Discussion paper
UNECE Machine Learning for Official Statistics Workshop 2023

Date
April 2023

Version
1

Authors
Jens Malmros, methodologist, PhD
Data division, methodology
Jens.malmros@scb.se

Imputation of occupation in the Occupational Register

Abstract

Statistics on occupation in Sweden are published for the gainfully employed population 16-64 years old. The statistics come from the Occupational Register, which contains information on the occupation of individuals. Because occupation is intermittently collected, 6 % of the register population has a missing value on occupation, and 4 % of the register population has an imputed value. The present imputation model is however obsolete.

Information on occupation from the Occupational Register will be included in the recently pre-released register-based labour market statistics, Population by Labour market status (BAS), during 2023. The population of BAS is (gainfully employed) individuals 15-74 years old. Because this population is larger than the population for the current statistics on occupation, the proportion of missing values may increase.

Because of these issues, a new machine learning imputation model for occupation is developed using data from the 2019 Occupational Register on the gainfully employed population 16-74 years old. Results from evaluation of a model using occupation as the response variable and multiple register variables as explanatory variables are shown with respect to individual level and class level predictive performance, and with respect to the effects on the statistics.

The accuracy of the model is 54 % and the macro F1 score is 44 %. The predictive performance of the model varied substantially between classes. However, almost all classes fulfil the quality guidelines for the Occupational Register. The distributional accuracy of the predictions depends on the method that is used.

Introduction

Between 2019-2021, Statistics Sweden carried out the Subject Area Design Labour Market project, with the aim to modernise and future-proof the Swedish labour market statistics. The project resulted in a plan for the development of the labour market statistics. A first result of the project is the introduction of the new register-based statistical product Population by Labour Market Status (BAS), which was released as statistics under construction[1] in May 2022. During 2023, BAS will be released as official statistics, and the scope of BAS will expand to include, e.g., statistics on occupation. The target population of BAS is (gainfully employed) individuals in the Total Population Register (TPR) 15-74 years old.

At present, the Occupational Register contains information on the occupation of individuals in Sweden. Information on occupation from the Occupational Register will be included in BAS. Because of quality issues related to the occupation variable, especially concerning the number of missing values, an imputation model for occupation is to be developed. The present paper describes the development of the imputation model for occupation.

The Occupational Register

The Occupational Register contains information on occupation for individuals in the TPR aged 16 years and older. In addition to occupational information, the register contains variables related to occupation on the individual and the employer level. The register is the source for the official statistics on occupation, with target population the gainfully employed population 16-64 years old.

Occupation is classified according to the Swedish standard SSYK2012, which is based on ISCO-08. The occupational code has four digits, where each consecutive digit represents an increasing level of detail. There are nine main categories corresponding to the first digit of the occupational code and 425 possible values of the complete four-digit occupational code. From here and onwards, we refer to the variable providing occupational information according to SSYK2012 as Occupation.

The Occupational Register was formed in 2001 because of the decision to replace the traditional Population and Housing Census with the register-based Population and Housing Census. According to the

[1] I.e., statistics, which are not official statistics, and which are under development. Sometimes denoted “experimental statistics”.

government bill on the Occupational Register, the quality of the Occupational Register should be similar to that of the traditional Census, i.e., the proportion of missing values should not be larger than 5-10 %, and the proportion of wrongly classified individuals should not be larger than 10-15 %.

For each individual in the register, the quality of Occupation is rated based on the relationship to the current employer, the information of which comes from the Labour statistics based on administrative sources. If Occupation comes from the current employer, the observation gets the highest quality rating. If this is not the case, the observation gets a lower quality rating. For some individuals, Occupation is missing, or Occupation was imputed using an old model for imputation.

Occupational information in BAS

Starting in 2023, BAS will include information on occupation. The target population of BAS is individuals 15-74 years old; hence, it includes more individuals than the target population for the current official statistics on occupation, i.e., individuals 16-64 years old. Because the proportion of individuals with missing values on Occupation is larger for the youngest and the oldest individuals in the target population for BAS, the proportion of missing values for Occupation will increase in BAS compared to the Occupational Register.

Imputation may be used to reduce the number of individuals with missing values on Occupation. In the current Occupational Register, imputation is made using an old model, whose primary purpose was to aid in the transformation between the previous classification of Occupation, SSYK96, and the current SSYK2012, and which is dependent on occupational information classified by SSYK96. Because the number of individuals with such information decreases each year, the model becomes less relevant with time.

