SELF PRESENTATION PAPER

OF

DOCTORAL DISSERTATION

The use of imputation techniques in estimating statistical output information and in the analysis of data structure in business statistics

Author:

Paweł Lańduch

Supervisor:

Andrzej Młodak – Associate Professor, Calisia University – Kalisz, Poland

Warszawa, maj 2023 r.

1. Objectives of the thesis

The principle of using data sources, other than information from respondents, in order to harness them to produce statistical surveys results, is one of the objectives of official statistics. In the case of business statistics, response burden is, among others, one of the problems. Although the participation in business surveys is mandatory, the problem of non-response increases. Additionally, due to the characteristics asymmetric distribution in the business population, large firms are included in virtually every business survey, increasing its administrative burden. On the other side, new data sources are emerging, such as big-data sets. Those kind of data could replace traditional sources, i.e. sets based on responses from establishments. In the case of short-term statistics one could mention another problems. Among them are the dynamic of the business population and the costs of conducting research (because of this micro firms are omitted, i.e. those with less than 10 employees). An important element of statistical survey is its quality. The concept is multidimensional. One of the most important qualitative aspects is the precision of the results. Applying non-statistical data sources for producing statistical results therefore requires the developments of methods, that would correspond to qualitative criteria as regards survey outcomes.

The integration of surveys and different data sources, including administrative registers and big data files, in order to produce statistical results, are the research field, that has long been a European Statistical System area of interest. However, the probability sample remains the gold standard in statistical surveys. In this dissertation the problem of using mass imputation for estimating variables totals in business surveys is undertaken. This is such a case of integration in which a probability sample and a non-probability sample are combined. This kind of integration is a new, emerging area of statistical research. This is the case of integration in which the response variables is only observed in a non-probability sample. Whereas, these variables will be transported to the non-probability sample by using common variables present in both data sources. Then, the analysis is carried out in a probability sample using available first order probabilities. Mass imputation, although it has been used previously, in the two-phase sampling context, has not been yet sufficiently studied as a method of combining a probability sample and a non-probability sample, i.e. used in the account of this paper.

The main objective of the work is to assess the effectiveness of the use of mass imputation methods, based on the integration of probability and non-probability samples, in the statistical business surveys.

The main objective of the analysis was achieved by the realization the following specific objectives:

1. Determination of global values of both analyzed features based on the Horvitz-Thompson estimator, assuming that their values are observed in a probability sample.
2. Determination of global values of both analyzed features based on the Horvitz-Thompson estimator after previous transmission of feature values to a probability sample by means of mass imputation based on the nearest neighborhood. To define the distance, an auxiliary variable is used, namely the number of employees in the enterprise. The donor is selected within given classes, where the class is a set of units whose type of activity belongs to the same division of the Polish Classification of Activities.
3. Determination of global values of both analyzed features based on the Horvitz-Thompson estimator after earlier transmission of feature values to a probability sample by means of mass imputation of the average based on the nearest neighbors. This is a different case from the one discussed in the first point, because the imputed value here is the average value of the feature for the designated subset of units closest to the given unit. As in point 2, the nearest neighbors are limited by a subset of units with the same type of activity in relation to the Polish Classification of Activities division.
4. Determination of global values of both analyzed features based on the Horvitz-Thompson estimator after previous transmission of feature values to a random sample by mass imputation carried out using a local regression model in a non-random set. Regression models were designated within classes for a given type of activity of the unit.
5. Determination of global values of both analyzed features based on the Horvitz-Thompson estimator after previous transmission of feature values to a probability sample by mass imputation, carried out using the spline regression models. Relevant models were designated separately for each class of units, with the class being units with the same type of activity, i.e. a division of the Polish Classification of Activities.

In accordance with the objectives set, the following main research hypothesis was formulated:

Mass imputation as a method of data integration, in terms of combining probability and non-probability samples, can be a useful tool – supporting or main – for obtaining main, fast estimates of comparable quality compared to business surveys based on a random and stratified

sample, when we assume that the values of the analyzed characteristics are observed only in the non-probability sample and can be taken from there.

Taking into account the above hypothesis, an additional hypothesis was formulated:

The integration of data using mass imputation can also be used in a comparable dimension regarding the quality of estimates, in a scheme of sampling units proportional to their size with a smaller sample size, but ensuring a greater representation of smaller units. In other words, based on imputation methods and the use of a proportional to size design sampling, the size of the sample could be reduced or the transition to lower levels of aggregation could be attained without increasing the sample. Currently, in the DG-1 survey conducted by public statistics, all large entities (employing 50 or more people) and 10% of medium-sized entities (10-49 employees) in each layer are included in the sample. Using mass imputation, for example, the threshold for dividing units into medium and large units (e.g. from 50 to 250 employees) could be raised, which would lead to a reduction in the sample size while maintaining high quality results.

