Measuring representativeness of different data sources connected with short-term statistics

Prepared by Mateusz Smektalski (m.smektalski@stat.gov.pl) and Alina Szkop (a.szkop@stat.gov.pl),

Poznań Statistical Office

Introduction

In the modern market economy, entrepreneurs need information to make decisions. The appropriate selection of accessible information of sufficient quality affects the validity of assessment of the current business situation and projections that can contribute to the success or failure of a company, an entire industry or even the national economy. For this reason, one of the key elements of any national policy is to maintain registers, which are used to collect, process and monitor many aspects of social life and economic events. In Poland such registers are maintained, among others, by the Central Statistical Office or Ministries of Finance, Health and Economy.

Official statistics provides guidelines for norms and standards, is the source of classifications and typologies of social, economic and political phenomena and relies on scientific methodology applied in each survey to guarantee the clarity, objectivity and confidentiality of statistical outputs [1]. However, the wide scope and high frequency of surveys in official statistics, which entails high respondent burden, and the very complexity of the problem are the reasons why less attention is paid to the important question of the verification of data quality.

The present study is aimed at comparing data collected by the Central Statistical Office (CSO) with data collected independently by other units of public administration. The study compares information reported in the monthly DG-1 business survey with data from tax statements submitted to the Ministry of Finance (MF). The comparison of datasets is an attempt to determine the representativeness and reliability of data, which are later aggregated, generalized and released to the public. The study is part of a methodological project entitled “The use of administrative data in the assessment of the current economic activity of enterprises”, which is being conducted by the Methodology and Programming Division of the Statistical Office in Poznań. The purpose of the project is to modernize the system of statistical production by making a better use of data from administrative registers, and the present study can provide support for the results of the project work.

The article has the following structure. The first part describes the approach adopted in the study and its scope. It provides characteristics of the CSO and FM datasets and the variables they contain. This part also includes a description of the methodology and statistical methods used to compare the parameters of the datasets. The second part presents results of an empirical analysis of the response rate and reliability of data reported in the DG-1 survey.

Description of the study

Despite appearances, data quality is a wide and problematic area. Usually the term quality refers to the selection of an appropriate method of data collection (survey technique); the preparation of tools, including the way questionnaire questions are formulated and the clarity of concepts; the correctly defined survey population and the sampling scheme; the training and supervision of enumerators; survey representativeness, which is defined by the size of nonresponse error; the correct implementation of the process of data input and processing; measurement precision and many other criteria.

Franciszek Sztabiński [2] has made an attempt to classify traditional ways of assessing data quality and distinguished two sets of alternative perspectives: internal vs. external, and indirect vs. direct. The internal perspective is based on the analysis of the survey process, its components and identification of errors found in the survey process. The external perspective in data quality assessment disregards the survey process and any potential errors and focuses on the analysis of data validity and reliability and seeks to determine their "truthfulness" in the process of verification and data validity. In order to narrow down the terminological scope of data quality used in this study, the external perspective is used.

As already mentioned, one of the datasets used in the comparison comes from the DG-1 survey. DG-1 is a monthly survey of business entities. The survey covers all entities employing 50 or more people and a 10% sample of businesses employing from 9 to 49 people. The survey collects information about economic activity, such as: net revenue from the sale of products, goods and materials, retail and wholesale sales, employees, wages. The survey covers legal entities and organizational units without a legal status and private individuals employing 10 or more people. The survey covers entities whose activity according to the Polish Classification of Activity (PKD 2007) falls into the following sections: from B to J, L, M (excluding divisions 72 and 75), N, R and divisions 02, 95, 96 and class 03.01.

Data from the DG-1 survey for a given month are stored in the B1 database, which is the provincial database of unit data. Another database labelled B3 contains aggregated data from B1, generalized data (i.e. aggregate data supplemented with a generalization of medium-sized units not included in the survey and active large enterprises, which have not completed a DG-1 reporting form), price indexes and correction indexes for generalized data.

For purposes of comparison, administrative registers used in the study included datasets of payers of personal income tax, reduced-rate tax on gross income (for specified occupations), lump-sum tax, and payers of corporate income tax.

This article does not include details connected with the integration of registers, since they were thoroughly described in the article by Grażyna Dehnel [3]. Therefore, the input dataset for the study is an intersection of the CSO and MF registers. December 2011 is the reference period.

The two datasets were compared using the methodology of The Five Steps of Statistical DQA [4]. The very process of selecting statistical methods comes from the publication of the United States Environmental Protection Agency [5].

The comparison of registers was treated as an analysis of two independent surveys, in which the same respondents are asked again to answer questions from the basic questionnaire (DG-1) in a control survey (FM), which corresponds to the direct approach [2].

Hence, despite the fact that the variables tested for reliability refer to the same quantity, i.e. revenue, they are treated as independent variables. The dataset was first analyzed using basic descriptive statistics and graphical data analysis. The histograms showed that the variables are not normally distributed and are strongly right-skewed. Moreover, there are differences in the counts of observations depending on the intervals of the variables of interest, which are evident in the histograms. A scatter plot shows a positive, monotonic (linear) correlation, which is an expected and desirable characteristic. Based on the initial assumptions, two-sample nonparametric methods for independent groups were selected. Their detailed description is presented below.

