“The Use of Administrative Data to Improve Quality of Business Statistics Concerning Micro-Enterprises”.

Paper prepared by
Regional Statistical Office in Łodź
on the base of project
“The *Implementation of a More Efficient Way of Collecting Data*”

Summary

The paper contains the description of the use of administrative data in Polish public statistics concerning micro-enterprises, ranging from their methodological conformity with the system of statistics, through their usage in the newly constructed data base and already existed Statistical Business Register to their role in estimating and assessing data on the example of the most extensive micro-enterprise survey in Polish statistics - SP-3 survey – works within project “The Implementation of a More Efficient Way of Collecting Data” carried out 01.09.2008- 31.08.2009 by the Central Statistical Office of Poland.

Introduction

In Poland, once a year micro-enterprises are examined by means of SP-3 survey. While conducting it Polish public statistics encounter many difficulties (among others: out of date records and data in Statistical Business Register and especially lack of contact with respondents, lack of responds, too small samples for low level of geographical and NACE aggregation).
Administrative data are currently used in realisation of the survey to improve its quality. The use mainly concerns the frame preparation and sampling process. On the base of administrative data from the Ministry of Finance the upper stratum has been defined, i.e. a subset of the most important units in the frame, which are completely surveyed. .
The project "The Implementation of a More Efficient Way of Collecting Data" was launched to recognize the possibilities of wider use of administrative data. It shot for:

- **Improvement of SBR** (Statistical Business Register)**quality:**

1. updated enterprise population
2. updated sampling frames
3. widen scope of statistical features (characteristics) i.e. revenues, remunerations, incomes,
4. implementation of new estimation methods of data from sample surveys to obtain data on low levels of geographical aggregation.

- **Quality improvement** of business statistics and **meeting users' requirements** by wider use of administrative data by means of the increase: of data relevance, timeliness, comparability and coherence, efficiency
- **Reduction of administrative burden** on enterprises and **lowering the costs** of conducting statistical surveys by: reduction of the scope of surveys and information from microenterprises, refraining from imposing the obligation to submit reports on these entities, on condition that a set of relevant data is obtained from administrative sources;

Besides, the usage of administrative sources is expected to:

1. provide us with data to estimates, in case the reporting obligation is not fulfilled,
2. eliminate or restrict the respondent's error,
3. be useful as auxiliary variables in intermediate estimation,
4. substitute the sample survey with a full-scope survey, in the case of a direct full-scope imputation,

To achieve above mentioned goals intensive work was done and experiments conducted. The following methods of application of administrative sources were assessed:

- administrative sources as the direct source of statistical data,
- administrative sources as auxiliary information for indirect estimation,
- administrative sources – as the source of information for SBR and sampling frame improvement
- administrative sources- as the additional/ auxiliary information enabling the increase of sampling efficiency.

The underneath assignments, in detail described later, were taken:

1. The analysis of the methodological consistency and conformity between administrative and statistical systems - the base of information about administrative systems – Mikro PIK;
2. Construction of data base of micro-enterprises containing information about them from

administrative as well as statistical sources, having practical application for further tasks;

3. Exploration of possibilities of administrative data use to enhance the quality of SBR and sampling frame.
4. Working up the method of supply statistical data with administrative sources by their direct use as well as indirect estimation and carrying out the experimental estimates;
5. The analysis of possibilities of administrative data use in sampling.

1. The analysis of the methodological consistency and conformity between administrative and statistical systems

The base of information about administrative systems **– Mikro PIK** is a base of terms and classifications used in administrative systems, establishing methodological conformity of administrative systems with the system of statistics. This system serves verification of terms and classifications between administrative and statistical systems and consists of 3 applications.

The following administrative systems, which may be a potential data source for the micro-enterprise statistics, were identified and described:

1) Information system run by the Ministry of Finance:
   PIT (personal income tax), CIT (corporate income tax), VAT (tax on goods and services-value added tax), KEP (taxpayer register),
2) Information system run by the Polish Social Insurance Institution – KSI ZUS
3) Information system run by the Ministry of Justice
   National Court Register – KRS
4) Information system run by the National Health Fund

About each system is provided a description of the system, information content of the system, the structure of data collections.