The purpose of the present work is to develop a new model for imputation of Occupation in the Occupational Register and eventually for use in BAS. The variable of interest is Occupation on the 4-digit SSYK12 level. Occupation is imputed on the individual level and evaluation of the model address the individual level performance as well as the class level performance and the effect on the distribution of Occupation.

Methods

Methodological overview

The present task is an imputation problem, i.e., a prediction problem. For observations with missing values on Occupation, we want to replace the missing values with values predicted by a model. Ideally, the imputed value should be equal to the true value, i.e., that which would have the highest quality rating.

Traditional imputation procedures such as mean imputation or hot-deck imputation often stem from practical considerations and may suffer from statistical flaws. For complex imputation scenarios where much auxiliary information is available, traditional methods will in general be inadequate. In such cases, modelling approaches are needed to utilize the potential of the auxiliary information for prediction.

Modern machine learning methods typically provide superior predictive performance compared to traditional regression and classification models for large datasets and complex relationships (Dagdoug, Goga, & Haziza, 2021). In recent years, machine learning methods have been used for imputation at several other NSIs (UNECE, 2022). In addition, the Operational plan for Statistics Sweden and related documents highlights the use of machine learning, and in particular its use for imputation. For these reasons, we choose a machine learning approach for our imputation problem.

Data

Data come from the 2019 Occupational Register on the gainfully employed population 16-74 years old. The data consists of 5 million observations, of which 78 % have a value on Occupation of high quality, 4 % are imputed, 12 % have a value on Occupation of low quality, and 6 % are missing. Only observations with high quality are used to train the model; the remaining observations are used in the evaluation.

Besides Occupation and its quality rating, data contain variables on both the individual level, for example, age, sex, and field of education, and the employer level, for example industrial classification and number of employees. The data also include a categorical variable formed by business name. For example, business names that includes “pizz” would make up a category primarily consisting of pizza restaurants.

Occupation has 425 classes, i.e., possible SSYK2012 values, and an imbalanced heavy-tailed distribution. For example, 3.3 % of the population has the most common occupation, assistant nurse, while the 120 least common occupations together comprise 3 % of the population.

Imputation model

The present work concerns a machine learning implementation to address an imputation problem. We limit ourselves to considering tree-based models, which often show good predictive performance on many problems. Two models are considered, random forests and gradient boosted classification trees. The performance of the models is very similar; however, less computational resources are needed to train the random forest model. We therefore only consider results for the random forest model in this paper.

Random forests (Breiman, 2001) are made up of many classification trees. Each tree is trained separately using a randomly selected subset of the explanatory variables and a randomly selected subset of the training data, which reduces the correlation between the trees. The prediction from the model is the majority vote from all the trees. We use the random forest R package *ranger* in our implementation (Wright & Ziegler, 2017).

We train the model on data with high quality. These data are split into 80 % training data and 20 % validation data. The remaining observations make up test data since predictions will be done on (a subset of) them when the model is used. Note that test data do not include the true values, i.e., values on Occupation with high quality.

Evaluation

Predictive performance

We evaluate the model with respect to predictive performance on validation data. The individual level predictive performance of the model is evaluated with respect to accuracy, precision, recall, and F1. We also evaluate the performance by comparing predicted values with true values from the 2020 Occupational Register.

The predictive performance is also evaluated for each class with respect to the quality guidelines given in the government bill on the Occupational Register. In the evaluation, we use validation data, in which we simulate missing values to mimic the distribution of missing values in the population. Specifically, an observation will be missing with a probability depending on the values on age and industrial classification. Simulated missing values are then replaced with imputed values and the proportion of correct values is computed using all observations belonging to the class. According to the quality guidelines, the proportion of wrongly classified values should not exceed 10-15 % for the general population; we perform this evaluation on the class level.

Effects on the statistics

We evaluate the effects on the statistics on Occupation by considering the effect of imputation on the distribution of Occupation in the general population. It is desirable that the distribution of imputed values mimic the distribution of true values. We use validation data with simulated missing values replaced by imputed values as described previously and compare the distribution of the true values on Occupation with the distribution of values on Occupation when simulated missing values were imputed using a permutation test and by evaluating the numerical differences between the distributions.

In addition, we also evaluate the distribution of Occupation when missing values were imputed in test data.

