The missing data problem in statistical surveys on the example of the SOF-1 report (Report on the activities of foundations, associations and similar social organisations)

Barbara Pawełek[1], Tomasz Sekuła[2]

Abstract

The problem of missing data is often encountered when carrying out statistical research. To properly address this issue, the factors contributing to the lack of data must be investigated. The analysis is based on data from a study conducted for 2014 via the SOF-1 Reporting Form. The purpose of this research is to establish the level of economic and social resources of the non-profit organisations operating in Poland and to provide a description of their activities. The aim of this work is to present the results of empirical research into existing correlations between the missing answers in the SOF-1 report and the characteristics of the research subjects. The added value of the work is to demonstrate that the method used to visualise the distribution of missing data in a database and the association rules can be successfully applied to the analysis of correlations between the missing answers in the SOF-1 report and the profile of the respondents. The identified correlations may help researchers achieve a higher level of completeness for the SOF-1 report in terms of answers given. The potential benefits of this approach include reduced study costs and more accurate generalisations. The findings of this analysis can be used to identify mechanisms which contribute to generating missing data and thus to select the right method for estimating missing information.

Keywords: *non-profit organisations, missing data, visualisation methods, association rules*

JEL Classification: *L310, C810, C100*

1. Introduction

The problem of missing data is often encountered when carrying out statistical research. A number of different reasons for missing information in statistical databases are reported by the literature (Eaton et al., 2005). The incomplete nature of data sets poses a major obstacle to conducting an effective statistical analysis, assessing the status and dynamics of the studied phenomenon and making rational decisions. For this reason, the public institutions involved in collecting statistical data strive to ensure a high level of completeness of the data sets intended for use by state and local government authorities, businesses and research centres.

The present study involving the SOF-1 Reporting Form, *Report on the activities of foundations, associations and similar social organisations*, is part of a wider research into the non-profit sector carried out by the Statistics Poland (*Główny Urząd Statystyczny*, GUS). The SOF-1 report is intended to address the increasing importance of civil society, provide data to assess the progress made in implementing public policies which support social economy and social capital and to monitor the activities of non-profit organisations in Poland.

[1] Corresponding author: Cracow University of Economics, Department of Statistics, 27 Rakowicka St., 31-510 Cracow, Poland, e-mail: barbara.pawelek@uek.krakow.pl.

[2] Statistical Office in Krakow, Research Centre for Social Economy, 3 Kazimierza Wyki St., 31-223 Cracow, Poland, e-mail: t.sekula@stat.gov.pl.

Once the entities (selected for research) which gave incomplete answers in the SOF-1 report are identified, we contact them to obtain the missing answers. Contact is made by phone or via e-mail. One of the objectives of the research undertaken by the Statistical Office in Cracow and the Department of Statistics at Cracow University of Economics is to define the profile of the entities/respondents who failed to provide certain answers in the survey conducted using the SOF-1 Reporting Form.

The aim of this work is to present the selected results of empirical research into existing correlations between the missing answers in the SOF-1 report and the characteristics of the research subjects. The added value of the work is to demonstrate that the method used to visualise the distribution of missing data in a database and the association rules can be successfully applied to the analysis of correlations between the missing answers in the SOF-1 report and the profile of the respondents.

To our knowledge, there is no previous record of the association rules being applied to the analysis of correlations between the missing answers and the profile of the respondents. Likewise, we are not aware of any reported case of the visualisation methods for distribution of missing data in a database and the association rules being used to examine the activities of foundations, associations and similar social organisations.

2. Data

Empirical data for analysis were collected via the SOF-1 Reporting Form as part of the study carried out in Poland for 2014 (database status as of May 20*th* 2015).

The SOF-1 report was targeted at selected non-profit organisations. According to the definition provided in the United Nations' Handbook on Non-Profit Institutions (*Satellite Account…*, 2018), the non-profit sector includes all entities that are: institutionalised to some extent (i.e. following registration with a competent public office) or have some degree of persistence of goals, structure and activities; institutionally separate from government; non-profit making (do not exist primarily to generate profits and do not return profits generated to their owners, employees, etc.); self-governing (able to control their own activities); involve some meaningful degree of voluntary participation.