The levels of compliance used were the following: identical, convergent, corresponding, different.

Having in mind microenterise surveys there were introduced:

- 97 administrative terms
- 58 statistical terms
- 6 classifications from administrative sources

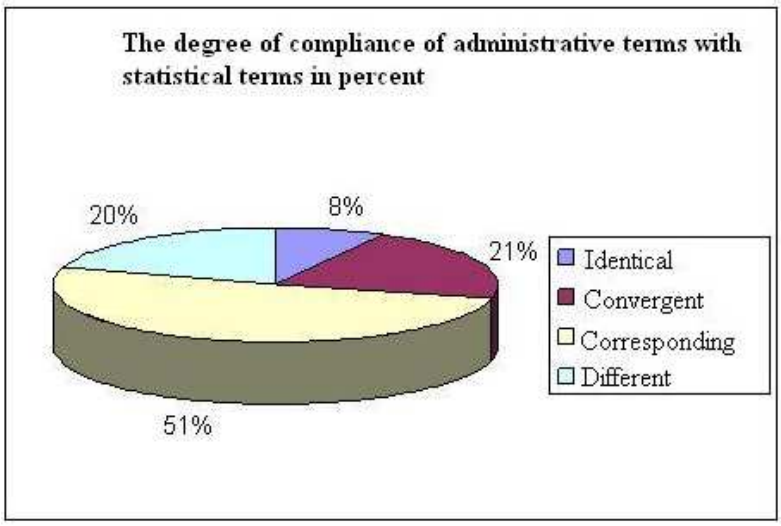
to analyse their methodological conformity.

The specification below shows the number of administrative terms registered in the MikroPik system with a division into administrative systems:

No.	Administrative system name	Number of registered administrative terms
1	**KSI ZUS**	***32***
3	**CIT**	***5***
4	**PIT**	***33***
5	**KEP**	***13***
6	**VAT**	***14***

7	**KRS**	***7***

There were identified 131 linkages of administrative terms, yet the level of compliance is usually not high, which is shown in the chart: Level of compliance of administrative terms with statistical terms.
The chart below displays the compliance level in percentages.



2. Data base of micro-enterprises

The core of the base of micro-enterprises is the information from our SBR (Statistical Business Register) in order to select micro-enterprises connecting and matching all records from administrative sources by the use of ID numbers: REGON (The National Official Business Register) or NIP (taxpayer identification number) or PESEL (Common Electronic System of Population Register). The core base is supplied by the variables from administrative systems which are considered to be in accordance with statistical ones or additionally which would be right and appropriate in order to meet users' needs or variables calculated on the basis of the data previously collected.

Objective scope

The scope of the survey on micro-enterprises covers enterprises with number of employees up to 9 persons (not including private farms). The administrative source-based data sets contain information describing units having any of the legal forms i.e. natural persons conducting economic activity, legal persons and organisational units without corporate status.

If, in the administrative system, there is information on other (lower, local) level of aggregation than legal unit, then efforts are made to "achieve" data specific to legal unit prior to the data import to the database.

The following entities are indicated in administrative systems:

Table 1: Administrative systems and types of units (entities) registered – not including capital groups

System	Legal units	Legal units on lower (local) level of presentation	Natural persons (other than legal units)
KEP	X	X	X
PIT	X		X
CIT	X	X	
VAT	X	X	
ZUS	X	X	X
CWS	X	X	

The National Taxpayers Register (KEP) is a particularly important system, which may be viewed as superior to other administrative sources. It contains the most complete list of entities conducting economic activity, registered in administrative systems.

Variables

The database includes data collected from the statistical business register:

- identification features of legal units,
- other features related to economic activity conducted by units,
- attributes describing selected features, e.g. source and data of the latest update

and:

- features and variables derived or calculated on the basis of data from administrative sources
- units' features calculated on the basis of features describing units in systems or in the statistical business register
- information on the presence of legal unit in an administrative source.

