Machine Learning Methods for Assessing the Consistency and Integration of Statistical and Administrative Data

Author: PROF. DR. Elena Zarova

Submission ID: 479

Format: CPS Paper

Reference Number: 479

Presentation File

abstracts/ottawa-2023_8e77dfd5f12a274a04051a10ba58bf80.pdf

Brief Description

Based on the generalization of the methods proposed in publications for integrating administrative data into the practice of national statistical offices, as well as taking into account the results of scientific developments in this direction, a set of methods for assessing the consistency of official statistics and administrative data has been developed (using the example of wage statistics).

On the basis of a complex of "traditional" statistical methods and machine learning methods, the necessary components for ensuring consistency are identified.
A set of statistical and machine learning methods for assessing the consistency of administrative and statistical data used in official practice at the regional level has been tested on real wage data.
The proposed methods for integrating administrative and statistical data at the regional level are relevant for the statistical practice of various countries.

Abstract

An important direction in the development of national statistics in most countries is the use of administrative data, along with data obtained from observations of national statistical offices.
Recommendations on the use of administrative data for the purposes of business statistics and a summary of the best foreign practices of national statistical offices in Europe on the use of administrative data are presented on the Eurostat website [1]. Based on the results of the implementation of projects on the basis of the UN European Commission, recommendations have been developed: “MIAD - Methodologies for an Integrated Use of Administrative Data in the Statistical Process” [2].
Examples can be given of the many scientific papers on the use of administrative data in the production of official statistics.
When solving scientific and practical problems of using administrative data for the purposes of official statistics, most authors propose a set of actions, summarized and very clearly presented in the UN Statistics Division presentation on the use of administrative data in the development of SDG indicators [3].
This set of actions includes:
“1. Inventory of all administrative registers available
2. Mapping of administrative entity types to statistical units
3. Mapping of administrative variables to statistical variables
4. Establishing relationships among administrative registers
5. Development of statistical registers
• Base statistical registers
• Primary statistical registers (directly based on administrative registers)
• Integrated statistical registers (derived from primary registers)
6. Linking microdata from statistical registers and other data sources” [3].
While agreeing with these "steps" to integrate administrative data into official statistics, we note that they do not include actions to ensure the consistency of the statistical distributions represented by these data. Register matching does not solve this problem because units with the same register codes can generate different statistical distributions in administrative microdatabases and statistical microdatabases. For example, the distribution of populations of units with the same registry code reported in payroll tax data may differ significantly from the distribution of those units in the Labor Force Survey (LFS) database of the National Statistical Office. There may be several reasons for this. Of these, two are the main ones: (1) the administrative data represent a complete observation, while the LFS data are the result of a sample observation (and there may be a sampling bias effect); (2) there are qualitative differences between the observed units: not every person observed in the LFS is a taxpayer; (3) there are differences in the “program” of observation: the administrative data on wages do not depend on the age of the taxpayer, the data of the statistical observation of the LFS are the population aged 15+.