Jacek Kowalewski
Statistical Office in Poznań

Model of optimization of statistical surveys

Dynamical development of information economy and information society is related to an appearance of many various challenges. One of them is a quickly growing demand of wide set of various users for information. The fulfilment of this demand is connected with undertaken trials of collecting data using various methods (e.g. by sampling surveys or usage of administrative registers). Both the variety of ways leading to required information receipt and the necessity of reduction of respondents' burdening result in the complex coordination of statistical surveys and are connected with a dilemma concerning choice of optimal method of their collection.
The process of optimization of statistical surveys requires a creation of some map of the full information system, including precise definitions of the whole set of input information, recognition of possible sources of data collection and available methods of transformation.
This article is a trial to precise a theoretical model, which would allow to performance of optimization of the process of statistical surveys.

1. Premises of the model

The premies of the model of survey are as follows:

1. It is possible to define the ***Y*** vector containing the variables the values of which are the results of the whole statistical survey process. We shall refer to this vector as **the output vector**. Hence the basic purpose of a statistical survey process is to estimate the *n* value of various output variables.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} \tag{1}$$

2. The Y output variable values are estimated directly upon **the input variables** (or primary variables) which will be represented by an *m*-member *X* vector.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_m \end{bmatrix} \tag{2}$$

3. The process of converting the *X* input variables into the *Y* output variables shall be referred to as **the transformation process** of data and denoted by Ω.

$$\Omega : X \times \dots \times X \rightarrow Y \tag{3}$$

The transformations may have varying character and the value of a specific y_i output variable can be estimated by transforming one or more input variables.

$$y_i = f(X) \tag{4}$$

4. There can be many methods of estimating the same y_i output variable upon different input variables. It is assumed that there is a finite number of transformations (qi) leading to the estimation of the i-th input variable which can be described in the set of Sy_i (Fig. 1.)

$$Sy_i = \{f_1^i(X),.., f_{qi}^i(X)\} \tag{5}$$

where:

$f_k^i(X)$ denotes the k-th method of estimating the y_i output variable based on the X vector of input variables.

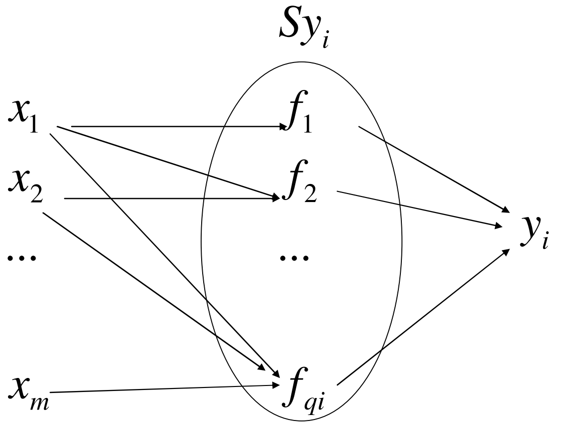

Fig. 1. An example of estimating the i-th output variable
Source: Own development

5. The estimations of the X input variables are the result of **surveys**. It is assumed that there is a finite set of r possible and different surveys, which can be represented as a vector:

$$B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \dots \\ b_r \end{bmatrix} \tag{6}$$

A b_l survey is an isolated and uniform data acquisition process. It can be a complete survey, a sample survey or exploitation of administrative registers or other sources of data.

6. Relation Φ of estimating the X input variables upon the data from the B surveys shall be referred to as **the data acquisition process**, which can be denoted as follows:

$$\Phi : B \to X \tag{7}$$

It is assumed that estimation of the x_j variable value is conducted directly in the survey (without further transformations). It is also assumed that it is possible to estimate the same specific x_j variable in different surveys (Fig. 2).

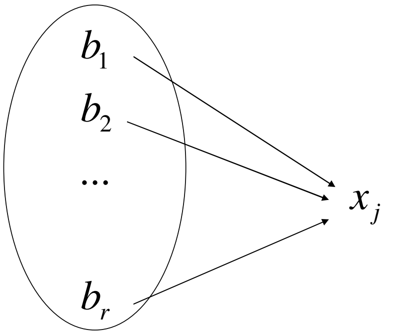

Fig. 2. The method of estimating the *j*-th input variable upon various surveys
Source: Own development

7. For every *l*-th b_l survey, the following can be assigned:
- KC_l – a constant denoting total survey cost independent of the amount of data acquired in the survey,
- bx^l – an *m*-member vector defining the variables acquired in the *l*-th survey

$$bx^l = \begin{bmatrix} bx_1^l \\ bx_2^l \\ bx_3^l \\ \dots \\ bx_m^l \end{bmatrix} \tag{8}$$

where:

$$bx_j^l = \begin{cases} 1 & \text{, when the } x_j \text{ input variable value is estimated in survey } b_l \\ 0 & \text{, in an opposite case.} \end{cases} \tag{9}$$

8. It is assumed that every bx_j^l variable can be uniquely assigned with the be_j^l coefficient, which describes the quality of estimating the x_j variable in a b_l survey. For the sake of this work it is assumed that the coefficient in question will have the following values:

$$be_j^l \geq 0 \tag{10}$$

The smaller is the coefficient in question, the higher is the estimation quality of the *j*-th variable in the *l*-th survey. It is assumed that the coefficient is relative, i.e. it conveys quality irrespective of the estimation level.