"Database elements" presents a detailed list of features and variables attributed to units of the database.

Table 2: Data sources and the scope of their application

Source	scope of data from the administrative source
KSI ZUS	employed persons, employees, students (for the purposes of the SP3 survey), wages and salaries
VAT	revenues, costs, taxes, outlays on fixed assets, gross value of fixed assets
PIT, CIT	revenues, costs, taxes, information on specialist questions
KEP	Legal form (basic, special) Information on connections between units, Type of unit in the database
CWS	Information on performance of health care services funded by NFZ

In case of information derived from distinct administrative sources - information is assumed to be written separately for each of them.

Database imputations will take place once a year according to a scheme described for the first imputation. The database will comprise data from next imputation years; therefore it will be possible to browse information included from different states and sources.

3. The use of administrative data to enhance the quality of SBR and sampling frame

The National Official Business Register constitutes the base of SBR. Each entity of the national economy is obliged to submit to the National Official Business Register an application (for assignment of identification number, changes in features of already registered entity, removing entity from the register) accordingly to their actual and legal situation.

The quality of the statistical register of enterprises - strongly impacts on the quality of the results obtained from surveys. Increasing the completeness and relevance of the statistical business register is extremely important.

The table below presents the range of characteristics covered by the statistical business register, as well as sources of data updates from the administrative systems recognised as the most useful, i.e. the fiscal system of the Ministry of Finance (PIT, CIT, VAT, KEP), the National Court Register (KRS) of the Ministry of Justice and the social insurance system of the Social Insurance Institution

The use of administrative registers for updating the statistical business register is possible when the terms/features and their definitions selected from administrative systems are consistent with those in the statistical system. The table below shows a plan for updating the statistical business register with administrative variables of the basic features

Range of characteristics covered by the statistical business register	**Sources of data updates**		
	from the fiscal system	**KRS**	**from the** social insurance system
NIP	**X**		
PESEL	**X**		
registered office address	**X**		
address where business activity is run	**X**		
other places for running activity	**X**	**X**	**X**
address for correspondence	**X**		
Telephone	**X**		
Fax.	**X**		
e-mail	**X**		
the name under which the company operates		**X**	
legal status		**X**	
the type of legal and economic activities	**X**	**X**	**X**

information about liquidation		X	
information about dissolution or invalidation of the entity		X	
information about merger or transformation		X	
information about bankruptcy proceedings		X	
information about composition proceedings		X	
the number of the employed			X
the number of the workers			X
class size			X
Income	X		
the kind of account books	X		
information about the number of local entities and groups of enterprises	X	X	X

The frequency of updating the statistical business register using administrative sources depends on the availability, relevance and adequacy of these sources. The programme of statistical surveys of official statistics contains a provision which will make it possible to periodically acquire datasets from their administrators, as well as to increase the frequency of transferring them, depending on their availability.

For each unit attribute value, the source from which it is derived is described, indicating the source-update date. Where the value of the date is higher than the last value stored in the statistical business register, the adjustment is accepted.

The main attributes updated in the statistical business register using administrative sources include the number of workers as well as the legal and economic activity. The attribute has a direct impact on the establishment of the reporting obligation as the burden on the reporting unit depends, on the number of workers defined on a yearly basis. The number of workers will be updated using data compiled on the basis of the information system of the Social Insurance Institution. The activity of units will be verified using data compiled on the basis of tax systems (CIT, PIT, and KEP).

4. Direct use and estimating data - methods and experiments based on SP-3 survey variables

The standard and the most extensive micro-enterprise survey in Polish statistics is SP-3 survey where one of subject of interest is income. Unfortunately, a relatively long period after which it is available significantly narrows the possibility of its current direct use. However, the use of administrative date can be an effective way to improve the precision of income estimates when we use indirect estimation or additional assessment. It also concerns number of active enterprises, costs and VAT. We made some attempts for fix assets and investments as well, however in these cases we did not obtained satisfying effects (results).

