Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Data editing for machine learning prediction models

Prepared by Susie Jentoft, Statistics Norway

I. Introduction

1. Imputation is the common practice of substituting in realistic values when there is missing data. It is both useful to have complete data in order to create nice tables but can also be used to improve estimations by correcting for various missing data patterns. It is a technique used in official statistics for completing both survey data and administrative data in many different contexts. The methods used for imputing are now expanding beyond the traditional statistical approaches into the area of statistical learning involving machine-learning algorithms.
2. One concern is the quality of the training data, for which the imputation model or algorithm is learnt. If this contains many erroneous values, we must expect many error-prone values to also be predicted. Even if the imputation method is highly predictive, if we are basing it on data with a high error rate, we are not going to achieve high accuracy. This is not easily solved by simply improving the imputation method.
3. Alternately, the training data may contain high-quality data but be skewed in a way that means it does not represent the group to impute. In this case, there are certain methodological approaches that can be taken to improve the predictive outcome. For example, if using hot-deck imputation, stratifying using variables related to the outcome and missing data pattern can improve the estimates compared to simple hot-deck imputation.
4. In this study, we look at outlier detection and imputation of a key variable for Statistics Norway's employment statistics: contractual working hours in terms of the contractual full-time equivalent percentage (FTE). This is the number of contractual hours as specified in their working contract divided by the number of hours in a full-time position, as a percentage. This variable provides the basis for estimating the number of full- and part-time workers in Norway. It is also used in calculating earnings statistics, which are published in terms of full-time equivalent wages. We describe the processes used to identify outliers in the data, comparing the method used in 2018 and changes in 2019. We look at how this change in the identification of outliers has affected the imputation groups characteristics and the resulting predictions for imputation. Finally, we investigate an alternative approach to learning the prediction model by synthetically balancing the training dataset to represent the imputation population.
5. Numbers in this paper may differ from those published by Statistics Norway. This is due to some differences in the underlying available data at the time of this study and some simplifications of the editing processes for analysis reasons. Statistics in this study should therefore not be used for official statistics use.

II. Outlier detection

A. Data source

6. Statistics Norway publishes employment and earnings statistics based on administrative data. All employers in Norway are required by law to report salary and working hour information, on a monthly

basis, into a centralized system. The system is jointly owned by the Norwegian Tax Department, Norwegian Labour and Welfare Administration and Statistics Norway (Skatteetaten, 2020). This system, called A-ordning, was established in 2015 and provides the data for employment statistics, currently published on a quarterly and yearly basis. The base unit in the A-ordning is job. A person may be included several times if they have several jobs. A variant of the register is used in this study which only includes one job per person (the main job) and only employees. A key variable from this data source is contractual FTE, which is used to determine the number of people in full-time work. This variable ranges from 0 to 120 as a percentage.

7. Since the establishment of A-ordning in 2015, Statistics Norway has seen a steady improvement in the reported data quality. Despite this, there are still values that are questionable or missing which we observe create a bias in the statistics, unless addressed. One issue relates to the default value in the reporting systems, which for many, are automated from their payroll system. The value generally defaults to a full-time position (100 percent) and so this value occurs more often than expected in the data.

8. The data for this study is based on that for the 4th quarter of 2018. Quarterly publications are based on information from the middle month of the quarter, therefore this work is based on data from November 2018. Additionally, yearly publications are also based on purely this month's data. The unit is persons whom are employees. For those with several jobs, only the main job is considered (Statistics Norway, 2020).

B. Description of outlier detection methods

9. In 2018, outlier detection methods were developed to identify which observations are believed incorrect. These values are then imputed based on a machine learning algorithm, called XGBoost, learnt on the observations which passed the outlier detection process. Adjustments to these outlier detections has been an area of development within Statistics Norway with regular improvements being made. However, we have noticed that small changes in these methods can led to reasonable differences in the values imputed and the statistics produced. This has led to the question on whether these changes are a direct result of the inclusion of outliers (and exclusion of correct values) in the training dataset (rubbish in, rubbish out) or whether it may be a result of structural differences between the training and prediction datasets that the prediction algorithm is not able to correct for.

