UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Data imputation for the purposes of statistical research with the use data from administrative registers

Paweł Murawski (Statistics Poland, Poland)

P.Murawski@stat.gov.pl

I. Introduction

1. The possibilities of using administrative data to data imputation for the purposes of statistical research will be presented. Theoretical issues and an example of the application of this type of imputation will also be presented. The case study will concern research on Internet access in households. Sources of data from administrative registers used during imputation will be indicated, as well as the methods of their processing and integration with the list of population survey.

2. The hot-deck imputation method used in the study will be presented. The results of the performed imputation will also be discussed, as well as the results of the research itself, along with their characteristics. Conclusions that resulted from the implementation of the study will be presented. Possibilities of wider use of administrative data for imputation for statistical research purposes will also be indicated.

II. Administrative registers

A. Quality of registers

3. The most important issue to for the use of data from administrative registers is their quality. Only the highest quality data should be used for advanced operations, and in particular for data imputation.

4. Statistics Poland has a specialized unit that deals with improving the quality of data from administrative and non-administrative data sets for the needs of all units of official statistics. Thanks to this approach, all administrative files are standardized in a uniform way, which at the later stages of statistical production allows for full data consistency.

B. Data integration

5. High-quality administrative data enables their further use for statistical purposes, including data imputation. Therefore, it is extremely important to perform the fullest possible integration of selected data sets. It is therefore necessary to follow the same approach when standardizing all datasets used in the process, in particular donor and recipient datasets.

III. Practical application - Research on the use of information and communication technologies – Information Society in Poland 2019 (SSI)

A. The aim of the survey is to collect information on the use of the Internet by households

B. The first stage of the research implementation

6. Import of 3 input tables into the OBM data processing environment:
   a. DANE_UKE_ZASIEGI (44 673 175 records) – data provided by Office the Electronic Communications regarding the reach of Internet Service Providers
   b. KARTOTEKA_SSI (12 150 records) – lists of the units participating in the survey
   c. DANE_SSI10G (11 706 records) – data collected during SSI survey
7. Generating the integration key (key_num = individual building id) and then splitting the DANE_UKE_RASIEGI set into 2 tables:
   a) individual address to the building level + from the building (8 535 026 records);
   b) Data on the possible range of Internet Service Providers + building id (44 673 175 records).
8. Standardization of address data in the table with individual building addresses (8 535 026 records).:
   - upcase – upgrade to the standard of capital letters;
   - separation of the street prefix into a separate column;
   - control and improvement of the address code to the standard used in official statistics
   - address consistency check.
9. Standardization of address data in the SSI study file (12,150 records) analogous to building addresses. Also, generate a street code for records that have a street name. Control and marking of address consistency down to street level.
10. Multi-stage connection of the building id to lists of the units participating in the survey (12 150 records) by address data.

    Results:
    a. individual building id from the table UKE data ranges was added to 12,100 records;
    b. 50 records have not been attached: 49 - no building number in the table UKE data ranges, 1 - no city in the commune in the UKE table ranges.

    Generating a list of unique values (57 records) for the TECHNOLOGY variable from the DANE_UKE_ZASIEGI set.

    Generating unique values of the TECHNOLOGY variable in order to classify them according to the type of Internet connection - DANE_UKE_TECHNOLOGIE_RODZAJE_POL data set. it contains a list of TECHNOLOGIES provided by by Office the Electronic Communications along with their appropriate classification according to the type of Internet connection: kind_pol_st_mob (FIXED / MOBILE) and kind_pol_band (BROADBAND / NARROWBAND) (57 records, 4 variables).

C. The second stage of the research implementation

11. To the households from the DANE_SSI10G set, information features from the following sets were attached:

    a. KARTOTEKA_SSI - address data (including the added address key key_num)
    after the HOMEID = id key;
    b. DANE_UKE_RASIEGI - variable TECHNOLOGY after the key_num key;
    c. DANE_UKE_TECHNOLOGIE_RODZAJE_POL - variables kind_pol_st_mob, kind_pol_band after the key TECHNOLOGY

12. The number of unique non-integrated records after the key_num key was 44 (for 26 of these records the variable WAGA = 0). The non-integrated records were subjected to the imputation process in the scope of the variables kind_pol_st_mob and kind_pol_band.

2. Data imputation

13. The hot deck imputation based on the nearest neighborhood method was used for the given criteria for distinguishing imputation classes - variables WOJ and KLM.
14. In this method, the imputed value for the feature y for the kth object is

$$\hat{y}_k = y_{l(k)}$$

where l (k) is a donor for this object randomly selected from among all objects with a complete data record or from among those objects that belong to the same class imputation. The distribution of the value of the feature y after such a supplement to the missing data looks quite "natural", but it may still differ significantly from the distribution of the feature, which would be obtained if all the test units from the sample s answered the question relating to the variable y. This shows from the fact that respondents and non-respondents may differ in terms of parameters such as mean, standard deviation, etc.

15. The available types of internet connections were defined on the basis of the TECHNOLOGY variable from the DANE_UKE_ZASIEGI set and the variables nazwa_pol_st_mob and kind_pol_band from the DANE_UKE_TECHNOLOGIE_RODZAJE_POL set.
16. The data set was transposed so that one row of the table corresponded to one household (unique value of the variable NRG - household number) and the household was assigned all its Internet connections on one row (11,706 records).
17. The given weights for individual records, defined by the WAGA variable from the DANE_SSI10G set, were applied. These are weights adjusted for the level of contact and research implementation.