The description of proposed methods and algorithms:
An experimental estimation of selected economic categories was carried out, on the basis of data obtained from statistical surveys and administrative systems.
In the table there is information whether the method requires considering information from a traditional survey (i.e. whether it requires further conducting of sample survey, or whether it is solely based on administrative data) and whether due to the dates of gaining the used

administrative data it enables the creation of estimates in time not exceeding previous dates for data publication. The possibility of obtaining results in prevailing data depends on which administrative systems are used – due to different time of obtaining data from the system. In acceptable time data from VAT and KEP sets are available, whereas date for obtaining data from PIT/CIT sets disables receiving estimates in prevailing date. That is why methods meeting this condition are the methods not requiring current information from PIT/CIT. Therefore particularly interesting will be those methods which do not require data from a survey and allow making estimates in the hitherto prevailing time. **(N, Y – notes in bold)**. Additionally, it was marked whether for a given estimation method an experimental estimate was carried out.

Statement of developed and suggested methods of counting and estimates in the field of basic figures on microenterprises

Examined variable	**Calculated variable / estimate (symbol)**	**Used administrative sets**	**Type of method**	**Necessity of surveying on a form**	**Results in prevailing date**	**Experimental estimate**
Number of enterprises	pvcit_n	PIT/CIT VAT	direct counting	**N**	N	Y
	sz_n	$(PIT/CIT, KEP)_{t-1}$[1] KEP	estimate	**N**	**Y**	Y
Nett incomes from the entire activity	pcit_przych	PIT/CIT	direct counting	**N**	N	Y
	eblu_przych	PIT/CIT VAT	indirect estimate	Y	**Y**	Y
	sz_przych	$(PIT/CIT, KEP)_{t-1}$ VAT, KEP	estimate	**N**	**Y**	Y
Costs from the entire activity	pcit_koszt	PIT/CIT	direct counting with estimate	**N**	N	Y
	eblu_koszt	PIT/CIT VAT	indirect estimate	Y	**Y**	N
	sz_koszt	$(PIT/CIT, KEP)_{t-1}$ VAT, KEP	estimate	**N**	**Y**	N
VAT tax due	**vat_pnalezn**	VAT	direct counting	**N**	**Y**	Y
Accrued VAT tax	**vat_pnalicz**	VAT	direct counting	**N**	**Y**	Y
Gross value of fixed assets	eblu_srtrw	PIT/CIT VAT	indirect estimate	Y	**Y**/N	N
Outlays on fixed assets	eblu_nakl	PIT/CIT VAT	indirect estimate	Y	**Y**/N	N

Listed variables (results of estimates) match the use of different methods of calculation, estimate, or assessment which use administrative data or jointly administrative data and data from a survey. They are defined in the following way:

[1] Value as regards the previous year

- In the field of the number of entities:
  - pvcit_n – the number of entities for which some information was found in PIT, CIT or VAT sets of the Ministry of Finance (coupled with PIT, CIT or VAT) belonging to a give domain (grouping class)
  - sz_n – estimate of the number of entities running economic activity obtained by correction of the pvcit_n value from the previous year by the change of the number of active entities in the KEP register. While estimating, collectiveness is divided into strata according to the type of activity, the type of entity (legal/natural person) and APE (activity coefficient is differentiated in relation to these features), so the obtained estimate also considers changes in the structure of the sampling frame as regards these features.

