Machine Learning Methods for Processing and Analyzing Integrated Datasets of Official Statistics

Elena Zarova

Analytical Centre by Moscow City Government, Russia, Moscow,

Zarova.ru@gmail.com

Abstract

The development of machine learning methods for solving non-trivial tasks of processing and analyzing statistical data is driven by the needs of big producers of statistical information, and, above all, national statistical offices. This need is due to the increasing use of new statistical data resources in official statistics: administrative data, big data, information from corporations and producers of unofficial statistical information. At the same time, the national statistical offices set the task of optimizing the observations, reducing the burden on the respondents. The solution of these tasks requires the inclusion of unstructured, semi-structured, heterogeneous and non-harmonized data in the production of official statistical information. For their effective processing and analysis, it becomes necessary to use machine learning and data mining methods, which, on the one hand, tend to the rapid increase in the volume of new information (the content of which also is rapidly changing), and, on the other hand, to optimize statistical observation. The advantage of using machine learning methods in the practice of official statistics is that they allow, on the basis of data mining approaches, to extract hidden, not based on previously accepted hypotheses, information patterns in integrated statistical data arrays, which makes it possible to quantitatively characterize new phenomena and processes that arise and developing in reality, but not provided for by the a priori given structure of the officially collected information.

The experience of applying machine learning methods in the processing and analysis of statistical data is presented on the websites of the national statistical offices of the USA, Canada, Great Britain, Germany, and a number of other countries, as well as on the Eurostat website. This article is based on the generalization of this experience, as well as its theoretical assessment and development. The novelty of the idea includes substantiation of the theoretical and practical significance of aggregation of statistical matching and machine learning methods in the practice of official statistics. The example of integrating microdata arrays of a statistical labor force survey and survey income of households is presented. The result of the study was tested on published data by the Federal State Statistics Service of Russia. This approach and the need for its methodological and practical support are due to the "Development Strategy of Rosstat and the system of state statistics of the Russian Federation until 2024".

The purpose of this approach is to obtain useful information about the phenomena hidden from traditional statistical observations, as well as recommendations for optimizing the content of official questionnaires.

Keywords: machine learning, statistical matching, aggregation, income survey, labor force survey, syntactic array, optimization.

Introduction

Currently, in the official statistics of most countries, the tasks of increasing efficiency are being solved, that is, increasing the volume and quality of information extracted from the observed data. To solve these problems in the practice of official statistics, two new areas are being actively formed to improve the work with observable data.

These two areas include:

1) applying methods and technologies of statistical matching (SM) to integrate datasets from sample surveys to produce synthetic datasets that are more informative than any dataset alone;

2) the introduction of data mining procedures based on machine learning (ML) methods in the processing and analysis of observed data in order to extract non-trivial, previously unknown useful information.

Examples of the development and implementation of statistical matching technologies, that is, obtaining aggregated information collected from several data sources obtained from the same population, are the work of the National statistical institutes of the EU countries on integrating statistics on income and living conditions (EU-SILC) and the household budget survey [1], datasets of EU-SILC and labor force survey (LFS) [2], statistics of EU-SILC and European quality of life surveys (EQLS) [3].

Confirmation of the relevance of the second area of development of official statistics is the machine learning project, which was launched by the UNECE High Level Group on the Modernization of Official Statistics (HLG-MOS) in 2019. According to their publication, "The project aimed to demonstrate the added value of ML, i.e. whether its enables to production of more relevant, timely, accurate and trusted data in an efficient manner. The project also aimed at increasing the capability of NSOs to use ML by identifying and addressing some common challenges encountered when incorporating ML in organizations and their production processes"[4].

These directions for improving official statistics are still being developed by national statistical institutes on an experimental basis. Nevertheless, even at this stage of the implementation of SM and ML methods, a more significant effect of extracting new useful information from aggregated statistical survey data is obvious, providing

users with new “non-programmable” knowledge about the objects observed by statistics.

Non-programmable knowledge in this case is the aggregated results of statistical observation in official statistics, not based on a priori hypotheses about their structure and relationships, which determine the forms of the output tables and, as a result, the structure of published statistical data for users.

The publication is devoted to the tasks, information support and results of combining SM and ML methods for processing and analyzing data sets of official statistics of Russia.

Methodology

For the purpose of obtaining additional information, methods and algorithms for the formation of a synthetic array based on the integration of microdata arrays of two sample surveys have been developed. These are surveys conducted by the Federal State Statistics Service (Rosstat):

1) Labor Force Survey (LFS);

2) Sample survey of incomes of the population and participation in social programs, or the survey of income of the population (ODN - transliteration of the Russian abbreviation).

To clarify the problem statement, it is necessary to consider the general and specific variables of the ODN and LFS arrays. These two arrays have common variables characterizing the respondents (people in the sample): gender, age, type of settlement, marital status, educational level, actual work hours per week, employment category (employment / self-employment), and individual weight coefficient of the respondent.