The SOF-1 report was used to gather information on the profile of selected non-profit organisations, determine their economic and social potential, development strategies and forms of activity and customer profiles. The form was divided into eight sections:

- field of activity, scope, form and nature of activity the reporting entity's (Section I),
- participation in the projects implemented under the European Social Fund (Section II),
- the type and number of beneficiaries of the activities carried out by the reporting entity and the type of activities undertaken for the benefit of individual consumers (Section III),
- members and volunteers, their social work (Section IV),
- employed on the basis of employment contract and civil law contracts (Section V),
- operating revenues and expenses (Section VI),

- fixed assets held, capital expenditure incurred and in-kind gifts received (Section VII),
- operating conditions (Section VIII),
- contact details.

The Database of Statistical Units (*Baza Jednostek Statystycznych*, BJS) served as a sampling frame for the statistical research involving the selection of non-profit organisations having legal personality, having a legal form which meets the target group criteria for the study and which are recorded in the BJS as having a specific legal and economic status of: an active entity engaged in an activity; an active entity not engaged in an activity, but preparing to engage in one.

The entities were selected for inclusion in statistical records using the following methods:

- purposive sampling was used for a highly varied sample of entities. This type of sampling was applied to: entities with registered offices, entities registered as Public Benefit Organisations, project promoters for the European Social Fund (ESF), entities employing more than 5 people, denominational entities engaged in social activities;
- stratified random sampling with proportional allocation was used for other entities. The strata represented Polish voivodships with a focus on the largest cities: Warsaw, Cracow, Lodz, Wroclaw and Poznan; legal form: foundations, associations and similar social organisations. The following samples were taken from the stratum which represented Polish associations: volunteer firefighters, hunting clubs, sports associations and other organisations.

The research involving the SOF-1 report was conducted in the spring of 2015 by the Statistical Office in Cracow. Data for research were collected from 24 000 entities, out of the 33 500 that qualified for the research. As of December 31st 2015, the level of completeness for the study as whole was recorded at 85%. In addition, the entities that the authors had been unable to contact the time before that took part in a registration survey carried out in September and October 2015 aimed at bringing greater precision to the generalisation of the research data. The results of the registration study, as well as the administrative data on employment from The Polish Social Insurance Institution (*Zakład Ubezpieczeń Społecznych,* ZUS) and the revenue data from the Ministry of Finance, were used to define the sampling weights for each result set, allowing for generalisation of the data collected during proper research.

3. Methods

When identifying the most effective ways of dealing with the problem of missing data in a database, the contributing factors should be investigated. Rubin (1976) defined three kinds of missing data mechanisms: missing completely at random, missing at random and missing not at random. Identifying a mechanism which contributes to generating missing data is not an easy task. One approach is to visualise the distribution of missing data in a database to exclude the possibility of missing not at random (Schafer and Graham, 2002).

The graphical methods for data presentation are widely used in statistical surveys (Kelleher and Wagener, 2011). These methods are also used to visualise the distribution of missing data in a database (Templ et al., 2012).

Another tool used in the study is 'association rule' (Kotsiantis and Kanellopoulos, 2006). Association models represent the co-occurrence of values/variants in a dataset. Association rule is an expression $A \Rightarrow B$ which states that if A occurs, then B occurs with a certain probability. A number of measures can be used to assess the degree of confidence in the association rule, including:

- $\text{support}_{AB} = P(A \cap B)$ – provides information about the probability of co-occurrence of both A and B;
- $\text{confidence}_{AB} = P(B|A) = \frac{P(A \cap B)}{P(A)}$ – provides information about the conditional probability of B in sets that contain A;
- $\text{lift}_{AB} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{\text{support}_{AB}}{P(A)P(B)} = \frac{\text{confidence}_{AB}}{P(B)}$ – provides information about the change in the probability of B given the presence of A.