2. The formulated problem and the assumptions of the analysis.

The problem with non-probability data source is, among other things, that the mechanism of selecting units for it is usually unknown. Treating this kind of a data source as a probability sample can induce bias. Secondly, to estimate the propensity score participation require auxiliary information at the population level. In this thesis mass imputation will be used as a integration of a probability sample and a non-probability sample in such way where the response variable is observed only in the non-probability sample, i.e. the data for this variable are imputed for all records. The non-probability sample will be the donor data set, the training set with the use of which mass imputation is performed.

Let $\mathcal{F}_N = \{(X_i, Y_i): i \in U\}$, where $U = \{1, \dots, N\}$ and $X_i = (X_i^1, \dots, X_i^p)$ is a p-dimensional vector of covariates, whereas Y_i is the study variable. The objective is to estimate the global value of Y. Additionally, it is assumed, that $\mathcal{F}_N$ is a random sample from a superpopulation ζ and N is known. The goal is to estimate a global value of parameter $Y = \sum_{i=1}^{N} g(Y_i)$ for a known function $g(\cdot)$. It is assumed that two kinds of data set are available: a probability sample denote as A and another source of data, denote as a "large" file or a big data set, in which the inclusion probability cannot be determined, denote as B. The table 1 describes the initial situation. The data set O_A contains information from a sample A:

$O_A = \{(d_i = \pi_i^{-1}, X_i): i \in A\}$ where the sample size A is $n = |A|$ and the inclusion probabilities are known, that is $\pi_i = P(i \in A)$

The data set O_B, in turn, contains information from a non-probability sample B: $O_B = \{(X_i, Y_i): i \in B\}$ where the size of the sample is $N_B = |B|$

Table 1. Schematic presentation of two sources of data before their integration.

Sample type	Unit number in the sample	Weights $d_i = \pi_i^{-1}$	Covariates X	Study variable Y
Probability sample O_A	1	✓	✓	?
	.	.	.	.
	.	.	.	.
	.	.	.	.
	n	✓	✓	?
Big data sample O_B	1	?	✓	✓
	.	.	.	.
	.	.	.	.
	.	.	.	.
	n	?	✓	✓

A sign ✓ indicates the presence of data and a sign ? means missing data

Source: Yang, S., Kim, J. K., Hwang, Y., (2021) Integration of data from probability surveys and big found data for finite population inference using mass imputation, *Survey Methodology, Vol. 47, No. 1, pp. 29-58, Statistics Canada, Catalogue No. 12-001-X*

Assumption for mass imputation:

Let $f(Y|X)$ be a conditional density function of Y given X from a superpopulation model ζ. Let $f(X)$ and $f(X|\delta_B = 1)$ be the density function of X in a finite population and in the sample B respectively, where δ_B is the indicator of the selection of a unit to sample B. The following two assumptions are made.

Assumption 1.

The conditional density of Y given X in the sample B follows the superpopulation model, that is:

$$f(Y|X;\delta_B = 1) = f(Y|X)$$

The assumption 1 implies that the selection mechanism in the data set B can be ignored conditionally with respect to the vector of variables X. This assumption leads to the conclusion that $P(\delta_B = 1|X,Y) = P(\delta_B = 1|X)$. This equality will result in fact that missing data in the explained variable are determined by the MAR (missing at random) mechanism.

Assumption 2.

For a vector of covariates $X \in \mathbb{R}^p$ there are constants C_l and C_u such as $C_l \leq f(X)/f(X|\delta_B = 1) \leq C_u$ almost surely.

The assumption can also be formulated as $P(\delta_B = 1|X) > 0$ for every value of X. Assumption 2 would not be fulfilled if there were units that would never be selected for the sample B. The judgement of the fact belongs to a subject matter specialist. The comparison of the results of the application of the methods of mass imputation were done with a direct Horvitz-Thompson estimator. This means, that is assumed, if values of a variable Y are observed in sample A then the population global value is:

$$\hat{Y} = \sum_{i \in A} \pi_i^{-1} Y_i.$$

The variance of the estimator is determined using Deville formula. It is useful in practice variance estimator, especially in a sample design with different probabilities of inclusion. In order to estimate the variance the second order probabilities are not required. Variance of the estimator is expressed by the formula:

$\widehat{var}(\hat{Y}) = \sum_{i \in s} \frac{c_i}{\pi_i} (Y_i - \hat{Y}_i^*)^2$, where

$\hat{Y}_i^* = \pi_i \frac{\sum_{i \in s} \frac{c_i Y_i}{\pi_i}}{\sum_{j \in s} c_j}$, where in turn $c_i = (1 - \pi_i) \frac{n}{n-1}$.

