Jacek Kowalewski
Statistical Office in Poznan
(representing the Central Statistical Office of Poland)

Possibilities of exploiting administrative data in short term statistics in Poland

A dynamic development of new techniques, an ever closer integration within the European Union as well as a necessity to reduce respondent burden all call for a modification of the data collection system in the field of economic statistics, which includes most surveys conducted by Central Statistical Office. A team of employees of Statistical Office in Poznan, aided by employees of Statistical Office in Katowice and statisticians from University of Economics in Poznan, working under the grant agreement "**Use of Administrative Data for Business Statistics",** conducted a research program aimed at exploring the possibilities of using existing administrative registers in short term statistics. The scope of work comprised an assessment of potential usefulness of administrative sources, an analysis of ways to decrease response burden for some companies and improve data completeness and quality. The last objective involved mainly assessing the possibilities of exploiting register data for purposes of calibration in the case of incomplete data.

Provision of short-term statistics in Poland

The monthly DG1 survey ("business activity report") is the basic source of short-term information about businesses, which is intended to include all units employing 50 or more people and a 10% sample of units with 10-49 employees, comprising economic activity sections B÷J, L, M (excluding divisions 72 and 75), N, R and divisions 02, 95, 96 and class 03.11. The report contains data about economic activity: sales revenue (from products and services), the volume of wholesale and retail, excise tax, specific subsidies and data about employees, the average number of employees and salaries. The DG1 survey is intended to create a system of providing monthly information about basic measures describing economic activity of businesses.

The use of data from administrative registers

The most important and potentially useful sources of information to be used for purposes of business statistics are tax and social insurance information systems. Both systems were reviewed with the following objectives in mind:

1. to recognise and describe regulations concerning the founding principles, system maintenance and data accessibility, the purpose of the system, the technology of system maintenance, the system's information scope, sources and frequency of updating, time and forms of data publishing, administrator, data sources, data scope;
2. to compare the scope of data used in the registers with those used by public statistics;
3. to compare definitions of concepts and classifications used in the administrative systems with those used by public statistics. This task was carried by means of an application (the PIK system) used in statistics, which stores information about description structure and

principles of comparing concepts and classifications used in the administrative systems with corresponding concepts used by public statistics and the typology of degrees of equivalence along with their interpretation. The degree of equivalence was determined using the following scale: identical, convergent, corresponding, different;
4. to study methodological compatibility;
5. to evaluate the quality of administrative systems to determine their usefulness as data sources for business statistics and predicted forms of usage (Table 1). The quality of registers was assessed to be high and very high;
6. to specify the scope of data to be used.

Table 1. Evaluation of the quality of administrative systems (max 49)

No.	Database / Register	Evaluation of system quality
	Tax system	
1.	Database of taxpayers of the personal income tax PIT as a source of data in the field of labour market	41
2.	Database of taxpayers of the personal income tax PIT as a source of data in the field of revenues and costs of activities as well as taxes	34
3.	Database of taxpayers of the corporate income tax CIT as a source of data in the field of revenues and costs of activities as well as taxes	41
4.	Database of taxpayers of the value added tax VAT as a source of data in the field of revenues and costs of activities as well as taxes	44
5.	National Register of Taxpayers (KEP) as a source of data in the field of labour market and revenues and costs of activities	49
	System of social insurance	
1.	Central Register of Contribution Payers as a source of data in the field of labour market	39
2.	Central Register of the Insured as a source of data in the field of labour market	37

