concept of "survey measurement error", ground it in the larger "total survey error" framework, and argue why it is important to account for. Then, we describe how many large survey organizations account for measurement error through an editing process, and why machine learning is a particularly useful in this process. We proceed by explaining the case – the R&D survey in Statistics Norway – and describe how machine learning could be used to ease this survey's editing process. Last, we offer preliminary findings for our analysis and some thoughts on the way forward.

2 Survey measurement errors

Survey error is defined as a deviation of a survey response from its underlying true value (Biemer 2010). It can incur in any stage of the survey process, from the design, to the collection, to the processing. Survey designers have been aware of the potential for error for decades, and many studies have gone into conceptualizing and categorizing different types of error (Biemer and Lyberg 2003; Groves and Lyberg 2010). For example, in 1944, Deming identified thirteen potential sources of error, among them being non-response, sampling and interviewer effects. Subsequent research has identified many more, and today, the concept of "total survey error" includes many categories of errors.

Groves and Lyberg (2010) organize the total survey error into two broad components based on their inferential properties. The measurement component refers to errors occurring in the inferential step one makes for an individual survey response, including anything from failing to capture the underlying construct to add deviating processing steps, such as misconstructed weights. The representation component refers to errors occurring in the inferential step from a sample to a population, including coverage, sampling errors and non-response. In sum, total survey error refers to all the sources of errors that contaminate a survey's ability to find the *true* value in its population. This framework is illustrared in figure 1.

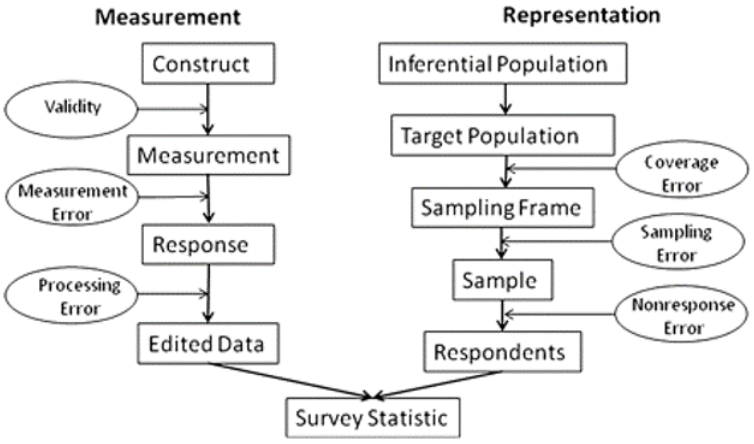


Figure 1: Total Survey Error Components (grovesTotalSurveyError2010, adapted from Groves et al. 2004).

Our method is specifically focused on measurement errors. Measurement errors is an extensively studied type of error that plagues most surveys to some degree. This error is caused by interviewers, respondents,

questions or overall survey design. It could also stem from contextual factors such as respondents' insufficient information systems, the platform where the survey is conducted, or the degree of sensitivity on the topics (Biemer 2010).

Measurement errors have important consequences for the research conducted using the data, and this research is substantial. Between 1994 and 1995, approximately 30 percent of the articles published in the *American Journal of Political Science*, *The American Political Science Review* and *the Journal of Politics* used surveys as data. The number was even higher for research within economy, sociology and psychology, and for articles published in *Public Opinion Quarterly*, as much as 95 percent of the research reported use of surveys. The number does not seem to have diminished by 2011 (Saris and Revilla 2016). Many of these studies used surveys administered by NSOs (Saris and Gallhofer 2014). At the same time, even though most researchers acknowledge that measurement errors exist, they are seldom corrected (Saris and Revilla 2016). Some researchers assume that measurement errors are small, and thus pose a negligible threat to inferences, but this is a bold assumption to make given the numerous ways measurement errors can arise.

The consequence of measurement errors in research range from inflated variance to downright bias. In practice, measurement error can lead to diminished strength of correlations, masking of real effects, false correlations and even reversed signs of correlations (Bound, Brown and Mathiowetz 2001). Both type I and type II errors are possible. To the extent that decision-makers rely on statistical data, measurement error can also have real consequences for policy making.