$$SZ\_N_{t,d} = \sum_{w}(PVCIT\_N_{t-1,w} * (\frac{KEP_{t,w}}{KEP\_N_{t-1,w}})$$

    where w denotes strata being part of the domain p, KEP_Nt,w – the number of entities active in the KEP register in the year t for the stratum w.
- In the field of revenues from the entire activity:
  - pcit_przych – estimate directly calculated as a sum of incomes from PIT and CIT sets for entities belonging to the domain. Additionally, it is increased by the sum of the turnover from the VAT set for the entities recording income tax and VAT tax, yet for which there is lack of information on incomes (fixed amount tax).
  - eblu_przych – estimate obtained by joining information from a survey and information from administrative sets with the use of indirect estimates. The predictive estimator type EBLUP2 is used, together with the use of a mixed model at the level of the domain. For the carrying out of experimental estimates the following were used as explanatory variables:
    - in the 1st variant: the income determined on the basis of PIT/CIT sets, as well as the participation of entities uncoupled with PIT/CIT sets
    - in the 2nd variant: the turnover determined on the basis of VAT set, the participation of entities recording PIT/CIT and not recording VAT, as well as the participation of entities uncoupled with PIT/CIT sets (the use of VAT data instead of PIT/CIT data enables to obtain results in the prevailing date of realisation SP-3)
  - sz_przych – estimate of the sum of incomes obtained by correction of information on income from the previous year determined on the basis of PIT/CIT sets (pcit_przych) by a coefficient describing the dynamics of turnovers in relation to the previous year which was determined on the basis of information from the VAT set. While doing the estimate, collectiveness is divided into strata according to the type of activity, the type of entity (legal/natural person), as well as APE – these features strongly differentiate the value of income for the entity. The algorithm for calculation of estimates can be presented by the formula:

[2] The description of the method: J. N. K Rao: *Small Area Estimation*, Wiley & Sons, 2003

$$SZ\_N_{t,d} = \sum_{w} \left(\frac{PCIT\_PRZYCH_{t-1,w}}{KEP\_N_{t-1,w}} * \frac{\frac{VAT\_SPRZ_{t-1,w}}{VAT\_N_{t-1,w}}}{\frac{VAT\_SPRZ_{t,w}}{VAT\_N_{t,w}}} * KEP\_N_{t,w} \right),$$

where w denotes strata being part of the domain d, VAT_N – the number of entities accounting VAT tax, VAT_SPRZ – the sum of sales (turnover) defined on the basis of VAT set.

- In the field of costs for the entire activity:
  - pcit_koszt – the estimate is calculated as a sum of costs from PIT and CIT sets for the entities belonging to the domain, while for the entities for which there is lack of information on the value of costs (PIT28) the proportion of costs to income (costs coefficient) is assumed as identical as for the entities for which there is information of the same type (natural/legal person) and of a similar activity (estimate in strata)
  - eblu_koszt – it can be proposed to apply an indirect estimate which uses, e.g. variables occurring in administrative sets, cause and result or correlatively related to the cost of running an economic activity, such as: the tax value, accrued VAT tax, employees, wages and salaries (ZUS). Estimates carried out in this way have never been conducted; it can be a subject for further research.
  - sz_koszt – it can be proposed to use the costs coefficient from the previous year to estimate costs on the basis of the current cost estimate (sz_przych) to calculate the value of the variable solely on the basis of current VAT sets, without the necessity to wait for PIT/CIT data. Estimate carried out in this way has never been conducted
- In the field of VAT tax:
  - vat_pnalezn – the sum of due VAT , calculated directly from the VAT set
  - vat_pnalicz – the sum of accrued VAT , calculated directly from the VAT set
- In the field of the value of fixed assets and outlays on fixed assets
  - eblu_srtrw, eblu_nakl – it can be proposed to apply indirect estimates using information on the value of purchased goods and services included in fixed assets from the VAT set as auxiliary variables together with other variables indicating correlations (e.g. income, number of employees). Such an estimate has never been carried out; it can be a subject for further research.