The study involved applying the methods used to visualise the distribution of missing data in a database and the association rules to define the profiles of the entities/respondents which failed to provide certain answers in the study conducted using the SOF-1 Reporting Form.

The analysis uses the R package *VIM* and the *Statistica* program.

4. Results

The database contained 11 345 records (the number of analysed research subjects) and about 300 variables. The variables represented the questions contained in the SOF-1 form. The following presentation of the results was limited only to questions from Section VIII (Operating conditions). Section VIII contained the following questions:

- Question d8_1_lok: As of December 31st, 2014 – did the entity have access to the premises? d8_1_l1: was the entity the owner (co-owner) of the premises – in m^2; d8_1_l2: did the entity use the premises on a rental basis – in m^2; d8_1_lok_n: did the entity use the premises free of charge (e.g. lending for use) – in m^2.
- Question d8_1_sam: As of December 31st, 2014 – did the entity have access to cars? d8_1_s1: was the entity the owner (co-owner) of cars – specify quantity; d8_1_s2: did the entity use cars on a rental basis – specify quantity; d8_1_sam_n: did the entity use cars free of charge (e.g. lending for use) – specify quantity.
- Question d8_1_kom: As of December 31st, 2014 – did the entity have access to computers? d8_1_k1: was the entity the owner (co-owner) of computers – specify quantity; d8_1_k2: did the entity use computers on a rental basis – specify quantity; d8_1_kom_n: did the entity use computers free of charge (e.g. lending for use) – specify quantity.
- Question d8_2: What were the most serious operating problems encountered by the entity in 2014? The question had 12 variants and more than one answer could be selected.

Table 1 shows the number of missing answers to questions from Section VIII.

Table 1. Number of missing answers to questions from Section VIII

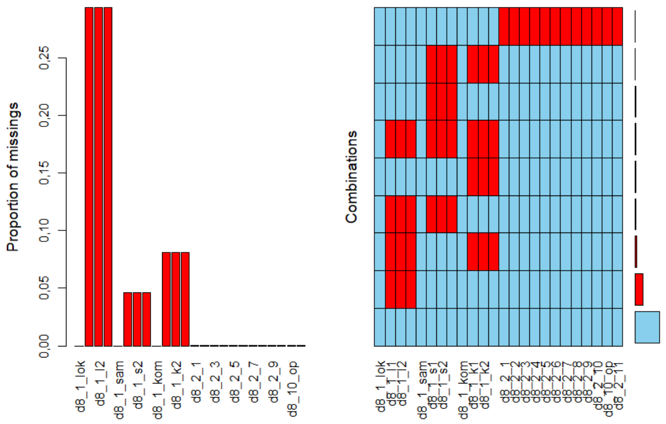
Question	Number of missing answers	Question	Number of missing answers
d8_1_lok	0	d8_2_1	8
d8_1_l1	3332	d8_2_2	8
d8_1_l2	3332	d8_2_3	8
d8_1_lok_n	3332	d8_2_4	8
d8_1_sam	0	d8_2_5	8
d8_1_s1	524	d8_2_6	8
d8_1_s2	524	d8_2_7	8
d8_1_sam_n	524	d8_2_8	8
d8_1_kom	0	d8_2_9	8
d8_1_k1	923	d8_2_10	8
d8_1_k2	923	d8_2_10_op	8
d8_1_kom_n	923	d8_2_11	8

Table 2 shows the combinations of given and missing answers to questions from Section VIII (1 – no answer, 0 – answer given). Data visualisations for respective proportions of specific types of combinations of answers in the total number of analysed research subjects can be found on the right-side graph in Figure 1.