10. The reporting system contains a series of automatic checks with instant feedback to users. Additionally, when data is received by Statistics Norway, it is controlled using automatic checks and pre-processed. Afterwards, there are three main statistical controls for detecting outliers in the FTE variable which are described here. Further details are described in Grini & Bakke (2019). These checks determine which observations are used in the training data for the prediction model and which will be imputed:

11. **Control 1)** For hourly-based employees, an iterative rate model is used for the contractual FTE against paid hours worked. A boundary of 2.5 times the studentized residuals was used in 2018 for determining the boundary for outliers which was adjusted up to 2.0 in 2019. This was adjusted as many reasonable observations (particularly at the upper end of the scale) were being identified as outliers and excluded. This resulted in fewer observations among those working full-time in the training dataset than expected. Additionally, this control was adjusted to an iterative regression model in 2019 due to investigation of the distribution of the residuals. The models are run within strata (industry groups x occupational groups) which were also adjusted in 2019 (earnings class x occupation groups) to create more homogenous groups. These changes have decreased the number of observations identified as outliers (see table 1). Observations within these limits move to the next control. Those identified as outliers at this stage are replaced by with the FTE base on payment hours instead of contractual and move to the second control.

12. **Control 2)** A fixed lower limit on the full-time equivalent wages is used to identify lower valued outliers. In addition, lower and upper limits for the hourly rate (for hourly-based workers) are used to determine outliers. All observations outside the limits are excluded from the training data and are instead imputed (see Grini & Bakke, 2019, for more details).

13. **Control 3)** An iterative regression is used for contractual FTE using the predictor paid earnings. Given the importance of paid earnings for tax purposes this is a variable of high quality and is more likely to be correctly reported than contractual FTE. In addition to paid earnings, variables including occupation, education and age are included for control in the modelling. In 2019, an additional cut-off was introduced. Those with contractual FTE of 100 percent and whom are in the top 2.5% (or 3.0% from 2019) for full-time equivalent earnings were not included in this final control. This was because many of these observations were taken out, only to be imputed with contractual FTEs at the same level as the

original observations, as there is an upper limit of 120 percent. This is indicative of an outlier detection method that is perhaps not performing correctly for the data and is an area of further investigation for Statistics Norway.

C. Characteristics of the accepted and rejected observations

14. Of the 2.8 million employees in 2018, around 2.65 million were accepted and used in the training dataset using the new 2019 control procedures. In comparison, 2.61 million were accepted using the procedure from 2018, i.e. an extra 40 thousand observations are now included. In this section, we compare the acceptance and outlier groups for the two methods for 2018 data.

15. In certain cases, the contractual FTE is not equal to that directly reported from the employer. The reported FTE was adjusted if certain criteria were fulfilled, for example, when the contractual FTE was reported as 0, this was converted to 100 percent, so long as the data passed all the outlier detection controls. Additionally, those missing contractual FTE were set to 100 percent for fixed-wage earners and to paid hours FTE for hourly-based workers. This is because some payroll systems have 0 as the default value for this variable and from experience, we know that most of these are actually full-time workers. The average reported FTE is shown in table 1 which gives the average value of the initially reported FTE into the system before processing (besides reducing very extreme values down to 150 percent).

Table 1. Summary of contractual FTE for accepted and rejected observations. 4th quarter, 2018

	Acceptance	Average contractual FTE	Average reported contractual FTE	Total number
New method (2019)	Accepted	83.0	80.2	2648270
	Rejected (outliers)	-	75.2	161130
	All (incl. imputation)	80.4	79.9	2809400
Old method (2018)	Accepted	81.1	81.0	2607102
	Rejected (outliers)	-	70.9	202298
	All (incl. imputation)	78.2	80.4	2809400

16. Table 1 one gives the percentage accepted under the old and new methods with break downs for age, education and occupation. For education level, the acceptance proportion is relatively evenly spread, with slightly lower rates of acceptance among the lower education levels. The changes in identification method were relatively evenly spread among the education levels. For sex, the acceptance rate among men is slightly lower than for females. Again, the changes in identification of outliers is relatively even between the two groups. For occupation groups, those with unknown occupation show the lowest levels of acceptance. This group corresponds to around 30 thousand employees. Next lowest are those working in areas relating to skilled agricultural, forest and fisheries work. There are around 27 thousand employees in this group and these types of jobs are often paid on an hourly basis without fixed hours, making reporting potentially difficult.