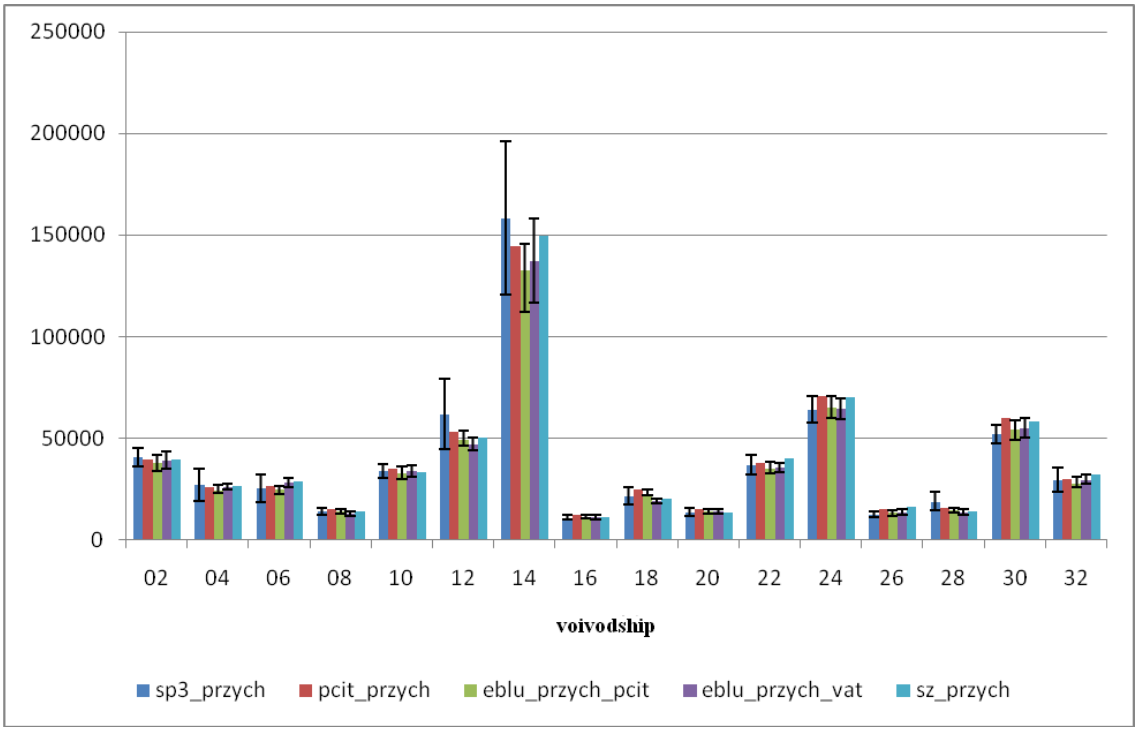
The results of experimental estimates

The results of the experimental estimates were compared with the results of the SP-3 survey, including information on sampling error which occurred in this survey (where it is available). The results of SP-3 were generalised directly from the survey set with the use of weightings included in this set in order to keep the coherence and comparability with the estimates conducted on the basis of the sampling frame SP-3. For the same reasons all the generalisations and estimates by sections are based on the information from the sampling frame and not from the information from the survey. In calculations and during the presentation of the results estimates of the precision of selected survey variables SP-3 defined by the Team of Sampling Methods of the Department of Methodology, Standards and Registers were used. Below there are presented the results of the conducted experiment in the form of diagrams illustrating comparison of estimates obtained with the help of various alternative methods for the sum of incomes by voivodeship and by NACE section with