3 The editing process and machine learning

Despite the important effects of measurement error, they too often remain unaccounted for. One reason is that they can be hard to track. Survey designers can uncover error that cause high variance by running the same survey several times, and they can discover bias by for example changing how questions are framed and calculate differences. Furthermore, methods such as Structural Equation Models (SEM) and multitrait-multimethod (MTMM) can be used to estimate the quality of questions (Saris and Revilla 2016). However, these methods require more questions per survey, thus increasing the costs and respondent burden.

In the wake of these challenges, NSOs have implemented numerous means to assure data quality. Many of these occur in the design phase, but a substantial part of the quality assurance methods occur after the data has been collected, in the processing stage. Quality assurance after data collection is called "data

editing". The Generic Statistical Data Editing Model (GSDEM) offers guidelines on the conduction of data editing (Statswiki 2019). It divides the editing process into functions and methods, where functions relate to any controls made by statistics producers to investigate potential errors, and methods are the specific routines used for error correction. In this setting, it is useful to distinguish between data validation and imputation. While data validation involves functions and methods used to find problematic units or variables, imputation involves correcting a problematic value. For example, finding outliers and investigating whether their deviating value is an error is part of the validation procedure, while correcting errors and changing the value is an imputation. Both of these procedures are part of data editing, which again is a part of the larger quality assurance framework. The relationship between these concepts are illustrated in figure 2.

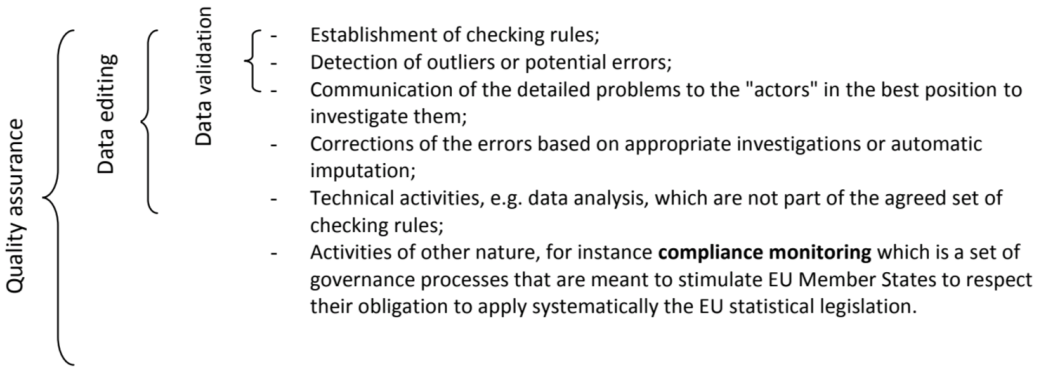


Figure 2: The relationship between validation, imputation, editing and quality assurance, from Di Zio et al. (2018).

Ideally, rule-based or data-driven methods exist to account for most editing functions. These could for example be logical controls that alerts when there are impossible inconsistencies, for example more female employees than there are total employees, or automated duplicate checks. However, editing of large survey data is a comprehensive process, and while some editing tasks can be captured in rules, some tasks are too complex to create rules for. Therefore, many statistic production processes include a lengthy stage of manual editing. This is what causes data editing to become so costly, and is the background for the introduction in the GSDEM manual: "The topic of data editing attracts considerable interest in the context of modernising official statistics, because it is traditionally one of the most expensive and time-consuming parts of the statistical production process and is prone to be influenced by innovative procedures like machine learning" (Statswiki 2019, p. 3).

In many NSOs, machine learning methods are becoming increasingly popular. The vast amount of high-quality statistical data combined with the ability to attract relevant competency make NSO's ripe for machine learning applications. Following a series of working packages, United Nations Economic Commission for Europe (UNECE) tested and outlined what potential machine learning holds for NSOs. Among them are textual classification, imagery analysis and – indeed – editing and imputation (Julien et al. 2022).

