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

Changing Data Sources in the Age of Data Science for Official Statistics

Cedric De Boom, Michael Reusens (Statistics Flanders, Belgium)
cedric.deboom@vlaanderen.be

Abstract

Data science has become increasingly essential for the production of official statistics, as it enables the automated collection, processing, and analysis of large amounts of data. With such data science practices in place, it enables more timely, more insightful and more flexible reporting. However, the quality and integrity of data-science-driven statistics rely on the accuracy and reliability of the data sources and the techniques that support them. In particular, changes in data sources are inevitable to occur and pose significant risks that are crucial to address in the context of data science for official statistics.

This paper gives an overview of the main risks, liabilities and uncertainties associated with changing data sources in the context of data science for official statistics. We provide a checklist of the most prevalent origins and causes of changing data sources; not only on a technical level, but also regarding ownership, ethics, regulation, public perception, etc. Next, we highlight the repercussions of changing data sources on statistical reporting. These include technical effects such as concept drift, bias, availability, validity, accuracy and completeness, but also the neutrality and potential discontinuation of the statistical offering. We offer a few important precautionary measures, such as building robustness in both data sourcing and statistical techniques, and thorough monitoring. By doing so, official statistics based on data science can maintain their integrity, reliability, consistency and relevance in policymaking, decision-making, and public discourse.