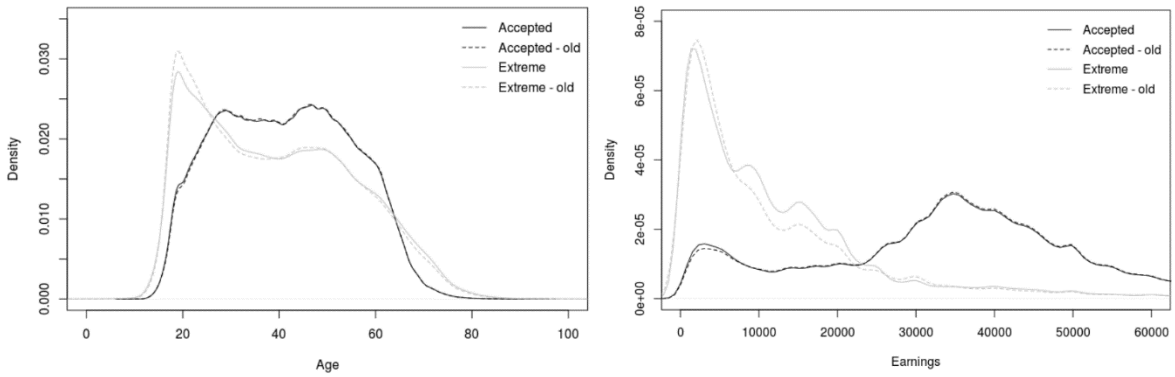
Table 2. Percentage of acceptance by education, sex and occupation. 4th quarter, 2018

		Acceptance percentage (2018 method)	Acceptance percentage (2019 method)
Education level	0 Primary, none/missing	90.5	91.8
	1 Lower secondary	89.1	91.3
	2 Upper secondary	93.3	94.7
	3 Post-secondary	94.3	95.6
	4 Tertiary undergraduate	94.6	95.6
	5 Tertiary postgraduate	91.3	92.8
Sex	Male	92.5	93.9
	Female	93.1	94.7
Occupation	Armed forces	98.4	97.5

Managers	91.9	94.5
Professionals	95.3	96.3
Technicians and associate professionals	94.1	95.4
Clerical support workers	93.3	94.2
Service and sales workers	91.1	93.7
Skilled agricultural, forest and fisheries	86.3	88.2
Craft and related trades workers	95.9	96.2
Plant and machine operators and assemblers	92.7	93.1
Elementary occupations	90.3	91.6
Unknown	56.0	56.6

17. The distribution of age and earnings for accepted and outlier observations are shown in figure 1. From this we see that the proportion of younger and older people is higher in the rejected/outlier group. The new method of identifying outliers appears to have reduced the proportion of young people within the rejected group slightly while increasing the proportion of older people a little.
18. The distribution of earnings shows there is a much higher rate of people with low earnings in the rejected group which is then imputed.

Figure 1. Density distribution of age (left) and earnings (right) for accepted and rejected groups using 2018 (old) and 2019 methods. 4th quarter 2018.



III. Imputation

A. Description of XGBoost

19. Since the publishing of 2018 employment statistics, Statistics Norway has used an XGBoost algorithm for predicting contractual full-time equivalent percentages (FTE). This method is based around gradient tree boosting described in Freidman et. al. (2000). In addition, the algorithm uses weighted quantile sketches for determining split points and sparsity-aware split finding (Chen & Guestrin, 2016). These additions have reduced the run-time for the algorithm and added subtle changes in how it deals with the bias-variance tradeoff. It results in an algorithm that is very fast and has a huge predictive power in a broad range of problems. Its popularity has grown in the last few years and has been the go-to algorithm for large-scale predictive problems, used in many of the top machine learning competitions (Nielsen, 2016). At Statistics Norway, this imputation method is used for contractual FTE for all observation that are identified as outliers in the previously described outlier detections. However, as seen in figure 1, the group which is to be imputed (outlier /extreme group) varies from that of the accepted group, particularly with regards to earnings.

B. Test of balancing the training dataset

20. In classification problems, it is common for machine learning algorithms to struggle to correctly predict rare classes when there is a high imbalance in the outcome classes. The algorithm will often lean towards the most common class as they aim to minimize the overall error rate, however we are often more interested in the "rare" class (Chen et. al, 2004). One proposal that has shown to work well is to balance the training data through re-sampling methods. The aim is to increase the proportion of the "rare" class in the training dataset in order to increase the prediction accuracy of that group (Batista, et. al, 2004).