9. The estimation quality vectors are isolated for both input (*Xe*) and output (*Ye*) variables.

$$Xe = \begin{bmatrix} xe_1 \\ xe_2 \\ xe_3 \\ \dots \\ xe_m \end{bmatrix} \qquad Ye = \begin{bmatrix} ye_1 \\ ye_2 \\ ye_3 \\ \dots \\ ye_n \end{bmatrix} \tag{11}$$

10. The costs of data transformation processes (Ω) are negligible and thus omitted in the model.

2. The mathematical model of statistical survey process

Basing on the premises above we may propose a model of statistical survey process in the form of a two-stage problem represented in the diagram below (Fig. 3).

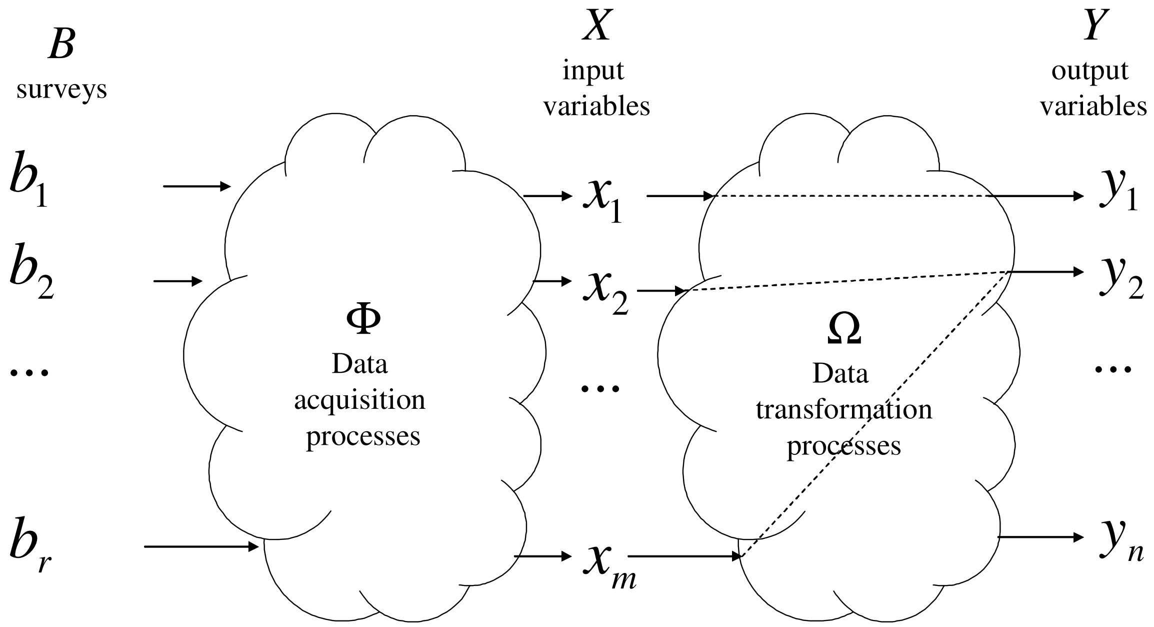

Fig. 3. The statistical survey model
Source: Own development

Solving the problem comes down to establishing the organization of data acquisition processes (Φ) and data transformation processes (Ω) in a way that allows an effective estimation of the input variables vector (*Y*). It is worth noting that estimating a part of assumed input variables (*X*) may become unnecessary and that not all of the surveys planned (*B*) will have to be conducted.

The quality of estimated variables (*E*).

The key value of the optimization process is the quality of output values (*Y*). Upon premises (7) - (9), the purpose criterion which we are interested in may be ultimately formulated as follows:

$$E = \sum_{i=1}^{n} ye_i \tag{12}$$

where:

ye_i is the estimation quality coefficient of the y_i output variable,
n is the number of estimated output variables.

The criterion above allows only to determine the total 'averaging' quality of the whole statistical survey process. It may turn out that in case of specific variables the estimation quality can be different and at variance with the expectations. That is why it seems rational to include an additional criterion (*Ey*):

$$Ey = \max_{i=1,..,n} ye_i \tag{13}$$

where ye_i and n are defined as in equation (12).

The whole problem may be relayed in short as determining the following set:

$$Z = \{Y, BX, Sy^*\} \quad ,$$

which will provide the following:

$$\begin{aligned} &K \rightarrow \min \\ &E \rightarrow \min \\ &Ey \rightarrow \min \end{aligned} \tag{14}$$

where: *K* – total survey costs,
E – total survey process quality (i.e. estimate error),
Ey – the maximum estimate error for an individual output variable.

It is a two-stage multi-criterion problem which can be solved with the methods of multi-criterion programming. Optimization of the first stage - data acquisition processes (Φ) - depends on solving a binary problem, while the second stage - the data transformation processes (Ω) - is a combinatorial problem.

References

[1] Galas Z., Nykowski I., Żółkiewski Z., Programowanie wielokryterialne, PWE Warszawa 1987

[2] Kordos J., Jakość danych statystycznych, PWE, Warszawa 1988

[3] Oleński J., Ekonomika informacji. Metody. PWE, Warszawa 2003

[4] Roy B. Wielokryterialne wspomaganie decyzji, WNT Warszawa 1990

[5] Zarkovich F., Quality of Statistical Data, FAO, Rome 1966