Table 2. Combinations of given and missing answers to questions from Section VIII

Combination of answers	Number	Share (in %)
0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0	7677	67.669
0:1:1:1:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0	2407	21.216
0:1:1:1:0:0:0:0:0:1:1:1:0:0:0:0:0:0:0:0:0:0:0:0	516	4.548
0:1:1:1:0:1:1:1:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0	239	2.107
0:0:0:0:0:0:0:0:0:1:1:1:0:0:0:0:0:0:0:0:0:0:0:0	213	1.877
0:1:1:1:0:1:1:1:0:1:1:1:0:0:0:0:0:0:0:0:0:0:0:0	170	1.498
0:0:0:0:0:1:1:1:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0	91	0.802
0:0:0:0:0:1:1:1:0:1:1:1:0:0:0:0:0:0:0:0:0:0:0:0	24	0.212
0:0:0:0:0:0:0:0:0:0:0:0:1:1:1:1:1:1:1:1:1:1:1:1	8	0.071

Note: 1 – no answer, 0 – answer given



Note: missing answers are marked in red, answers given are marked in blue.

Fig. 1. Distribution of the missing answers for questions from Section VIII by question (left graph) and proportions of the combinations of answers and missing answers for questions from Section VIII in the total number of research subjects

The following section of this study shows the results of applying the methods used to visualise the distribution of missing data in a database and the association rules to the analysis of correlations between the missing answers in the SOF-1 report and the characteristics of the respondents, using question d8_1_l1 as an example.

Figure 2 shows the distribution of missing answers to question d8_1_l1 by three selected characteristics of the respondents:

- ‘Organisation’ variable: 1 – Foundations, 2 – Volunteer firefighters, 3 – Physical culture associations and sports associations, 4 – Hunting clubs, 5 – Typical associations and social organisations, 6 – Denominational entities engaged in social activities;
- ‘Revenue class’ variable: 1 – no revenue, 2 – with revenue below PLN 1 000, 3 – with revenue of PLN 1 000 to 10 000, 4 – PLN 10 000 to 100 000, 5 – PLN 100 000 to 1 million, 6 – over PLN 1 million;
- ‘Urbanisation’ variable: *ug* – urban gmina, *cps* – city with powiat status, *urg* – urban-rural gmina, *rg* – rural gmina.

Employees from the Statistical Office in Cracow selected these characteristics drawing on their knowledge and experience. To provide greater clarity of the following graphical presentation of results, only two types of organisations were examined, i.e. foundations (variant 1) and typical associations and social organisations (variant 5). The selected types of organisations represent the largest share of the studied population. The dimensions of the rectangles (width and height) depend on the number of units of a particular type in the total number of analysed research subjects, while the area of the rectangles represents the number of possible answers to question d8_1_l1 (i.e. the number of records in a database). The rectangle area marked in red represents the number of missing answers in the total number of analysed research subjects. The rectangle area marked in blue represents the share of answers given.

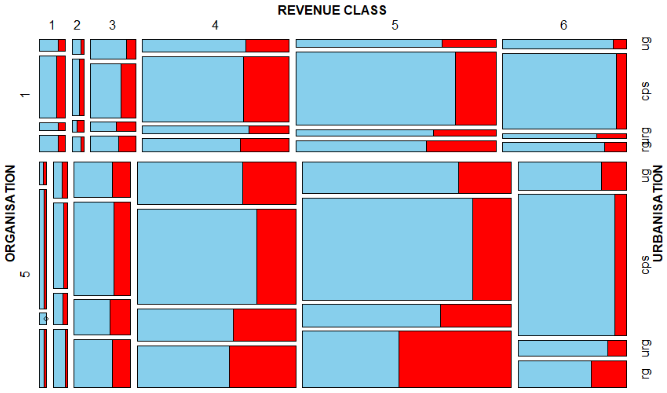


Fig. 2. Distribution of missing answers to question d8_1_11 by three selected characteristics of the respondents

Figure 2 shows that the highest level of completeness in terms of answers given can be observed for the reporting forms which have been completed by foundations (variant 1) generating more than PLN 1 million in revenue (variant 6), based in the cities with powiat status (variant *cps).* Whereas the lowest level of completeness in terms of answers given was reported for typical associations and social organisations (variant 5) with a revenue of PLN 100 000 to 1 million (variant 5), which are based in the rural gminas (variant *rg*).