Results

Predictive performance

Individual level predictive performance

The accuracy of the model is 53.7 % and the macro F1, i.e., the mean F1, is 44.3 %. The macro precision is 45.7 %, and the macro recall is 44.2 %. The performance varies substantially between classes.

Table 1 shows the five classes with the highest F1, and Table 2 shows the five classes with the lowest F1. For each class, the SSYK code, its description, the number of observations, precision, recall, and F1 is shown. The tables show that the difference between precision, recall, and F1 between the best predicted classes and the worst predicted classes is large. Note that two classes, to which no individuals are predicted, are left out from Table 2.

Table 1
The five classes with highest F1.

SSYK	Description	Number of obs.	Precision	Recall	F1
9621	Newspaper carriers[2]	1030	92.3%	97.7%	94.9%
2613	Prosecutors	198	89.4%	97.5%	93.2%
2260	Dentists	1208	90.8%	95.6%	93.1%
2222	Midwives	1288	88.6%	93.6%	91.0%
5141	Hairdressers	1149	89.8%	91.9%	90.8%

[2] Ad-hoc translations of descriptions of occupation.

Table 2
The five classes with lowest F1.

SSYK	Description	Number of obs.	Precision	Recall	F1
1351	Property service managers, level 1	119	1.8%	0.8%	1.1%
1241	Information, communication, and PR managers, level 1	254	2.7%	0.8%	1.2%
1421	Preschool managers, level 1	83	10.0%	1.2%	2.2%
1311	IT managers, level 1	607	8.3%	3.0%	4.4%
1222	HR managers, level 2	947	9.2%	3.0%	4.5%

Figure 1 shows the distribution of F1 by the first digit of SSYK[3]. We see that the F1 scores varies between the categories, and that it is especially low for the value 1, which represents managers. Because managers are difficult to separate from the other occupational categories using our explanatory variables, this is not surprising.

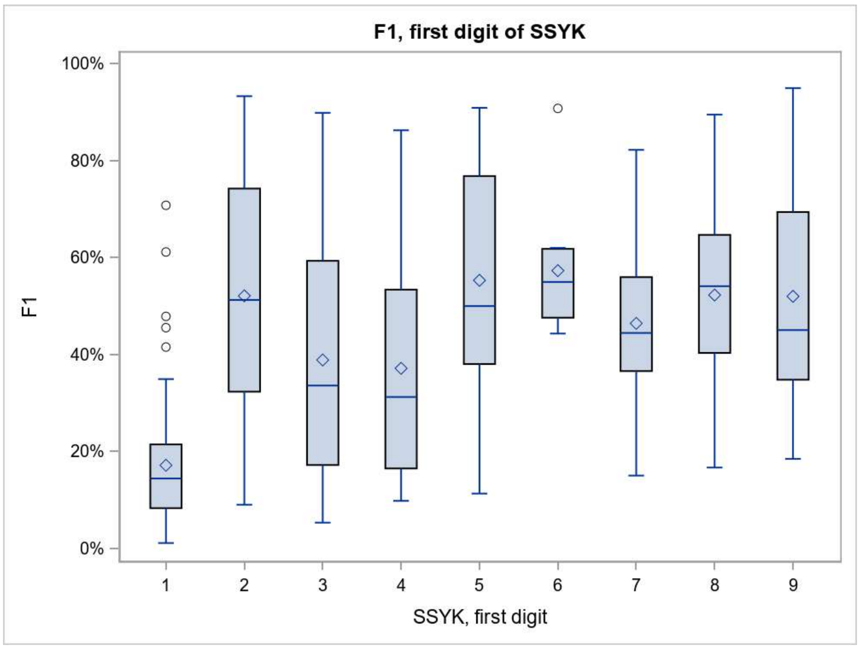

Figure 1
F1 by first digit of SSYK.

Class level predictive performance

When simulated missing values are replaced by imputed values, the proportion of wrongly classified values within each class is less than 10 % for 402 classes, and it exceeds 15 % for three classes, which has 16 %, 16 %, and 25 % wrongly classified values, respectively. Hence, the

[3] Four-digit SSYK is predicted, and the first digit of SSYK is extracted from the predicted values.

majority of classes fulfil the quality guidelines for the Occupational Register.