21. This idea has led to this investigation on whether the same could be done to improve the accuracy of prediction for our numeric variable contractual FTE. As shown in figure 1, there are significant difference between the distributions of some of the variables between the training and predictive datasets. By resampling certain groups which are underrepresented in the training dataset, are we able to better predict these groups? We are particularly interested in improving the accuracy based on the characteristics of the group which is to be imputed (low earning and the young and older people. Additionally, by doing so for both outlier detection methods (2018 and 2019), we can see whether this method may reduce the sensitivity of the outcome, depending on which observations are identified in the controls.

22. For testing, we created a 75:25 split of the data among observations that were accepted. The 75 percent was used as the training data set for learning the XGBoost algorithm. We then created a "balanced" training dataset through the following procedure:

a. Calculate decile boundary values for age and earnings based on the predictive dataset (those observations that will be imputed/were rejected).
b. Use the values to determine which age and earnings groups the observation in the training dataset belong to.
c. Cross the age and earnings groups to determine the strata (h) for each observation in the training dataset and determine the observed strata sample sizes n_h
d. Determine a fixed strata size (n_{strata}) by taking the total training dataset size (n_{tr})/ the number of strata (m = 100).
e. Adjust the strata sizes in the training data by sampling without replacement (if $n_{strata} < n_h$) or with replacement (if $n_{strata} > n_h$) so that each stratum has the same number of observations ($n_{strata} = n_h$).

23. This procedure means that the distribution of the training dataset is similar to that of the prediction group (those rejected) with regards to age and earnings, two variables know to play an important role in the prediction model. This procedure was tested for both the 2018 and 2019 outlier detection methods, however, value for n_{strata} was set from the newer 2019 methods. This meant that there were the same number of observations in the training data for comparison between the methods.

24. After the "balancing", we learnt the model based on the training data and predicted on the test group. The root mean squared error for the training and test was calculated separately as

$$RMSE = \sqrt{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

where $\hat{y}_i$ is the predicted contractual FTE for observation i, and y_i is the observed contractual FTE. The bias was calculated as

$$\widehat{Bias} = E[\hat{y}] - y$$

25. The results from this show the training RMSE was lower for the balanced training compared to the original distribution dataset. The test RMSE, however, was considerably more for the balanced group (table 3). This I believe indicates that there is overfitting occurring in the balanced dataset. By replicating values among the groups which are under-represented in the training data we are giving more weight to them in the model. However, the model is also learning too much from the error assosciated with these observations, which is then not observed in the test dataset. The bias based on the test is relatively similar, slightly less in the balanced data.

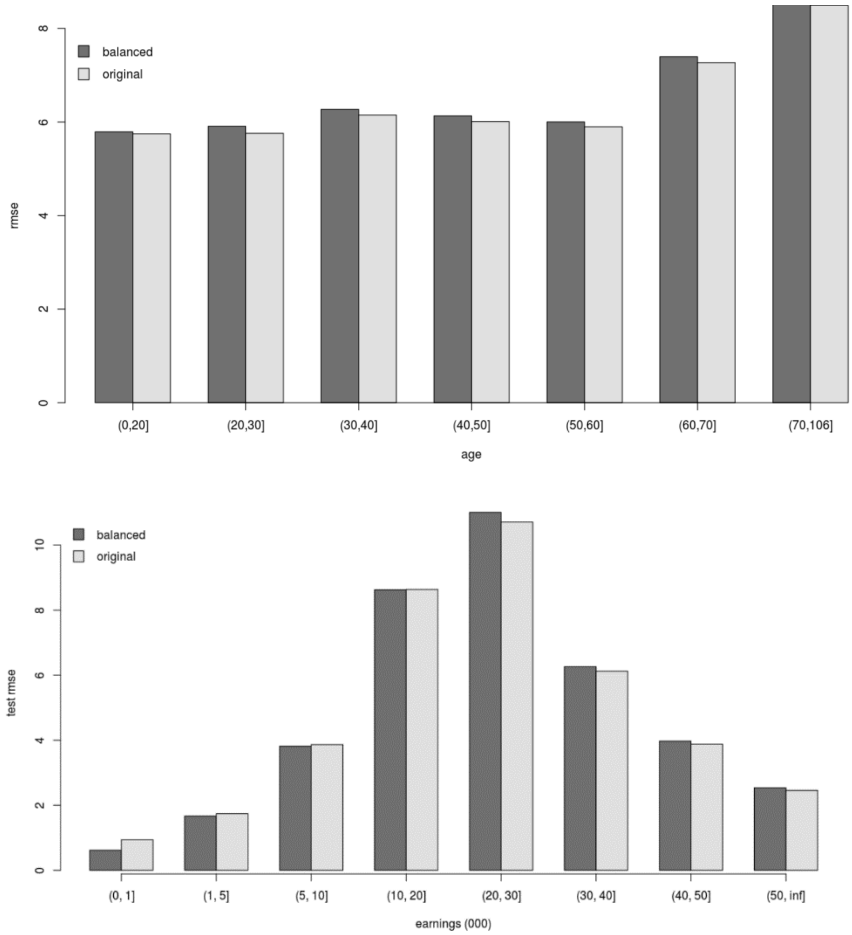
Table 3. Training and test results from balanced and original training datasets. 4th quarter, 2018 using new (2019) outlier detection method