The correlation between two variables was measured using Pearson's r, Spearman's rho (Spearman's rank correlation coefficient) and Kendall's tau coefficient.

Further analysis involved calculating the R-indicator for the two registers. The R-indicator was calculated using the following formula [5]:

$$\hat{R} = 1 - 2S_{\hat{\rho}} = 1 - 2\sqrt{\frac{1}{\sum_{i=1}^{n} w_i - 1}\sum_{i=1}^{n} w_i(\hat{\rho}_i - \bar{\hat{\rho}})^2} \qquad (1)$$

where w_i is the sample design weight for unit *i*, $\hat{\rho}_i$ are the response propensities (estimated using a logistic model).

All statistical tests were conducted using procedures implemented in the SAS software [6].

Results

The graphical analysis of the DG-1 and FM datasets (Figure 1) clearly indicates the lack of normal distribution, which is the basic assumption of most parametric tests. The distribution of observations is skewed positively to a very high degree, that is as revenues increase their number drops. Figure 1 shows that the largest group of enterprises (50%) is found in the revenue interval from 0 to PLN 1,333.90 (median).

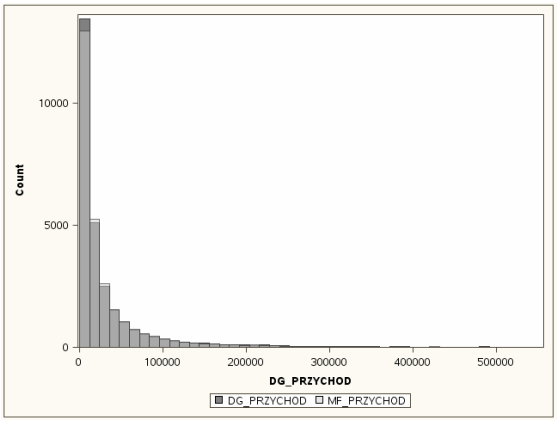

Figure 1. Histogram of variables DG_ REVENUE and MF_ REVENUE.

Variable	DG_REVENUE	MF_REVENUE	diff_MF_DG
N	28 535	28 535	28535
MEAN	78 823.93	83 262.48	4 438.55
MEDIAN	13 333.90	14 061.60	313.40
MODE	651.00	1 234.70	-
STD EROR MEAN	4 842.47	4 761.08	717.27
SKEWNESS	90.34	76.43	13.24
KURTOSIS	11 278.59	8 503.56	3 348.36

Table 1. Basic statistics.

The scatter plot reflects the relationship between two variables: DG_REVENUE and MF_REVENUE. The plot shows a positive correlation, which means that an increase in DG_REVENUE is accompanied by an increase in MF_REVENUE. The boxplot shows the basic descriptive statistics of the two datasets.

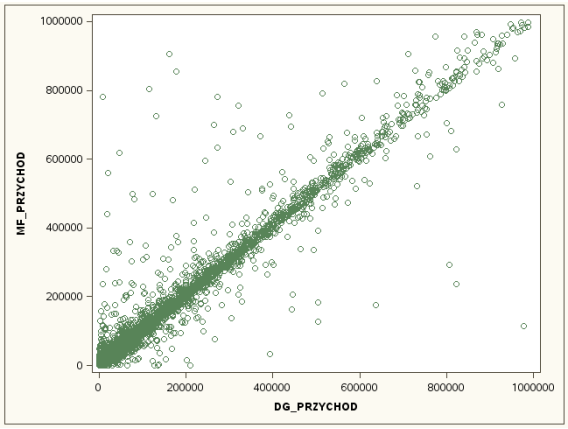

Figure 2. A scatter plot showing the correlation between DG_ REVENUE and MF_ REVENUE.

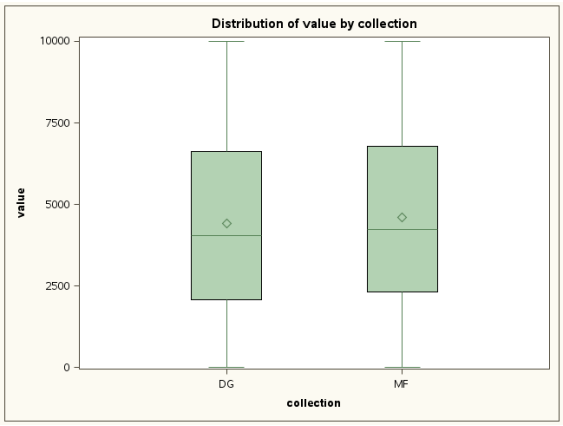

Figure 3. A box plot for DG_ REVENUE and MF_ REVENUE.

The strength of correlation between the two variables was tested using Pearson's r, Spearman' rho and Kendall's tau. Each of the tests can take values in the range <-1; 1>, but their interpretation are somewhat different. The results of the test are presented in the table below.