Effects on the statistics

Distribution of imputed values

The null hypothesis that the distribution of true values on Occupation and the distribution of values on Occupation when simulated missing values were imputed are the same is rejected in a permutation test where either the Kolmogorov-Smirnov statistic or the distance of total variation is used (p=0 %). If we consider the actual difference between the distributions, it is however small for most classes. The largest difference in population proportion is 0.08 percentage units for software engineers. For 394 of the 425 classes, the difference is 0.01 percentage units or less. The bias is larger for classes with many observations in data.

The predicted value is the majority vote of the predictions from the individual decision trees in the random forest. However, one may consider the distribution of the individual tree predictions as a probability distribution on occupation for an individual (Laitila, 2011). If the predicted value on Occupation is randomly chosen from this distribution, the distribution of true values on Occupation and the distribution of values on Occupation when simulated values were imputed will be similar and the null hypothesis will not be rejected. This is further elaborated in the Discussion.

Imputation in test data

We impute values on Occupation in test data, i.e., for individuals, who previously had missing values. There are 229 100 such individuals in our data. The difference in population proportion before and after imputation is small for most classes. Table 3 shows the five classes with the largest difference between the population proportion before and after imputation. Note that missing values are left out from the calculations on the population proportion before imputation.

Table 3
The five classes with largest absolute difference in population proportion.

		Before imputation		After imputation				
SSYK	Description	Number of observations	Population proportion	Number of observations	Population proportion	Difference population proportion	Relative difference within SSYK	F1
9412	Restaurant and kitchen assistants	71122	1.6%	97477	2.1%	0.5%	37%	48.6%
9111	Cleaners	77575	1.8%	87855	1.9%	0.1%	13%	68.0%
5321	Assistant nurses	132252	3.1%	133138	2.9%	-0.1%	1%	74.0%
2341	Primary school teachers	108770	2.5%	109632	2.4%	-0.1%	1%	77.4%
5222	Shop salesmen, everyday commodities	86873	2.0%	96069	2.1%	0.1%	11%	88.2%

Evaluation on true values from 2020

Some individuals, which had a missing value on Occupation in 2019, will have a value with high quality on Occupation in 2020. For them, the values predicted for Occupation in 2019 may be compared to the true values from 2020. Of the 229 100 individuals, which have a missing value on Occupation in 2019, 56 900 have a value on Occupation with high quality in 2020, 27 200 have a value on Occupation with low quality in 2020, and 145 000 have a missing value on Occupation in 2020.

For individuals, which have a missing value on Occupation in 2019 and which have a value with high quality on Occupation in 2020, the predicted SSYK value for 2019 is the same as the true value on Occupation for 2020 on the four-digit level for 16.9 % of the individuals and the same on the one-digit level for 25.8 % of the individuals. If we consider the subset of these individuals which have the same value on industrial classification in 2019 and 2020, the predicted values for 2019 are the same the true values for 2020 on the four-digit level for 52.6 % of individuals and on the one-digit level for 68.9 % of individuals.

Discussion

In this paper, we present ongoing work on imputation of Occupation in the Occupational Register. Several issues remain to be investigated, for example:

- The predictions from the model are biased towards common occupations, which is likely caused by the imbalanced class distribution in training data resulting in few observations for small classes. As previously discussed, an alternative to imputing the majority vote from the random forest is to impute from the distribution given by the individual tree predictions. This will mitigate the bias; however, the predictive performance of the model will worsen. In this case, the accuracy drops from 53 % to 44 %. It remains to choose which prediction to use.
- The proportion of trees which predict the same value on Occupation may be seen as a measure of the certainty of the predictions. We may set a lower threshold on the proportion of trees which predict the same value such that observations with a value below the threshold will be set to missing. This will result in higher quality of imputed values, but it is currently unclear how and if this possibility should be used.
- The data collection for the Occupational Register may be re-designed to facilitate the use of the imputation model. For example, it may be beneficial for the quality of the predictions to increase the collection of data from workplaces where small occupations are represented.

References

Breiman, L. (2001). Random Forests. *Machine Learning*, Vol. 45, pp. 5-32.

Dagdoug, M., Goga, C., & Haziza, D. (2021). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. *Journal of Survey Statistics and Methodology*, Vol. 00, pp. 1-48.

Laitila, T. (2011). *On imputation of binary variables in registers.* Ljubljana: Conference of European Statisticians.

UNECE. (den 11 11 2022). *Machine Learning for Official Statistics.* Hämtat från UNECE: https://unece.org/statistics/publications/machine-learning-official-statistics

Wright, M., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, Vol. 77, No. 1, pp. 1-17.