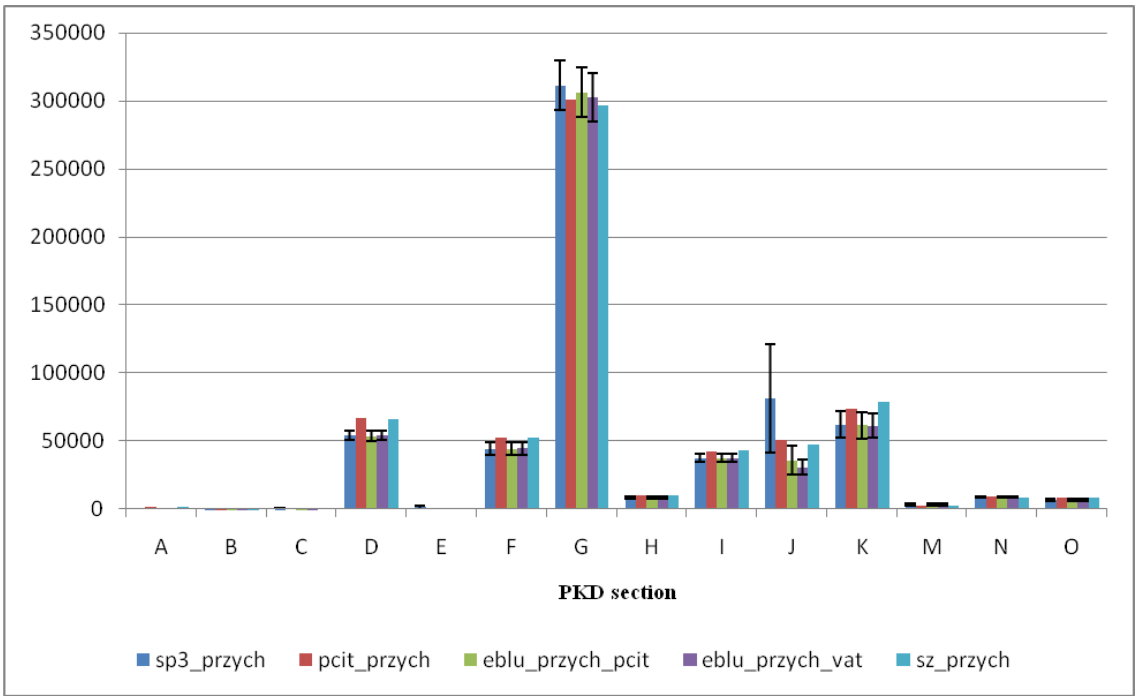
generalisation from the SP-3 set. In case of methods with probabilistic foundations on the diagrams estimates of precision were marked as error bars which correspond to 95% confidence intervals.

There is also a table showing the relative deviations of the estimates obtained by the application of the proposed methods from the results of the SP-3 survey (comparison of percentage differences in the estimates the sum of incomes by voivodeship in 2005).

Comparison of the estimates of the sum of incomes of micro enterprises in mln PLN by voivodeship in 2006 based on the SP-3 survey and with the use of the proposed methods



Comparison of the estimates of the sum of incomes of micro enterprises in mln PLN by NACE section in 2006 based on the SP-3 survey and with the use of the proposed methods



Comparison of percentage differences in the estimates of the sum of incomes by voivodeship in 2005

voivodeship	Estimate				Sampling error SP-3
	SP-3	pcit_przych	eblu_przych (reg. PIT/CIT)	eblu_przych (reg. VAT)	
in relation to SP-3					
02	x	-2.1%	-1.6%	-2.8%	20.8%
04	x	3.1%	1.4%	3.8%	15.9%
06	x	-7.1%	-7.1%	-8.3%	15.3%
08	x	28.3%	6.8%	6.1%	9.9%
10	x	-3.4%	-2.5%	-0.4%	11.6%
12	x	-3.4%	-6.8%	-5.4%	19.9%
14	x	7.7%	0.0%	-1.8%	9.0%
16	x	-4.2%	-2.3%	-7.1%	13.3%
18	x	-12.0%	-14.1%	-17.8%	46.7%
20	x	-4.6%	-8.8%	-8.1%	36.9%
22	x	7.9%	1.3%	4.8%	13.2%
24	x	3.2%	-0.4%	-1.4%	9.3%
26	x	6.8%	3.8%	2.4%	18.5%
28	x	-17.5%	-17.3%	-15.5%	22.5%
30	x	9.0%	1.0%	2.4%	7.9%
32	x	-11.8%	-10.3%	-3.2%	30.4%
Total	x	1.2%	-2.8%	-2.5%	
in relation to accrued value on the basis of the data from sets regarding income tax (pcit_przych)					
02	2.1%	x	0.5%	-0.8%	x
04	-3.0%	x	-1.6%	0.7%	x
06	7.7%	x	0.1%	-1.3%	x
08	-22.1%	x	-16.8%	-17.3%	x
10	3.5%	x	0.9%	3.1%	x
12	3.5%	x	-3.5%	-2.0%	x
14	-7.2%	x	-7.2%	-8.8%	x
16	4.3%	x	1.9%	-3.1%	x
18	13.7%	x	-2.4%	-6.5%	x
20	4.8%	x	-4.4%	-3.6%	x
22	-7.3%	x	-6.1%	-2.9%	x
24	-3.1%	x	-3.5%	-4.4%	x
26	-6.3%	x	-2.8%	-4.1%	x
28	21.2%	x	0.3%	2.5%	x
30	-8.3%	x	-7.3%	-6.1%	x
32	13.4%	x	1.7%	9.7%	x
Total	-1.2%	x	-3.9%	-3.7%	x