Machine learning is a particularly relevant method for data editing for at least three reasons. First, the processes of correcting survey measurement error through manual validation and imputation lead to an abundance of training data. This training data is often of high quality as the institutions providing the survey data has a high stake in assuring its quality. Second, the questions that are prone to manual editing are often too complex for rule-based methods. However, the training data contains embedded rules on what constitutes a measurement error, which machine learning algorithms can learn and apply to new data. Third, access to numerous data sources, as is often the case within survey institutions and which becomes even more relevant with the advent of big data, can be effectively utilized by machine learning algorithms. Thus, machine learning algorithms can be used to predict a value that would otherwise be manually corrected. This could lead to a semi-automation of the editing process even for complex tasks, making the process of minimizing measurement error more efficient.

The added advantages of using machine learning in the editing process go beyond efficiency. Machine learning would also improve reproducibility and transparency where tasks are otherwise too complex to be delineated into clear-cut rules. Moreover, survey quality is multifaceted. As outlined in figure 1, error affecting survey quality can incur in all stages of the survey process. In addition, other factors such as ethical considerations, timeliness (Weisberg 2009) and ensuring international comparison (Smith 2011) are also important. Thus, improving efficiency in addressing measurement errors would free up time to address other types of survey errors and considerations, thus increasing the overall survey quality.

4 The research and development (R&D) survey

To illustrate how machine learning can be used in the editing process, we employ the survey on research and development (R&D) in the Norwegian business sector[2], administered by Statistics Norway. This survey

[2]https://www.ssb.no/en/teknologi-og-innovasjon/forskning-og-innovasjon-i-naeringslivet/statistikk/forskning-og-utvikling-i-naeringslivet

measures the R&D activity of Norwegian enterprises, including their of R&D personnel, R&D expenditures, R&D acquisition and so forth, and it is used frequently by both policy makers and researchers. The survey runs annually and contains several complex questions. As such, its editing process is quite resource heavy. As shown in table 1, from 2011 to 2021, roughly 43 percent of all enterprises were on average validated. To estimate the cost of this, if validating one enterprise takes 10 minutes and 2000 enterprises are validated for an average survey, this constitutes about 330 hours of work, roughly 0,2 man-hours.

Table 1: Percentage (and total) of validated enterprises in the R&D survey, 2011 - 2021

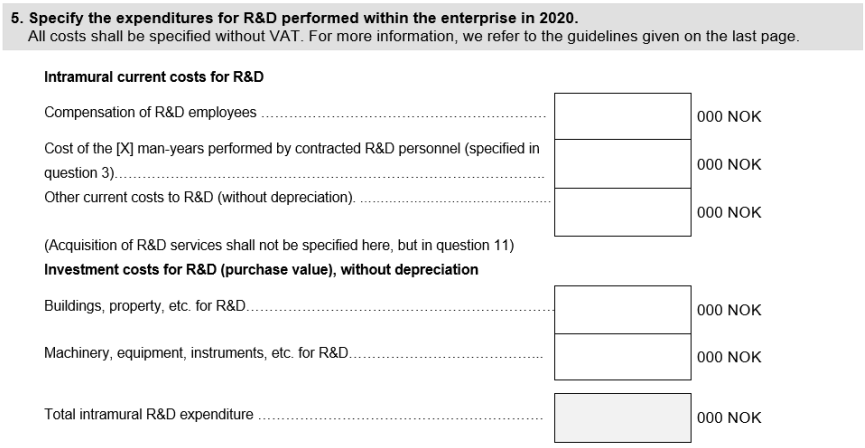
Validaton status	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Not validated	44% (2,141)	44% (2,741)	48% (1,963)	37% (1,546)	58% (3,174)	44% (2,125)	71% (4,225)	67% (3,559)	64% (3,971)	72% (3,754)	65% (4,119)
Validated	56% (2,719)	56% (3,468)	52% (2,148)	63% (2,667)	42% (2,276)	56% (2,709)	29% (1,736)	33% (1,718)	36% (2,234)	28% (1,466)	35% (2,223)

The sample covers large segments of the Norwegian business sector. All enterprises with more than 50 employees are included, as well as all enterprises that reported having R&D last year. A random sample is drawn for enterprises with 10-49 employees. Every other year, the study also includes a random sample of enterprises with 5-9 employees, also counting small enterprises that reported R&D in the last survey. Because the enterprises have a legal obligation to answer the survey, the response rate is close to 100 %. Because of this institutional asset, survey errors relating to representation in figure 1 are most likely small.