The methods used to visualise the distribution of missing data in a database provided a valuable first insight into existing correlations between the missing answers in the SOF-1 report and the profile of the respondents. A graphical analysis can be used to select the potential characteristics of the respondents that can be helpful in defining the profile of respondents who failed to answer certain questions included in the study conducted using the SOF-1 Reporting Form.

The next step was to apply association rules based on the assumption that the consequent must be a missing answer variable for question d8_1_11, while the antecedent must be variants of three selected characteristics of the examined entities. The minimum support for association rules was set at 5%, and the minimum confidence – at 20%. The results of the analysis are presented in Table 3.

The highest lift value (1.523) was reported for the association rule where the antecedent is variant 6 of the ‘Organisation’ variable (denominational entities engaged in social activities). The association rule involving variant *rg* of the ‘Urbanisation’ variable (i.e. rural gmina) was ranked second in terms of lift value (1.476). A lift value of more than 1 was reported also for the association rules with the following antecedents: a revenue of PLN 10 000 to 100 000, a revenue of PLN 100 000 PLN to 1 million, and typical associations and social organisations with a revenue of PLN 100 000 to 1 million.

The results obtained using association rules are consistent with those of the graphical analysis. The reason why entities with the following three characteristics: typical associations and social organisations (variant 5) with a revenue of PLN 100 000 to 1 million (variant 5) which are based in the rural gminas (variant *rg*) did not co-occur in the antecedent – despite this being supported by the visualisation of missing answers – is that the minimum support was adopted for the association rule, i.e. 5%.

Table 3. Association rules for question d8_1_l1

Antecedent	⇒	Consequent	Support (%)	Confidence (%)	Lift
Organisation = 5	⇒	d8_1_l1 = *NA*	13.7	28.2	0.959
Urbanisation = *cps*	⇒	d8_1_l1 = *NA*	11.7	22.5	0.767
Revenue = 5	⇒	d8_1_l1 = *NA*	11.4	30.9	1.051
Urbanisation = *rg*	⇒	d8_1_l1 = *NA*	9.5	43.4	1.476
Revenue = 4	⇒	d8_1_l1 = *NA*	9.0	33.0	1.123
Organisation = 5 Revenue = 5	⇒	d8_1_l1 = *NA*	5.6	30.5	1.037
Organisation = 6	⇒	d8_1_l1 = *NA*	5.0	44.7	1.523

By establishing contact with the entities selected on the basis of analysis, we can improve the level of completeness in terms of the answers given to question d8_1_l1.

Conclusions

The methods used to visualise the distribution of missing data in a database and the association rules can be successfully applied to examine the correlations between the missing answers in the statistical research and the profile of the respondents.

The method of analysis outlined in this paper can be performed for all questions in the SOF-1 form. The results of such analysis can be used to increase the level of completeness of the answers given in the SOF-1 report, resulting in reduced study costs, more accurate generalisations, etc.

The research findings can be used to identify data missing mechanisms and thus to select the right method for estimating missing information.

Further research is planned, this time including a breakdown into voivodships. If continued, this research will foster cooperation between the Statistical Office in Cracow and the Department of Statistics at Cracow University of Economics.

Acknowledgements

Publication was financed from the funds granted to the Faculty of Management at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

References

Eaton, C., Plaisant, C., & Drizd, T. (2005). Visualizing Missing Data: Graph Interpretation User Study. In: Costabile, M.F., Paternò, F. (eds.), *Human-Computer Interaction – INTERACT 2005. INTERACT 2005*, Lecture Notes in Computer Science, vol. 3585, Springer, Berlin, Heidelberg, 861–872.

Kelleher, Ch., & Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, 26(6).

Kotsiantis, S., & Kanellopoulos, D. (2006). Association Rules Mining: A Recent Overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 71–82.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3).

Satellite Account on Non-profit and Related Institutions and Volunteer Work (2018), United Nations, New York.

Schafer, J.L., & Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2), 147–177.

Templ, M., Alfons, A., & Filzmoser, P. (2012). Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, 6(1), 29–47.