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

Elena Zarova, Dr., Professor, Plekhanov Russian University of Economics, ISI Council member (2021-2023)

Zarova.ru@gmail.com

Introduction

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+.

Method for analysis

Figure 1 shows comparisons of the distributions of employees by wage indicators, built according to the data of the Federal Tax Service (left) and the Federal State Statistics Service (Russia) (right) for two years: 2017 and 2021. The comparison is presented for two wage indicators: (1) wages of employees of enterprises (E) and (2) wages of employees with formal and informal (main) employment (ZNR). It can be concluded that these distributions, harmonized by the characteristics of units and by observable characteristics, differ significantly.

The data of the tax service on the wages of employees come from legal entities, individual entrepreneurs and registered self-employed, and the data of the national statistical service on wages are formed based on the results of the mandatory reporting of employees of large and medium-sized enterprises, reporting data from small enterprises that fell into the sample and on employees informal employment on the basis of a sample survey of households.

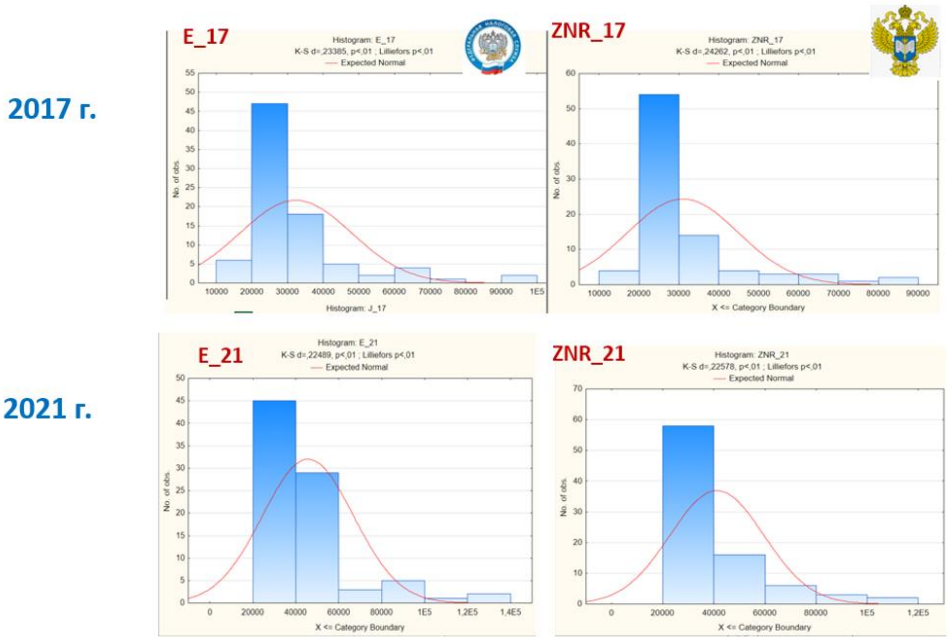
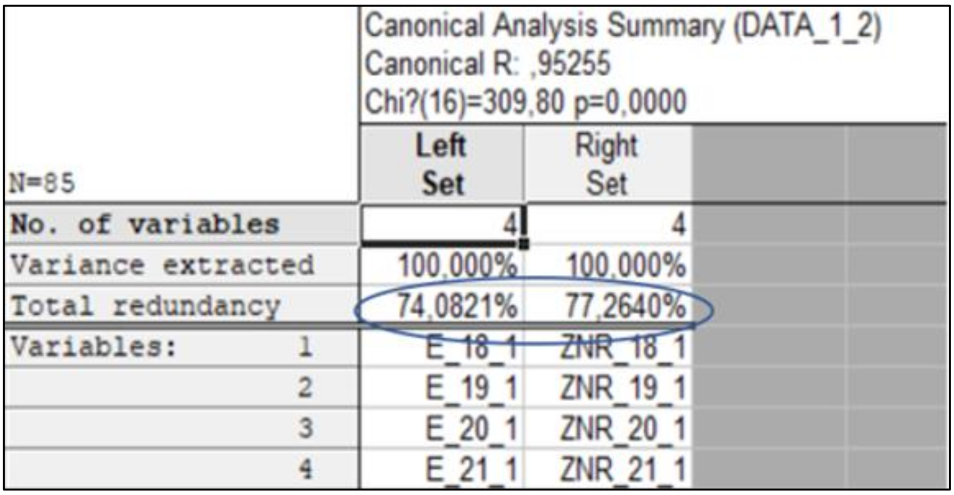


Fig. 1. Distributions of employees by wage indicators, built according to the data of the Federal Tax Service (left) and the Federal State Statistics Service (Russia) (right): (1) wages of employees of enterprises (E) and (2) wages of employees with formal and informal (main) employment (ZNR), 2017, 2021.

The use of similar data over a number of years shows a strong canonical correlation, and redundancy characteristics indicate that administrative wage data explains the distribution of official wage statistics by 74%. At the same time, the inverse conditionality of the distribution of tax data on wages by the corresponding distribution of official statistics data is by 77%. It can be concluded that official statistics are more reliable in this case.



Canonical Analysis Summary (DATA_1_2)
Canonical R: ,95255
Chi?(16)=309,80 p=0,0000

N=85	Left Set	Right Set
No. of variables	4	4
Variance extracted	100,000%	100,000%
Total redundancy	74,0821%	77,2640%
Variables: 1	E_18_1	ZNR_18_1
2	E_19_1	ZNR_19_1
3	E_20_1	ZNR_20_1
4	E_21_1	ZNR_21_1

Fig.2. The results of the evaluation of the canonical correlation of data from the tax service and data from the national statistical office on wages.

Conclusion

The use of similar data over a number of years shows a strong canonical correlation, and redundancy characteristics indicate that administrative wage data explains the distribution of official wage statistics by 74%. At the same time, the inverse conditionality of the distribution of tax data on wages by the corresponding distribution of official statistics data is by 77%. It can be concluded that official statistics are more reliable in this case. However, this conclusion needs to be supported by additional analysis. based on the data mining approach. Namely, the use of, for example, ensembles of trees of classification and neural networks to identify hidden relationships in both data sets to address the issue of their comparability and the possibility of integrated use in official statistics. The results of this work will be presented in a presentation at the 64th World Statistical Congress.

References

1. Use of administrative sources for business statistics purposes: Handbook on good practices (1999).
2. MIAD - Methodologies for an Integrated Use of Administrative Data in the Statistical Process
3. Use of administrative data for official statistics: The Global Perspective UN Statistics Division/DESA (2018).