In addition to units, the number of questions vary as well. In odd years, the survey contains 15 questions and in even years, there are 10 questions. In this paper, we focus on one of the most important variables, *intfou*, which asks respondents to specify the expenditures to R&D performed within the enterprise in the given year. This question is asked every year. The *intfou* variable is the sum of (1) compensation to R&D employees, (2) cost to R&D contracted personnel, (3) other current costs, and R&D costs for (4) buildings, property, etc., and (5) machinery, equipment, instruments, etc. The question is illustrated in figure 3, where the *intfou* variable is the number given to "Total intramural R&D expenditure".

Survey designers have taken several steps to reduce survey error, for example adding controls that alert the respondent if they have reported inconsistent values. Yet, the complexity of the survey prevents extensive controls in the design phase. The "contact person" is responsible for filling out the survey for the enterprise, and self-reported numbers indicate that the contact person spends about 75 minutes on the survey, a number

that increases to almost three hours if the enterprise reported having R&D[3]. Meanwhile, around 10 percent of the contact persons that answered a follow-up question on survey quality, reported that they found the survey "hard" to complete[4].



5. Specify the expenditures for R&D performed within the enterprise in 2020.
All costs shall be specified without VAT. For more information, we refer to the guidelines given on the last page.

Intramural current costs for R&D

Compensation of R&D employees		000 NOK
Cost of the [X] man-years performed by contracted R&D personnel (specified in question 3)		000 NOK
Other current costs to R&D (without depreciation)		000 NOK

(Acquisition of R&D services shall not be specified here, but in question 11)

Investment costs for R&D (purchase value), without depreciation

Buildings, property, etc. for R&D		000 NOK
Machinery, equipment, instruments, etc. for R&D		000 NOK
Total intramural R&D expenditure		000 NOK

Figure 3: The question in the R&D survey that measures total intramural R&D expenditure.

With this in mind, measurement errors are far from unlikely. Some common causes of *intfou* measurement errors from the respondent side are adding three extra zeroes to the reported number ("thousand error"), lacking accurate information on the enterprises' R&D, adding daughter companies' R&D to the total, and adding purchased R&D together with intramural R&D. Overall, these measurement errors lead to inflated raw numbers on intramural R&D expenditures. Therefore, after validation and imputation, the processed number of R&D expenditures in Norwegian enterprises is much lower than the raw numbers. From 2011 to 2021, the statistic on R&D expenditures in Norwegian enterprises was about 15 percent of what was reported in the survey. There is, however, a lot of variation between the years, as shown in figure 4.

[3]This includes time spent finding the necessary information, time spent filling out the survey and time other people have spent on helping.

[4]Around 40 percent said they found it "both easy and hard", and 50 percent found it "easy". These numbers are from 2017 to 2021 where between 1000 and 1500 contact persons answered the follow-up question.

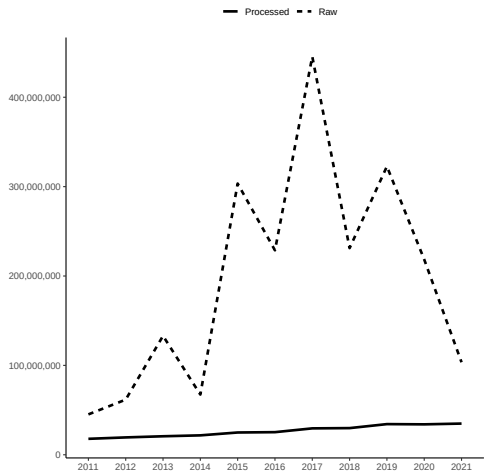


Figure 4: Difference between processed and raw data in the R&D survey, 2011-2021.