3. Empirical analysis

For the purposes of the analysis an artificial population of establishments was generated. As a reference structure, the frame of a short term survey that is permanently conducted for many years by the Central Statistical Office was used to generate the set. This is a survey called – The short report about business activity, carried out on a monthly basis. Relevant data are collected from establishments by using the questionnaire denoted – DG-1. The subjective range of the survey is the enterprise sector. Since the enterprise sector is very wide (basically the entire Polish Classification of Activities 2007) for the analysis a very important of it was selected, namely industry (the following type of activities belong to the industry category: mining and quarrying, manufacturing, electricity, gas, steam and air conditioning supply, water supply; sewerage, waste management and remediation activities). From the thus generated population two data sources were drawn: a probability sample and a non-probability sample, that is the sample is not design based probability sample. A probability sample is a stratified sample with different probabilities of inclusion of units. Due to the skewness of variables distribution in business population a proportional to unit size sampling was used. The strata of sampling was generating based on the PKD second level, that is the division level. Therefore, the PKD division is assumed as a sampling layer. The non-probability sample was generated also as a strata based, where stratum was PKD division level but the inclusion to the sample was determined by logistic linear model in one version, while in the other – also on the logistic model, nut in a non-linear version. For the analysis two variables from the survey questionnaire was chosen, namely:

- Sw_1b – net revenues from the sales of the products (products and services of own production)
- Wb_1b – gross salaries for the average number of employees.

The number of employed persons was used as a matching variable. This number of the employed persons is available for every unit in the survey frame. After the transmission of the values of the characteristics of units from the nonprobability sample to the probability sample the global values were estimated by using direct Horvitz-Thompson estimator, taking the probability of inclusions from the adopted sampling scheme, and also its variance was determined. The analysis of the four methods of mass imputation were conducted: two non-parametric and two semi-parametric. The following mass imputation methods were used (the first two are nonparametric):

- Mass imputation based on the nearest neighbor selected at random: for each unit in the probability sample, using a common variable that occurs in both data sources, that is the number of persons employed, in the nonprobability sample the subset of units is found with the smallest distance from it. To determine the distance the Euclidean metric was used. If the set contains more then one unit, the unit is randomly chosen. Then the value of the considered characteristic of the chosen unit is imputed for the target unit in the probability sample. After the transmission, the global value is estimated by using a Horvitz-Thompson estimator in the probability sample.
- The method of imputation of the average based on the nearest neighbors: it differs from that described in the previous section in such a way that instead of randomly selecting a given unit from the subset of units for which the distance from the unit $i \in A$ is the smallest, the average value of the examined feature for this set is calculated. This calculated value is then imputed to the right place.
- The mass imputation method based on local regression: to conduct the imputation was used the method defined as local regression or local polynomial regression. This is a generalization of the moving mean and polynomial regression. A weighted polynomial regression model is used here, where the weight is determined by the α parameter ($0 < \alpha \leq 1$), i.e. at a given model matching point, only information distant less than α from it are taken for account. An iteration method was used to resolve the parameter α. The linear model was used. The explanatory variable is always the number of persons employed. The value obtained from the model is transferred to the probability sample.
- Mass imputation method based on spline modelling - it is assumed, that the aim is to find such a function f that would represent the dependence of the outcome variable Y and the independent variable X. A domain is partitioned into K separate ordered intervals called knots. In each interval classical regression methods are involved. Among the most popular splines are cubic splines with no interior knots and boundary knots at the range of the X variable. In this analysis applied those kind of splines were adopted.

