UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna Austria

12 July 2024

Random forest imputation of nutritional information for statistics on food consumption in Norway

Magne Furuholmen Myhren & Susie Jentoft (Statistics Norway, Norway)
Magne.Myhren@ssb.no, susie.jentoft@ssb.no

Abstract

In 2024, Statistics Norway will publish average nutritional consumption for the first time, based on large, new data sources. Scanner receipt data from the largest supermarket chains is used primarily for the Consumer Price Index but provides the main data for the new statistic. This is linked using individual item keys (GTINs), to food databases and web-scrapped data containing nutritional values. Not all food items can be directly linked to nutritional information and a sequential imputation method was developed using random forest models, to complete the data. The models were trained on pre-processed food item descriptions, internal food group labels from the supermarket chains and COICOP tokens as features. Ten nutritional values were imputed for each item, including energy, fat content, sugar, salt and protein. The implemented random forest models performed above a baseline model (average values within COICOP groups) for all but one nutritional value (alcohol). The models performed similarly to a previously tested, more advanced imputation method using random forest imputation within nearest neighbour data pools, based on Jaccard similarities. The implemented imputation method appears to perform well and allows a complete data source to create nutritional statistics. Further work on tailoring and tuning the model for specific groups (including alcohol content) is planned for future rounds of publications. As the data sources do not rely on expensive survey collection, more frequent publications are feasible.