3. Results characteristics

18. According to the given guidelines, the possibility of connecting to the Internet was examined - the possibilities of Internet service providers, not the actual connection used.
19. It was assumed that due to the incompleteness of the set with UKE and the imputation of non-integrated records, all address points will be shown as "with the possibility of access to the Internet". Therefore, there will be no addresses "without the possibility of access to the Internet".
20. The number of surveyed households amounted to 12,635,252 and is consistent with the number of households with people aged 16-74 from the SSI 2017 survey.
21. There were no households with the possibility of access to the Internet through a permanent connection, only a narrowband one.
22. There were no households with the possibility of accessing the Internet through a narrowband only mobile connection.
23. There were no households without broadband Internet access.
24. It was found that the number of households with the possibility of access to the Internet through a permanent connection is equal to the number of households with the possibility of accessing the Internet through a permanent broadband connection.
25. It was found that the number of households with the possibility of access to the Internet through a mobile connection is equal to the number of households with the possibility of accessing the Internet through a mobile broadband connection.
26. It was found that the number of households with the possibility of access to broadband Internet is equal to the number of all surveyed households.
27. Internet access in households. Table in layout according to the class of the place of residence

Description		**Overall**	**Class of the place of residence**			
A – in absolute numbers			**Total cities**	**Cities with a population**		**Village**
B – in prcentage				**Over 100 000**	**Up to 100 000**	
Households with persons aged 16-74	A	12 635 470	8 475 205	4 307 610	4 167 595	4 160 266
	B	100,0	100,0	100,0	100,0	100,0
Including households:						
- with the possibility of access to the internet by permanent connection	A	11 756 575	8 354 410	4 273 342	4 081 068	3 402 166
	B	93,0	98,6	99,2	97,9	81,8
- with the possibility of access to the internet by a permanent broadband connection	A	11 756 575	8 354 410	4 273 342	4 081 068	3 402 166
	B	93,0	98,6	99,2	97,9	81,8
- with the possibility of access to the internet by a permanent narrowband connection	A	7 880 825	5 729 313	3 236 969	2 492 344	2 151 512
	B	62,4	67,6	75,1	59,8	51,7
- with the possibility of access to the by a permanent connection, only narrowband	A	35 234	0	0	0	35 234
	B	0,3	0,0	0,0	0,0	0,8
- with the possibility of access to the internet by mobile connection	A	12 296 792	8 245 373	4 230 305	4 015 068	4 051 419
	B	97,3	97,3	98,2	96,3	97,4
- with the possibility of access to the internet by mobile broadband connection	A	12 296 792	8 245 373	4 230 305	4 015 068	4 051 419
	B	97,3	97,3	98,2	96,3	97,4
- with the possibility of access to the internet by mobile narrowband connection	A	8 406 736	5 757 427	3 248 172	2 509 256	2 649 309
	B	66,5	67,9	75,4	60,2	63,7
- with the possibility of access to the internet by only mobile narrowband connection	A	x	x	x	x	x
	B	x	x	x	x	x

28. Internet access in households. Table in layout by household income groups after calculating the quartile ranges

Description		**Overall**	**Class of the place of residence**			
A – in absolute numbers			**Total cities**	**Cities with a population**		**Village**
B – in prcentage				**Over 100 000**	**Up to 100 000**	
Households with persons aged 16-74	A	12 635 470	2 906 221	3 393 465	3 229 810	3 105 975
	B	100,0	100,0	100,0	100,0	100,0
Including households:						
- with the possibility of access to the internet by permanent connection	A	11 756 575	2 710 204	3 147 523	2 957 174	2 941 675
	B	93,0	93,3	92,8	91,6	94,7
- with the possibility of access to the internet by a permanent broadband connection	A	11 756 575	2 710 204	3 147 523	2 957 174	2 941 675
	B	93,0	93,3	92,8	91,6	94,7
- with the possibility of access to the internet by a permanent narrowband connection	A	7 880 825	1 767 153	2 072 391	2 025 410	2 015 872
	B	62,4	60,8	61,1	62,7	64,9
- with the possibility of access to the by a permanent connection, only narrowband	A	35 234	8 806	6 663	14 822	4 943
	B	0,3	0,3	0,2	0,5	0,2
- with the possibility of access to the internet by mobile connection	A	12 296 792	2 822 860	3 311 364	3 150 505	3 012 063
	B	97,3	97,1	97,6	97,5	97,0
- with the possibility of access to the internet by mobile broadband connection	A	12 296 792	2 822 860	3 311 364	3 150 505	3 012 063
	B	97,3	97,1	97,6	97,5	97,0
- with the possibility of access to the internet by mobile narrowband connection	A	8 406 736	1 877 440	2 243 761	2 185 361	2 100 174
	B	66,5	64,6	66,1	67,7	67,6
- with the possibility of access to the internet by only mobile narrowband connection	A	x	x	x	x	x

4. Conclusions

29. A similar pattern of action can be applied to other similar studies. Thanks to this, both the process of calculating the results and imputation of data will be standardized, which contributes to increasing the effectiveness of the activities carried out, accelerating the implementation of processes and improving the quality of the resulting data. It also lowers the costs of carrying out the research due to the fact that the test sample does not have to be increased in order to maintain the quality of the results at a sufficiently high level.

30. Due to the availability of high-quality data from administrative sources, data imputation may be more precise, which will enable output data to be generated at the lowest aggregation levels.

5. References

M.Szymkowiak (2018) Imputation and calibration - new estimation possibilities in statistical research with no answers in *Statistics In Management* pp.95