On the basis of 500 conducted simulations for both chosen variables in every strata, that is the mentioned divisions of the PKD 2007 classification as well as for each of the method separately for each variable, the values of the coefficient of variation was determined according to the formula:

$$CV_h = \frac{1}{500} \sum_{i=1}^{500} \frac{\sqrt{var(\bar{Y}_{h\,i})}}{Y_{h,i}} * 100, \text{ where } h = 1, \ldots, 25.$$

4. The results of the analysis

Analyzing the quality of all four used mass imputation methods can be spotted, that the results measured by coefficients of variation have values close to that produced from the gold standard, that is the case when the response variables are observed in a probability sample. The range of values of coefficients of variation is from 0,2 % to 14,9 % if all ranges are taken into account and for both analyzed variables in both versions of integration (i.e. generating a non-probability sample based on a linear and nonlinear model). Whereas, for the gold standard the scope was from 1,1 % to 11,6 %. Based on these results it can be seen that these ranges are comparable. The values of the coefficients based on mass imputation starts at a level impossible to achieve in the gold standard, although in the upper range it slightly exceeded it. Low layer counts had an impact on higher values of the coefficient of variation. The size of the probability sample was based on the size used in the actual Dg-1 business survey, that is 10 % of the medium size units in the frame stratum. Summing up the conducted considerations it can be seen that is difficult to indicate the advantage of one of the analyzed mass imputation method over another in the context of the level of their quality expressed by relative coefficients of estimation of global values of the analyzed variables. Table 2 presents the results of the coefficients for variable Sw_1b in case of integration of probability and non-probability samples generated in linear version.

Table 2. Values of coefficients of variation achieved using different mass imputation methods in the case of integration of a probability sample and a non-probability data source, generated in a linear model.

No.	PKD Division (layer)	HZ	KNS	NS	LREG	SPL
1	5	3,9	6,6	6,6	6,5	6,3
2	10	1,2	0,9	1,2	0,5	0,6
3	11	7,7	7,4	7,6	4,7	5,4
4	13	3,0	2,5	3,2	1,2	1,8
5	14	3,4	1,9	3,5	0,5	0,5
6	15	5,4	4,1	5,4	1,7	2,9
7	16	2,2	1,6	2,2	0,9	0,9
8	17	4,7	4,7	4,9	3,3	3,1
9	18	3,6	2,9	3,6	1,6	2,4
10	19	4,1	7,7	7,7	8,3	6,8
11	20	3,3	3,0	3,3	2,2	2,4
12	21	11,6	12,6	12,6	5,3	4,0
13	22	1,7	1,4	1,8	1,0	1,1

No.	PKD Division (layer)	HZ	KNS	NS	LREG	SPL
14	23	2,1	1,7	2,1	0,8	1,6
15	24	2,7	3,1	3,2	2,2	4,1
16	25	1,6	1,1	1,6	0,7	0,5
17	26	2,8	2,7	2,9	1,8	1,2
18	27	2,4	2,3	2,4	1,7	2,0
19	28	2,4	2,1	2,4	1,8	1,4
20	29	3,6	3,7	3,8	2,9	2,1
21	30	5,2	5,2	5,3	2,9	4,1
22	31	1,9	1,4	1,9	1,0	1,1
23	35	1,1	1,1	1,1	1,0	2,2
24	36	5,8	5,1	6,0	2,0	2,8
25	38	3,6	2,6	3,6	0,6	1,0

Note: column HT — variables observed in the random sample; KNS column – mean based imputation method based on nearest neighbors; NS column – mass imputation method based on nearest neighbor; LREG column – mass imputation methods based on the local regression model, and SPL column – mass imputation method based on spline model. The analyzed variable is the Sw_1b variable – net revenues from the sales of products, i.e. its global value for individual domains, where the domain is a division (second level) of the PKD classification.

5. Summary

Business surveys have their own unique features – to mention only the asymmetry of distribution of variables, the complexity of statistical units, intricate, technical definitions, difficult to find in non-statistical sources of data. Additionally, timeliness is often given priority over quality. Despite this, the possibility of using non-statistical data sources, even as a auxiliary source of data, would give new possibilities. In view of the growing number and scope of new, non-statistical data sources as well as increasing users expectations new methods of data integration - including those based on mass imputation should be taken into account as a useful tool to support production of the statistical output in the face of increasing user expectations.

In this work, the problem of using mass imputation methods in statistical production of business surveys totals is addressed. The mass imputation was applied as a method of data integration. Such an approach is not currently used in official statistics in Poland. The application of methods based on mass imputation as an integration of a probability sample and a non-probability data set is an emerging field of research. The usage of such methods could

enrich the possibilities of using e.g. administrative registers in statistical surveys, the analysis of which is carried out in this respect.

The extremely important element of small and micro business population is its large number and volatility. If mass imputation could be adopted for integration of such sources of non-statistical data, that contain such units, then this could enrich business surveys with small units which, due to cost, are not included in short-term business surveys.