5 Predicting measurement errors on the total R&D expenditures (*intfou*)

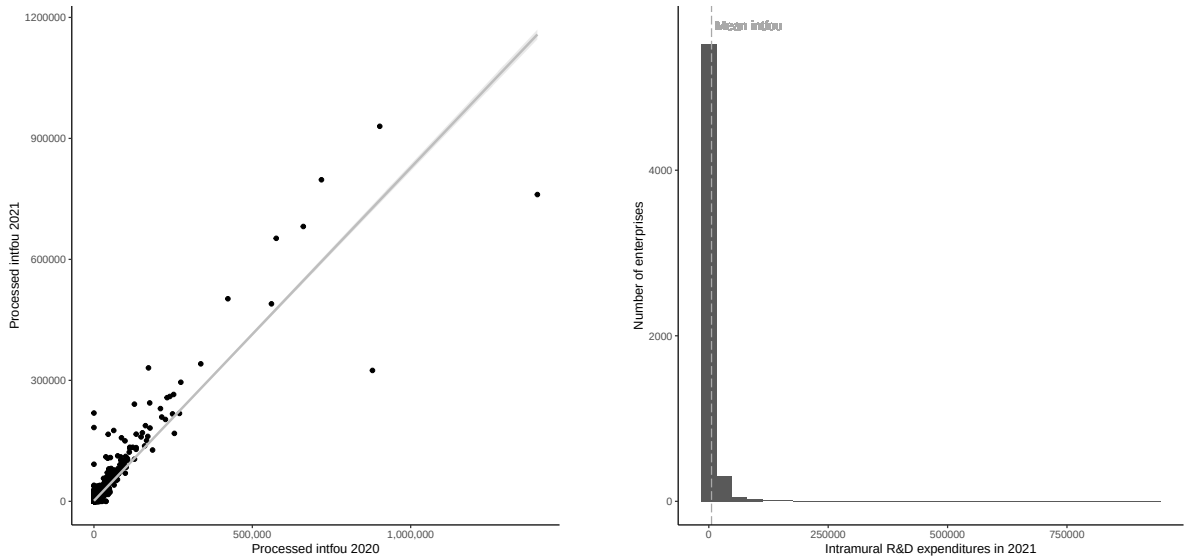
As mentioned above, machine learning allows the algorithm to learn latent patterns from training data. The training data, in this case, is the survey data after it has undergone validation and imputation. We refer to this data as "processed data". The data stemming directly from the survey, which contains measurement errors, is called "raw data". We use machine learning to predict the processed value of *intfou* for each enterprise given its previous processed values. The difference between the raw value and the processed value is the size of the measurement error.

Before we run the models, we perform data cleaning tasks that include merging and tidying datasets, identifying and removing duplicates, identifying and flagging outliers, and identifying missing values to either impute or remove them. After that, we split the data into training and test data. As test data, we use survey responses from 2021, and the training data constitutes data from 2015 to 2020.

The relationship between an enterprises' previous years' R&D expenditures and this years' R&D expenditures is highly linear, as shown in figure 5a, indicating that previous years of processed values should constitute good predictors for next year. The correlation between *intfou* in 2020 and *intfou* in 2021 is 0,92. The number drops to 0,89 for 2019 and 0,85 in 2018, but still remains substantially high[5]

Figure 5a also reveals a lump of enterprises with low amount of R&D expenditures. This reflects the skewedness of the outcome variable. As shown in figure 5b, the distribution is considerably right-skewed.

[5]Robustness checks indicates that the correlation is not due to multicollinearity.



(a) Processed *intfou* in 2020 and 2021 **(b)** Distribution of *intfou* variable in 2021

Figure 5

Models with all enterprises	**Explained variance**	**RMSE**
Linear	0.928	9664.27
Ridge	0.928	9723.31
Lasso	0.929	9651.92
Elastic Net	0.796	16296.93
Number of observations: 5957		
Models with only R&D-enterprises	**Explained variance**	**RMSE**
Linear	0.568	26452.50
Ridge	0.574	26271.46
Lasso	0.568	26465.62
Elastic Net	0.690	22400.24
Number of observations: 2174		

Table 2: Performance metrics for linear models with and without enterprises with zero R&D

While the skewedness is 17, the kurtosis is 392. Almost two thirds of the enterprises reported having 0 intramural R&D expenditures in 2021. On a scale that ranges to almost one billion NOK, the mean is 5700. This poses some challenges to the prediction models, and thus, in table 2, we show performance metrics for models both with and without enterprises that had 0 R&D expenditures in 2021.