	Balanced	Original
Training RMSE	5.5449	5.9892
Test RMSE	6.2045	6.0837
Test Bias	0.0124	0.0130

26. The table 3 results are based on the test data which has the same age/earnings distribution as the original accepted dataset. What we are really interested in is how is performs based on our prediction dataset (outlier values which will be imputed). We therefore breakdown the RMSE by age and earnings groups to observe how well the model is performing relative to our groups of interest. Figure 2 shows the test RMSE based on the balanced and original training datasets by age and earnings groups. From this we see that both perform similarly among the youngest and oldest age-groups, however, the original training dataset performs best (lowest RMSE values) in the middle age-group range. This can be explained due to the reduction in number of observations among these groups when balancing. The balancing however does not appear to improve the estimates among the youngest and oldest groups.

27. The test RMSE among the lower two earnings groups appears to be lower for that of the balanced training dataset. These are the groups that were most underrepresented in our original training dataset relative to group for prediction. The middle and higher earnings groups however, seem to be better predicted using the original training dataset.

Figure 2. Root mean squared error (RMSE) for the test data by age-groups (top) and earnings (bottom) for the balanced training dataset and original. 4th quarter, 2018 using new (2019) outlier detection method



C. Prediction of contractual FTE among the outlier group

28. While it is important to test how the XGBoost algorithm performs on known data, what we are really interested in is using the algorithm to predict for our rejected values/outliers. Here we don't have reliable observations to test against, but it is useful in the sense of observing its impact on the final statistics. Table 4 gives a summary of the calculated averages for contractual FTE for the outlier group and combined with accepted values. We see that for the outlier group, balancing the training dataset generally decreased the predicted contractual FTE under the 2019 outlier detection but increased it under the 2018 outlier detection method. This may simple be reflective of the stochastic nature of the algorithm or it may be caused by the underlying structural differences between these groups using the different detection methods.

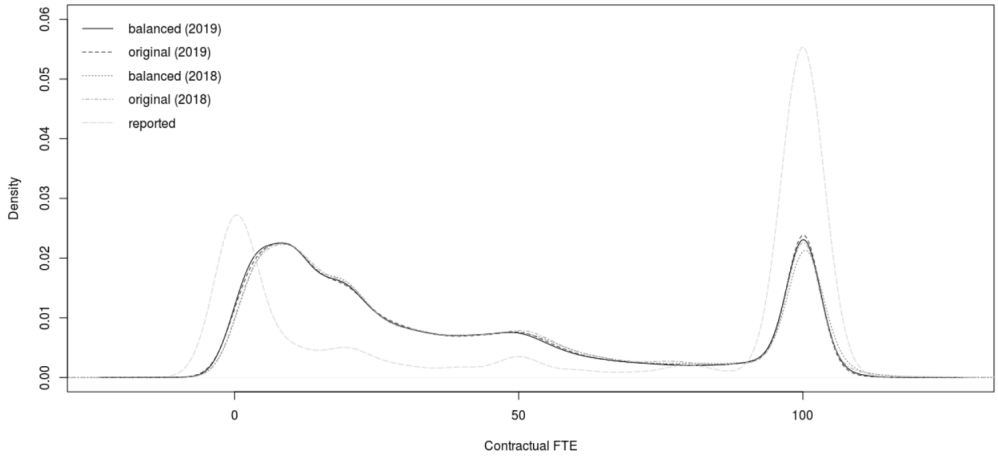
29. When observing the differences in imputation together with accepted values there is only very small changes at the overall level. It does appear however that balancing the training dataset is pulling the overall statistics towards each other. By that I mean, using the new 2019 outlier detection the estimate is lower using the balanced training dataset, in the direction of that of the average based on the 2018 outlier detection method. Similarly, the older 2018 outlier detection method, when the training dataset is balanced, the prediction and statistics are draw up towards the newer statistic.