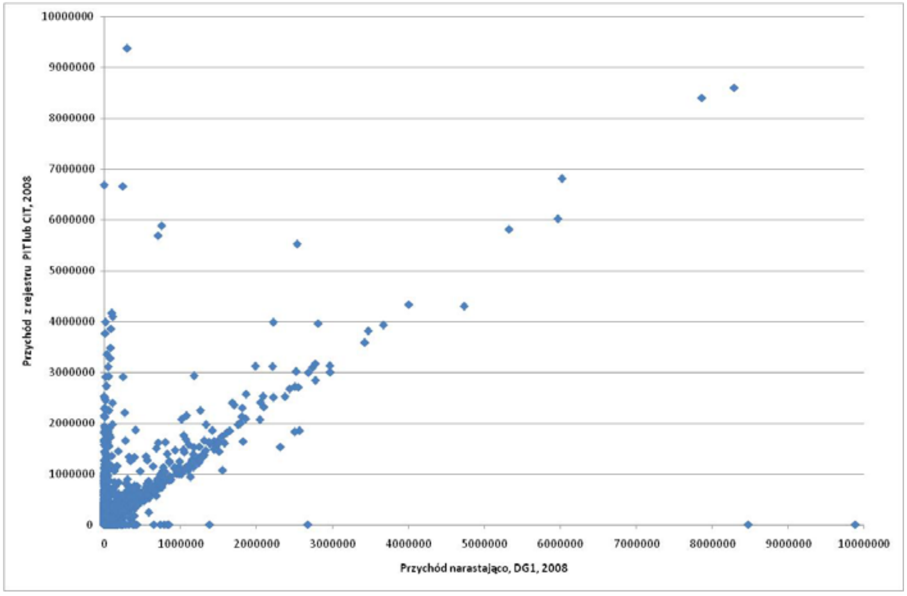
Source: Report

Relationships between variables from the DG1 survey and administrative registers

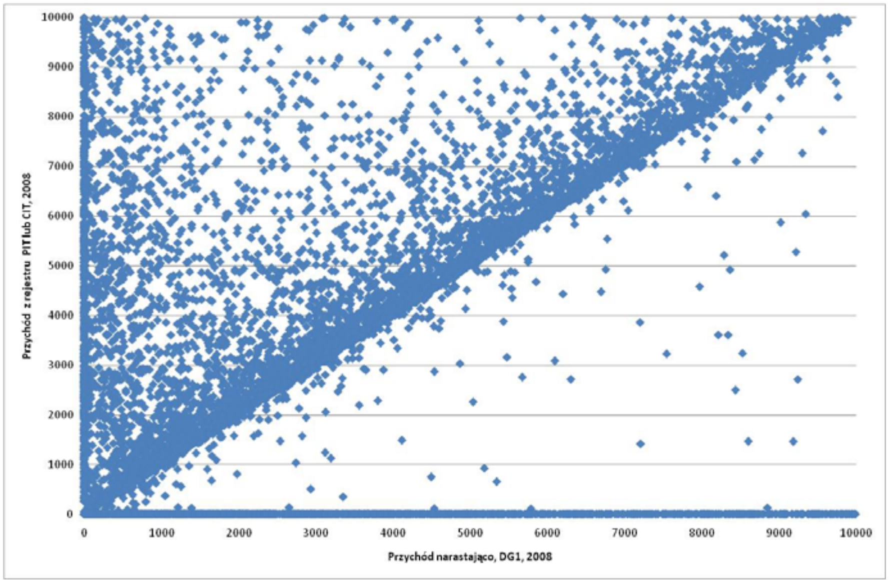
With access to data from administrative register, particularly those concerning tax information, it was possible to evaluate the reliability of information collected through statistical reporting in the DG1 survey.
One way of assessing reliability from different sources is to analyze their scatter plots. In the case of compatibility of information reported by a business in the DG1 survey and tax

returns, for example with respect to revenue, scatter plot data should form a band centered around the identity line, which can be described as the function y_1=$y_{2.}$

Fig. Relationship between the values of accumulated revenue - from DG-1, PIT or CIT register, 2008



Scale fitted to units with the highest revenue (limited to PLN 10 000 000)



Scale not fitted to units with the highest revenue (limited to PLN 10 000 000)

Having analyzed the scatter plots, one can notice that:

- In most cases (over 94%) businesses report lower values of variables in statistical reporting (DG1) than in tax statements. About 40% of enterprises report virtually identical values of variables under analysis.
- There is a sizeable group of enterprises, which reported non-zero revenue in the DG1 survey while the corresponding field in the tax register showed a missing value or contained zero. This phenomenon can be partly attributed to the discrepancy between the definition of revenue in the DG1 survey and the PIT/CIT tax return.