Given the linear relationship between the predictors and the outcome variable, we run linear regression models, specifically a linear model, lasso, ridge and elastic net. The advantages of these models compared to more complex models such as gradient boosting and neural networks, is that they are simple to run and easier to interpret. Simple models are easier to adopt into work processes, and more interpretable models provide more explainable machine learning, which creates transparency and helps foster trust around machine-aided decision-making (Rudin 2019). Thus, we fit the following model for each enterprise i, where t is 2021:

$$intfou_{it} = \alpha + \beta_1 intfou_{it-1} + \beta_2 intfou_{it-2} + \beta_3 intfou_{it-3} + \beta_4 intfou_{it-4} + \beta_5 intfou_{it-5} + \beta_6 intfou_{it-6} + \varepsilon$$

The results in table 2 show that the models predict extremely well when all enterprises are included. Even the most simple linear model achieves an explained variance of almost 93 percent. As both the lasso and ridge regressions give comparable results, we can be more confident that the performance is not due to overfitting. The elastic net performs worst, as illustrated by both the lower explained variance and higher RMSE.

However, the performance is highly driven by the skewedness of the outcome variable. Having no R&D in the previous years is a strong predictor that the enterprise will not have R&D in the upcoming year. When we fit models only to the enterprises that did have R&D in 2021, they perform considerably worse. Explained variance for the linear model is down to 57 percent. At the same time, the elastic net regression does considerably better, explaining 69 percent of the variance and now posing the lowest RMSE of the four models.

6 The way forward

Historical data has shown good performance for predicting R&D for the full sample of enterprises, but struggles with samples that include only R&D performing enterprises. The current models are very simple, and there are several steps we can take to boost performance.

First, historical data is not equally predictive for non-R&D- and R&D-performing enterprises. As ma-

chine learning models can be built on a multitude of variables, we take advantage of the large amount of statistical data that exists in NSOs and ask whether these can be used to predict *intfou*. Having collected and gathered data from various sources, table 3 gives an overview of potential variables.

Second, more complex models could yield better results. Different model specifications might fit the data better, such as a function that better accounts for skewedness. An example could be a two-stage model where we first predict whether the enterprise will have R&D, and secondly the R&D expenditure size.

Third, more complex models might yield better accuracy on the R&D performing sample. Given the results from the elastic net regression, this seems to be a promising avenue. More comprehensive model tuning might also help improve performance.

7 Conclusion

Accounting for survey error is vital to trust inferences made with the data, whether they are research or policy relevant. For large organizations that regularly employ surveys, such as NSOs, finding efficient methods to reduce measurement errors is particularly important. While it is possible to account for survey error in the design phase of the survey, not all error can be removed this way. Especially for complex questions, the data has to be quality checked after collection. This process, known as "editing", takes up a lot of time for many survey organizations. In this paper, we show how the existence of large amounts of training data due to diligent editing can be used to train machine learning algorithms.

Machine learning can be used to semi-automate the editing process. This will not only improve efficiency, but also reproducibility and transferability of the rules behind the editing process. Moreover, freeing up time from editing of measurement errors, statistics personnel can devote more attention to other types of errors, such as validity and processing errors.

The latter is particularly important, since, ironically, the process of correcting measurement errors might in turn introduce processing errors. While we can gauge the size of the measurement error of new survey data using machine learning methods, the processing errors are hidden to us. Since they are embedded in the training data, the machine learning model will reproduce them in any application. Thus, time freed from correcting measurement errors could be spent correcting processing errors, which would lead to an overall better quality on the survey data.