BY SECTION	R	rho	tau	p
A	-	-	-	-
B	0.98827	0.98494	0.93963	0.0001
C	0.99551	0.98758	0.94343	0.0001
D	0.99356	0.99033	0.95743	0.0001
E	0.97521	0.98183	0.93402	0.0001
F	0.97179	0.97745	0.92367	0.0001
G	0.98187	0.98924	0.95384	0.0001
H	0.99037	0.97943	0.92868	0.0001
I	0.9816	0.94187	0.80956	0.0001
J	0.98771	0.9669	0.92122	0.0001
L	0.94706	0.96829	0.90493	0.0001
M	0.54728	0.97474	0.9152	0.0001
N	0.73434	0.96551	0.88741	0.0001
R	0.99645	0.69711	0.52378	0.0001
S	0.98498	0.92257	0.87484	0.0001
TOTAL	0.98899	0.98036	0.9276	0.0001

Table 2. The correlation between DG_ REVENUE and MF_ REVENUE measured by Pearson's r, Spearman' rho and Kendall's tau

In each of the tests, the p-value exceeds the 5% significance level. On this basis we can reject the null hypothesis and accept the alternative hypothesis, which postulates the existence of a correlation between the two datasets.

In the second stage of comparison, the percentage differences between observations from the FM dataset (CIT/PIT) and the DG-1 dataset (difference = FM_REVENUE – DG_REVENUE) were grouped into intervals. The counts of enterprises in each interval are shown in Table 3.

BY SECTION	equal to	0-1%	1-5%	5-10%	10-25%	25-50%	50-100%	1-10x	10-100x	>100x
TOTAL	121	6262	11031	4301	3608	1629	1401	158	13	11
PERCENTAGE	0.42%	21.94%	38.66%	15.07%	12.64%	5.71%	4.91%	0.55%	0.05%	0.04%
CUMULATIVE	121	6383	17414	21715	25323	26952	28353	28511	28524	28535
CUMULATIVE PERCENTAGE	0.42%	22.37%	61.03%	76.10%	88.74%	94.45%	99.36%	99.92%	99.96%	100.00%

Table 3. Intervals of differences (FM_REVENUE – DG_REVENUE).

In order to measure the representativeness of the DG-1 and FM registers, a 10% sample of all companies (large and medium-sized) was drawn. The R-indicator was calculated according to formula (1). The indicator takes on values in the range <0;1>, where 1 denotes strong representativeness and 0 – a maximum deviation from strong representativeness. The response propensities used to calculate the R-indicator were estimated using a logistic model. The resulting values are presented in Table 4.

Register	R-indicator
DG-1	0.365
MF	0.883

Table 4. Values of the R-indicator.

Summary

The study involved comparing two registers. The first one – maintained by the Central Statistical Office (CSO) – contains information about economic activity of enterprises collected in the monthly DG-1 survey. The second register was made available by the Ministry of Finance (MF) for the purpose of assessing the quality of statistical data and contains annual information from PIT and CIT tax statements. The comparison of the registers was treated as an analysis of two independent surveys, in which the same respondents are asked again to answer questions from the basic questionnaire (DG-1) in a control survey (FM)

The comparative analysis revealed a strong correlation of over 90% between the two datasets. Over 61% of answers in the DG-1 survey differed from information indicated in tax statements by less than 5%. It can then be concluded that output statistics produced by CSO are based on reliable data.

The study of representativeness found a large disproportion between the datasets. The R-indicator for the MF dataset was equal to 0.883, while for the DG-1 dataset only 0.365. The MF dataset is more complete because it contains more observations as a result of the reporting obligation which is sanctioned by a fine or even imprisonment. The use of larger samples in the process of generalization and analysis certainly improves the quality of these statistical outputs.

In summary, the advantage of the DG-1 dataset is the frequency of obtaining new data and the survey's level of detail. Using the MF dataset as a control dataset, we have managed to demonstrate that revenue values obtained from the DG-1 survey do not significantly differ from those indicated in tax returns. Moreover, by conducting additional analysis and using other administrative registers, it is possible to improve data quality, e.g. by analysing distributions of variable (contained in MF registers), the number of employees from the registers of the Social Insurance Institution, the validity of address information from the Central Registration And Information On Business, which is maintained by the Ministry of Economy.

References

[1] Ustawa z dnia 29 czerwca 1995r. o statystyce publicznej (Dz.U. 1995 Nr 88 poz. 439)

[2] SZTABIŃSKI F., 2011, *Ocena jakości danych w badaniach surveyowych*, Wydawnictwo IFiS PAN, Warszawa, Polska, strona 19

[3] DEHNEL G., 2014, *The tax register and the social security register in estimation methodology of short-term statistics in Poland*, Konferencja Small Area Estimation 2014, Poznań, Polska

[4] United States Environmental Protection Agency (EPA), 2006, *Data Quality Assessment: A Reviewer's Guide*, Washington D.C., USA, strona 9

[5] United States Environmental Protection Agency (EPA), 2006, *Data Quality Assessment: Statistical Methods for Practitioners*, Washington D.C., USA, strona 51

[6] SAS Product Documentation *http://support.sas.com/documentation* PROC NPAR1WAY, PROC GPLOT, PROC CORR, PROC UNIVARIATE, PROC MEANS, PROC SURVEYSELECT