The calculations made indicate that the estimates obtained from administrative data describe the values in total, and also reflect properly the variety of features by voivodeship and by NACE section.

The error values vary from a few to several percent, though in most cases they are lower than, or equal to, the value of the SP-3 sampling error estimate. This means that the discrepancies recorded may result mainly from the sampling error in the SP-3 survey.
It is significant that the most considerable discrepancies occurred in the case of information concerning VAT, where we deal with both full consistency in terms of definition and full-scale population coverage with the administrative data. This situation suggests that the major causes for the discrepancies to arise should be attributed to the SP-3 sampling error.
Among the variables under consideration, emphasis was also put on the revenue from total activity; in this case, the number of estimate and estimation variants is the largest. The revenue - is of the most methodological consistency. However, in case of entities who settle their taxes using tax cards - the value of revenues of such entities is not provided. Such entities constitute an insignificant percentage, and usually indicate considerably lower revenues than other entities. The share of entities using tax cards in total revenues of all micro-enterprises, estimated on the basis of SP-3, does not exceed a few percent.
For revenues from total activity, (by indirect estimation) gave. The desirable effect i.e. considerable reduction of the assessments of the mean squared error was achieved. However, due to the fact that the estimates obtained using administrative data only, were of lower quality, these estimates should be considered preferable. Furthermore, the variance and mean squared error estimates can carry an even larger error than the estimates of the very same parameters; hence they should be approached with a certain reservation.
The results of applying the methodology making use of data from current VAT datasets, as well as of the information concerning income tax (PIT/CIT) for the previous year, compiled with the aim of avoiding the problem of delayed availability of the PIT/CIT sets, appear to be of good quality.
The results of estimates using administrative data appear very promising and create the real prospect of using these data in the micro-enterprise statistics (at least for some variables).

5. The analysis of possibilities of administrative data use in sampling.

The idea of using tax data in the sampling frame is related to the separation of the so called “upper stratum”, on which all the entities will be surveyed.

It is connected to the fact that also among enterprises with small employment there are entities having relatively high incomes. Therefore, they can significantly decide about the structure of economic activity for micro-enterprises. During sampling the upper stratum is selected, and includes automatically all entities with incomes above a given threshold. The number of upper stratum is determined numerically to achieve the assumed income precision estimation at the minimum size of sample and is optimized on the base of income from PIT/CIT or turnover form VAT.

The implementation of this method positively influenced on the precision of SP-3 survey results. The positive effect is the better the more valid data are used.

Conclusions:

- Data from administrative systems can positively influence the better sampling allocation (despite its size reduction the maintenance of assumed precision is provided).

- The increase of number in upper stratum seems to be a favorable solution to the results precision.
- Due to their proper validity VAT sets are recommended for use.

Summary

Generally, the results of work are considered to influence positively and beneficially on the survey results. Still, the implementation of proposed methods requires introduction of institutional changes concerning the survey organization and the administrative data acquisition. Desired effects are expected to be observed successively.

So far, changes aiming at broadening the scope of administrative data use in sampling and frame construction have been introduced. In 2010, VAT data from administrative sources are planned to be taken experimentally for direct counting.