TYPE	VARIABLE
Historical	(1) Y-variable in t-1, t-2, t-3, etc.
Accounting	(1) operating income (2) operating costs (3) salary (4) R&D as reported in the income statement
Structural	(1) number of employees (2) revenue (3) being part of an enterprise group (4) industry category
Employees	(1) employees' education level (2) employees' work category
Financial support	(1) support received from SkatteFUNN (2) support received from Norges Forskningsråd (3) support received from Innovasjon Norge
Respondent	(1) whether the respondent reported difficulties in the survey (2) whether there was a change in contact person from t-1 (3) reported time spent on collecting information (4) reported time spent on filling out the survey
Editor	(1) validation of enterprise (any variable) in t-1, t-2, t-3, etc. (2) who validated the enterprise

Table 3: Various variables that can be included in the prediction models

References

Asher, Herbert B. 1974. "Some Consequences of Measurement Error in Survey Data." *American Journal of Political Science* 18(2):469–485.

Biemer, Paul P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74(5):817–848.

Biemer, Paul P. and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. John Wiley & Sons.

Bound, John, Charles Brown and Nancy Mathiowetz. 2001. Chapter 59 - Measurement Error in Survey Data. In *Handbook of Econometrics*, ed. James J. Heckman and Edward Leamer. Vol. 5 Elsevier pp. 3705–3843.

Dai, Wei, Kenji Yoshigoe and William Parsley. 2018. Improving Data Quality Through Deep Learning and Statistical Models. In *Information Technology - New Generations*, ed. Shahram Latifi. Vol. 558 Cham: Springer International Publishing pp. 515–522.

Deming, W Edwards. 1944. "On errors in surveys." *American Sociological Review* 9(4):359–369.

Di Zio, Marco, Nadežda Fursova, Tjalling Gelsema, Sarah Gießing, Ugo Guarnera, Jūratė Petrauskienė, L Quensel-von Kalben, Mauro Scanu, KO ten Bosch, Mark van der Loo et al. 2018. "Methodology for data validation 1.0." *Essnet Validat Foundation* .

Groves, Robert M. and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5):849–879.

Hoque, Jesmeen, Jakir Hossen, Md Shohel Sayeed, C.K. Ho, K. Tawsif, Md. Armanur Rahman and Md Hossain. 2018. "A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics." *Indonesian Journal of Electrical Engineering and Computer Science* 10:1234–1243.

Julien, Claude, Claus Sthamer, Abel Coronado, Jimena Juárez, Siu-Ming Tam, Bart Buelens, Wesley Yung, Florian Dumpert, Gabriele Ascari, Fabiana Rocci, Joep Burger, Hugh Chipman, InKyung Choi and Claire Clarke. 2022. "Machine Learning for Official Statistics." *United Nations Economic Commission for Europe Statistics Publications* .

Kolluri, Johnson, Dr Shaik Razia and Soumya Ranjan Nayak. 2020. "Text Classification Using Machine Learning and Deep Learning Models.".

OECD. 2019. Hello, World: Artificial Intelligence and Its Use in the Public Sector. OECD Working Papers on Public Governance 36.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1(5):206–215.

Saris, Willem E. and Irmtraud N. Gallhofer. 2014. *Design, Evaluation, and Analysis of Questionnaires for Survey Research, 2nd Edition*. John Wiley & Sons, Inc.

Saris, Willem E. and Melanie Revilla. 2016. "Correction for Measurement Errors in Survey Research: Necessary and Possible." *Social Indicators Research* 127(3):1005–1020.

Smith, T. W. 2011. "Refining the Total Survey Error Perspective." *International Journal of Public Opinion Research* 23(4):464–484.

Statswiki, UNECE. 2019. Generic Statistical Data Editing Model (GSDEM). Technical Report 2.

Weisberg, Herbert F. 2009. *The total survey error approach: A guide to the new science of survey research*. University of Chicago Press.

Wirtz, Bernd W., Jan C. Weyerer and Carolin Geyer. 2019. "Artificial Intelligence and the Public Sector—Applications and Challenges." *International Journal of Public Administration* 42(7):596–615.