The variable "wages per hour worked, rubles." is specific to ODN: it is present only in this array and is absent in the LFS. Based on this, the task of integrating the ODN and LFS arrays is to obtain a synthetic microdata array containing both the employment characteristics for each respondent and the characteristics of labor remuneration (hourly wages). At the same time, the need for integration is due to the different statistical reliability of the surveys under consideration: LFS in Russia is carried out monthly, covering more than 960 thousand respondents aged 15+ on an annual scale; ODN is held once a year with coverage of 60 thousand households (and once every five years - with coverage of 160 thousand households.). Accordingly, in SM procedures, the ODN array should be considered as the donor array, and the LFS array as the recipient array.

In this case, the goal of integrating ODN and LFS can be concretized as obtaining a synthetic array of microdata, according to the reliability parameters of the

corresponding LFS, but containing data on wages (which are contained only in ODN).

There are many definitions of machine learning in the scientific literature. One of the most understandable for practical implementation is the following: “Machine Learning (ML) is the science of getting computers to automatically learn from experience instead of relying on explicitly programmed rules, and generalize the acquired knowledge to new settings”[5].

Machine learning methods are divided into two types: supervised learning and unsupervised learning, they allow you to solve two main interrelated problems: the classification problem and the regression problem.

Figure 1 shows new data for the official statistics of Russia, obtained as a result of solving problems based on the use of SM and ML methods for the processing and analysis of the integrated ODN and LFS data array.

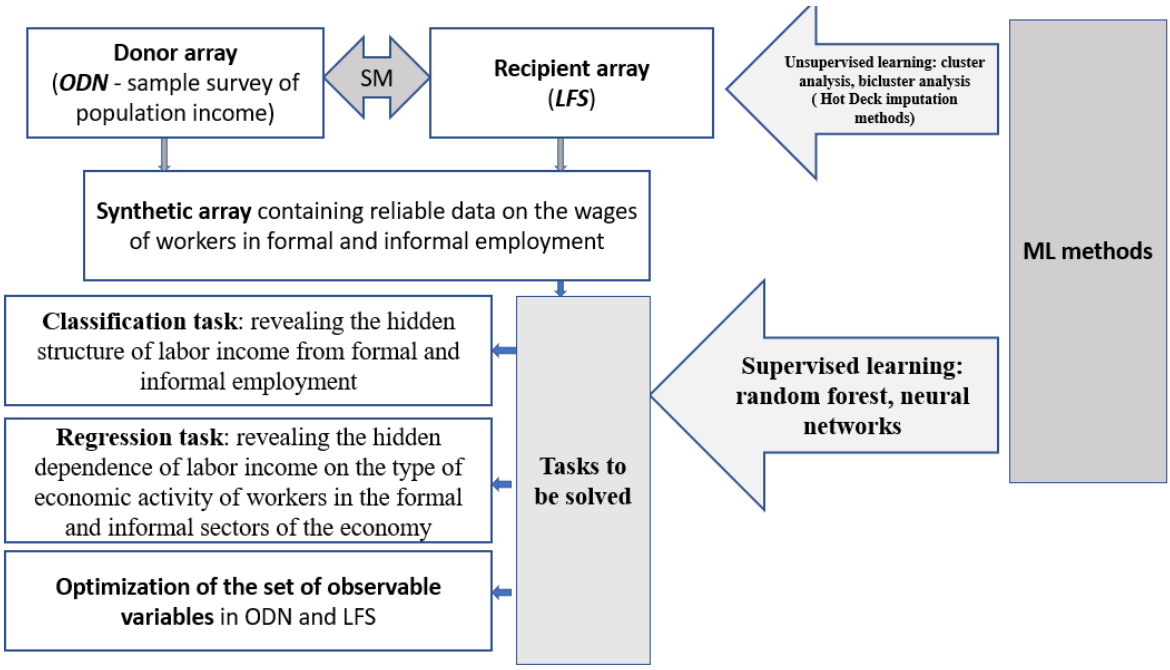

Fig. 1. New data obtained on the basis of aggregation of SM and ML methods for the processing and analysis of the integrated data set of sample surveys of official statistics of Russia (ODN and LFS).

Results

Practical results were obtained on the basis of ODN and LFS microdata arrays using the packages “StatMatch”, “randomForest”, “biclust” and some others of R system.

Figure 2 shows one of the results of the aggregation of the SM and ML methods for the purpose of analyzing the integrated data of the sample surveys of ODN and LFS. The results are presented on the example of one of the regions of the Russian Federation (Moscow city). It was found that the initial and final (identified on the basis of a synthetic array) ratio of income indicators of the population are significantly different. The statistical reliability of the final version (b) is much higher, since this array is much larger, and the integrated SM and ML methods applied to it made it possible to take into account the influence of hidden internal structural connections.

