UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Machine Learning for Official Statistics Workshop 2023
5-7 June 2023, Geneva Switzerland
04 April 2023

Imputation of occupation in the Occupational Register

Jens Malmros (Statistics Sweden)
jens.malmros@scb.se

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 of occupation, and 4 % of the register population has an imputed value of occupation. 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. Data come from the 2019 Occupational register on the gainfully employed population 16-74 years old. Several tree-based models using occupation as the response variable and multiple register variables as explanatory variables are evaluated. Because the response variable has many classes, and because the distribution is imbalanced, several measures are used to evaluate the performance of the models. The best performance is achieved by a random forest model. The accuracy of the model is 54 % and the macro F1 score is 44 %. The predictive performance of the model varied substantially between classes. The performance was also evaluated by comparing values on occupation imputed in 2019 with collected values on occupation in 2020 for the same individuals. The effect of imputation on the statistics is also evaluated.