30. The distribution of the imputed values under the different outlier detection methods and training datasets is shown in figure 3. Generally however, there is not significant differences between the methods.

Table 4. Average contractual FTE (percent) for imputed and total groups by training dataset type and outlier detection method. 4th quarter, 2018.

		Average contractual FTE among imputed group (outliers)	Average contractual FTE for all (accepted and imputed)
2019 outlier detection	Balanced	37.57	80.36
	Original	37.73	80.37
2018 outlier detection	Balanced	41.52	78.21
	Original	41.35	78.20

Figure 3. Distribution of imputed contractual FTE (percent) for balanced and original training datasets and by outlier detection method. Distribution of reported values (for those not missing) are also shown for comparison. 4th quarter, 2018.



V. Conclusions

31. It has been difficult to find literature around how to handle prediction problems using machine learning techniques where the observations for training the data differ from those we want to predict for. This is, however, a common problem in official statistics in both survey (due to non-response) and

administrative (due to coverage errors) based statistics. It is an interesting area to investigate further given the increasing use of these type of techniques in prediction problems.

32. Overall, the work here has highlighted some of the differences between the accepted and outlier groups for the contractual FTE variable. The technique of creating a "balanced" training data through re-sampling show some signs that it may improve estimates among groups of interest, however, it is far from conclusive and there was generally fairly insignificant differences. In this investigation we adjusted the training dataset to be of equal sample size to the original but increasing the size by including all observation and re-sampling with replacement for underrepresented groups could be an alternative to explore further. Additionally, we have only looked at the overall effect of the imputed values, whereas this type of technique may have more influence on smaller sub-populations. This may be useful for further investigation, especially with increasing demands for more detailed statistics.

33. The "balancing" process of the training dataset did not appear to explain the differences when considering changes in outlier detection method. For this reason, it is plausible to assume that the changes to outlier detection have not simply been structural but also improved through excluding more incorrect values while including more observation that are reasonable. Further work on improving these will likely impact positively on the quality of the statistics.

Acknowledgments

I would like to thank the division for labour statistics for their dedicated work on data editing in the A-ordning and preparing data for me. A special thanks goes to Knut Håkon Grini, Stine Bakke and Ingvild Johansen whom have worked vigorously on the changes to the outlier detection methods.

References

Chen, Chao; Liaw, Andy; Breiman, Leo (2004). *Using Random Forest to Learn Imbalanced Data.* Report nr. 666. Universtiy of California, Berkley

Chen, Tianqi; Guestrin, Carlos (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

Batista, Gustave E. A. P. A; Prati, Ronaldo; Monard, Maria C. (2004) A study of the behaviour of several methds for balancing machine learning training data. *SIGKDD Explorations*: 6 (1)

Friedman, J; Hastie, T; Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Annuals of Statistics,* 28(2): 337-407

Grini, Knut Håkon; Bakke, Stine (2019). *Predicting the contractual full-time equivalent percentage using XGBoost*. Paper for the Nordic Statistical Meeting, Helsinki 2019.

Nielsen, D. (2016). *Tree boosting with XGBoost. Why does XGBoost win "every" machine learning competition?* Thesis for Master of Science in Physics and Mathematics. Norweigan University of Science and Technology.

Skatteetaten, (2020). *A-melding* https://www.skatteetaten.no/en/person/taxes/get-the-taxes-right/employment-benefits-and-pensions/hobby-odd-jobs-and-extra-income/paid-work-in-the-home/salary-paid-over-nok-60000/a-melding/ (Accessed 14.02.2020)

Statistics Norway (2020). *About the statistics.* https://www.ssb.no/en/arbeid-og-lonn/statistikker/regsys (Accessed 17.02.2020)