- There is also a large number of business, which reported significantly higher revenues in tax returns than those indicated in the DG1 survey. Opposite cases, where revenue reported in tax returns was lower than that indicated in the DG1 survey were rare.
- high compatibility between the value of revenue reported in the DG1 survey and tax returns by medium-sized and large enterprises.
- high differentiation of the correlation coefficient depending on section. Pearson's correlation coefficient for section *trade* was equal to 0.4348 while for section *manufacturing* it was 0.9589. The difference was largely due to different definitions of sales revenue.

Possibilities of exploiting administrative data for calibration in business statistics

The study into the possibility of using calibration was conducted in the following way:

1. Transforming administrative data sources into statistical data sets. In the process of database integration a special **MEETS real data set** was created for purposes of the project study. It contained records about economic entities representing the four PKD sections of economic activity (*manufacturing, construction, trade, transport*), which participated in the DG-1 survey in December 2008 and which were successfully combined with information from the the KEP, CIT, PIT and ZUS databases. The database was treated as the general population to be sampled during the study.

2. Mean revenue was estimated using a calibration estimator with a known vector of the population total of auxiliary variables. In order to compare the results obtained in the study, the Horvitz-Thompson estimator was used as a benchmark.

3. During the simulation study, 5%, 10% and 15% samples were drawn from the MEETS real dataset, using simple random sampling without replacement. After obtaining a sample, information about revenue (dependent variable Y) for some enterprises was replaced with missing data. 3 different approaches were used to generate missing data. In the first one missing data were generated in a random fashion (option 1). In the second (option 2) and third (option 3), missing data were attributed to enterprises with the lowest and highest revenue respectively.

4. For each option, 500 iterations were performed to estimate the expected value of revenue, the expected value of the bias of the estimators and their empirical variance as well as relative estimation errors.

Table 2. The relative estimation error of estimators of the monthly enterprise revenue (in percent)

		Horwitz-Thompson estimator			calibration estimator		
sample size	% of missing data	option 1	option 2	option 3	option 1	option 2	option 3
5%	5%	28.37	28.71	8.94	14.40	14.35	12.25
	10%	29.78	26.61	7.51	14.89	14.06	9.79
	15%	32.43	27.49	7.42	15.08	14.60	9.10
10%	5%	20.33	19.32	6.03	10.77	10.57	7.95
	10%	19.65	18.94	5.45	9.73	10.04	6.19
	15%	20.05	19.74	5.12	11.19	10.27	6.07
15%	5%	14.69	15.11	4.67	8.42	8.15	5.29
	10%	16.40	14.15	4.24	8.67	7.87	4.75
	15%	16.49	15.30	3.81	8.62	8.03	4.50

Source: Report

Conclusions

The main problem connected with the direct use of administrative data in short-term statistics is the overly long time of data processing and a long period of waiting for data to be transferred by their administrators.

The definitional discrepancies found between the system of administrative registers and that public statistics as well as time restrictions connected with access to data should not exclude them as data sources for business statistics. Administrative databases can be used as direct data sources for surveys, as a source of information to complete missing data, as a source of data for comparison with information collected in statistical surveys conducted by means of statistical reporting forms. The databases are a rich source of potential auxiliary variables, which can improve the quality of estimations by reducing the negative effect of non-response in statistical reporting.

The use of all available information about selected variables that describe the economic activity of businesses can offset the negative effect of non-response by reducing the respondent burden and improving the efficiency of estimators. While the use of administrative data in producing business statistics may not completely solve all its problems (highly right-skewed distributions, very high differentiation and high concentration), it can significantly contribute to minimizing their negative effects.

As the simulation experiment conducted during the study indicates, the use of the calibration approach in a business survey can improve the quality of short-term estimation. The calibration estimation adopted in the study is a kind of remedy for the problem of non-response that statistical reporting suffers from and, combined with the methodology of small area statistics, can increase the range of estimation methods available in business surveys.