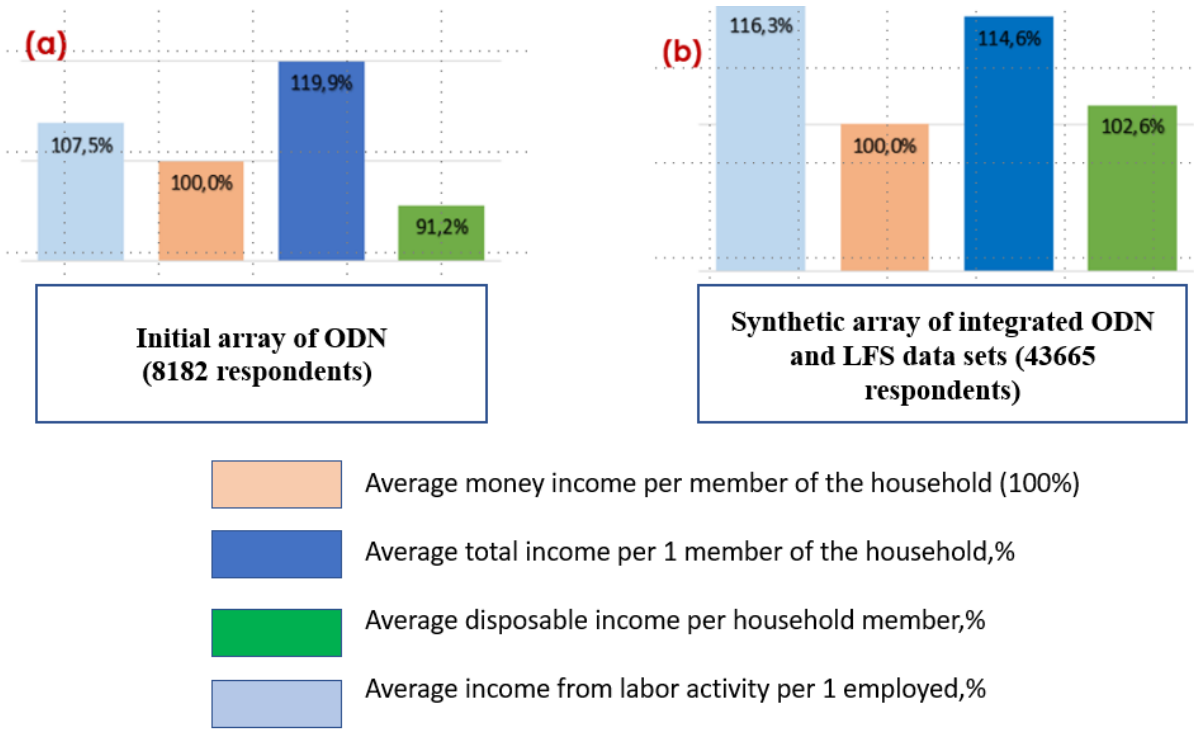

Fig. 2. Comparative results of the analysis of the ratio of indicators of incomes of the population, obtained on the basis of the initial data of the ODN and the integrated data of the sample surveys of the ODN and the LFS (Moscow city, 2019).

Discussion

The proposed approach is important for the development of official statistics as it has been tested and provided valuable new information about the sample survey microdata sets in the actual practice of official statistics in Russia.

The research carried out has practical value, requires further theoretical development, but at the same time identified several controversial points. Among them:

1) the extracted a priori unpredictable new knowledge when combining the application of the SM and ML methods to the survey results require interpretation, which increases the requirements for the theoretical knowledge of official statisticians; 2) the extracted a priori non-anticipated new knowledge requires an assessment from the standpoint of its usefulness for certain groups of users. This determines the need for them to be "advertised" by the national statistical institutes; 3) the issues of implementation of software systems containing modern and updated packages of SM, ML and other effective methods in the official statistical methodology require a special solution.

Nevertheless, the aggregation of SM and ML methods is a necessary tool for increasing the efficiency of national statistical services, as well as a methodological solution for the production of information when using new data sources in official statistics: administrative data and big data.

References

1.Serafino, P., Tonkin,R. Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey (2017). https://www.ine.es/e/essnetdi_ws2011/ppts/Leulescu_Agafitei_Mercy.pdf.

2. Statistical matching: a case study on EU-SILC and LFS (2019). Eurostat, European Commission, Luxembourg, https://www.gesis.org/fileadmin/upload/dienstleistung/daten/amtl_mikrodaten/europ_microdata/Abstracts_2019/Ahrendt_Leoncikas_Riob%C3%B3o.pdf

3. Statistical matching of EQLS and EU-SILC: A case study on public services. Daphne Ahrendt, Tadas Leoncikas, Irene Riobóo, Rey Juan, https://www.gesis.org/fileadmin/upload/dienstleistung/daten/amtl_mikrodaten/europ_microdata/Abstracts_2019/Ahrendt_Leoncikas_Riob%C3%B3o.pdf

4.HLG-MOS Machine Learning Project. https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project

5. The use of machine learning in official statistics – UNECE,

https://statswiki.unece.org