Poster Session #2
Time: Thursday, August 9, 2012 PM

Paper Prepared for the 32nd General Conference of
The International Association for Research in Income and Wealth

Boston, USA, August 5-11, 2012

Upper Tail of the Income Distribution in Tax Records and Survey Data: Evidence from Poland

Marek Kośny

For additional information please contact:

Name: Marek Kośny
Affiliation: Wroclaw University of Economics

Email Address: marek.kosny@ue.wroc.pl

This paper is posted on the following website: http://www.iariw.org

Upper tail of the income distribution in tax records and survey data. Evidence from Poland[*]

Marek Kośny[**]
Wroclaw University of Economics

Abstract

Growing interest in the situation of the rich and in changes in the upper tail of the income distribution entails quest for specialized sources of data. Majority of available data sets (e.g. within Luxembourg Wealth Study) come from surveys. Though such surveys are focused on the rich, they are burdened with the same problem as other surveys – reluctance of the rich to inform about their income and assets. This is an incentive to quest for alternative data sources. One of such sources are tax records. Irrespective of their credibility (limited by tax avoidance and tax evasion), the crucial feature of data collected by tax offices is its completeness for the population of taxpayers: refusal to file a tax return or giving there false information is punishable by law.

In this context the primary aim of the paper is the assessment of the reliability of the Household Budget Survey data, which is commonly used in research on affluence and income distribution in Poland. As a reference, complete dataset of tax records for a Dolnośląskie province (over 2 millions of taxpayers) was used.

Besides the analysis of usual measures used in wealth analyses, verification of the compliance with significant-digit law (Benford's law) was performed. Analyses showed significant discrepancies in results obtained for both datasets for upper tail of income distribution and higher reliability of tax record data. Observed differences were important for the assessment of the situation of the rich.

Keywords: affluence, income distribution, survey data, tax records

[*] This paper forms part of the research programme The importance of the shape of the income distribution to the economic growth in Poland (No. 3921/B/H03/2011/40) of the National Science Centre of Poland whose financial support is gratefully acknowledged. I would also like to thank the Directorate of the Tax Chamber in Wrocław for providing tax data for research purposes. Because of their openness and kindness, conducting this research was possible.

[**] Institute of Applied Mathematics, Wroclaw University of Economics, Komandorska 118/120, 53-345 Wroclaw, Poland, e-mail: marek.kosny@ue.wroc.pl

1. Introduction

Studies on the distributions of income focused primarily – for many years – on the situation of the poor. Vast literature on poverty issues discussed both the changes in the distribution of income – in particular the inequality and polarization, and social consequences of these changes – relative deprivation, social exclusion and lack of social stability (see, e.g., Ferrer-i-Carbonell 2005, Esteban and Ray 2011). The beginning of this century brought a significant increase in interest in the other areas of income distribution. The first is the income situation of the middle class. Particular interest in this area was related to the financial crisis that began in 2008. A significant increase in unemployment and limited access to credit contributed to a strong deterioration in situation of households classified as middle class. This process became the subject of detailed analysis not only from the point of view of changes in the level of income, but also – widely understood – economic security of households in this group (see, e.g., Osberg 2009; also Wheary et al. in 2007 for studies prior to the start of the crisis). The second area of research, which has become popular over the last decade, concerns analysis of the situation of the most affluent group. While the situation of the poor is of researchers' interest due to this group itself – to improve chances of its members for the normal functioning in society, assuring equal opportunities, ensuring the minimum resources necessary for life, or to provide at least basic health care - interest in the situation of the richest people, is mainly due to the importance of this group for the whole society. This group, although relatively small (it depends on the definition, but usually consisting of no more than 1% to 10% of the wealthiest individuals), generates a significant part of income and tax revenue. Because of this, it is able to accumulate significant savings and to invest. Through the process of jobs creation and influence on other aspects of the economy, the condition of the richest part of the population directly affects the situation of the less affluent.

Significant discrepancies between the level of income of the poor and the rich are causing problems with a description of the wealth and diversity of the wealth in the whole society. The arithmetic mean and the Gini coefficient, commonly used in assessing the average level and inequality in the distribution of income, depend largely on the structure of the upper part of the income distribution. Sometimes, a few very high incomes are able to significantly distort the results for the entire population. In this context, precise estimates of wealth of the richest parts of the population may have substantial significance for assessment of the situation - for example, in determining the level of inequality and the direction of its change. Therefore, the issue of assessment of the situation of the wealthiest people is inextricably linked with the search for data sources, which provide the most accurate and reliable information on this group.

Analyses that are presented in the literature are based on the data from two main sources. The official surveys, conducted by statistical offices and aimed at collecting data on the income and expenditure of households in the given country, are the first kind of such sources. For many years, the data of this type were the primary (and virtually only included in the study) source of information on

income distributions (see Human Development Reports 1990-2011 and Atkinson and Brandolini 2001 for review of use of "secondary" data sets). But the most serious drawbacks of survey researches concern the assessment of the situation of the wealthiest. The most important disadvantage of this type of research is the relatively small numbers of individuals (or households – depending on the measurement unit) with very high incomes, and refusals to participate in the study. An attempt to solve this problem is oversampling the highest income individuals (households). There are several such surveys (but none for Poland) – data from some periodic studies of this type has been gathered in the Luxembourg Wealth Study Database (see Sierminska et al. 2006 for a description). However, this only allows reducing, not eliminating, problems associated with surveying the most affluent part of population.

An attempt to solve these problems, related to survey researches, is the use of data from fiscal data sets. Although such data for decades were considered rather useless in the analysis of wealth, in the last decade tax records are one of the primary sources of data in this area (see, among others, Piketty and Saez, 2003, Leigh 2009, Atkinson et al. 2011). Despite the undeniable flaw, which is a burden on the data as a result of tax evasion and tax avoidance, their main advantage is the availability for long periods of time. In the long run (in the case of some countries covering more than 100 years) the availability applies only to the basic characteristics of the distribution (for example, the quantiles, which are very useful in the process of approximation of the original distribution of income). But some tax data is also available in the form of individual records. Information on the distribution is then complete (assuming that the data cover the entire population), but such data are typically available for much shorter periods (which is related to the development of information technology). The main advantage of the fiscal data is the completeness of the data on taxpayers' income (and sometimes also on some other variables, such as age, place of residence) for the population of taxpayers[1]. Unfortunately, these sets have same disadvantages as the aggregate tax data.

Both the survey data, and those from tax returns, are used as an independent basis for assessing the situation of the affluent people (an extensive description of the empirical research on the wealth presents Leigh 2009). However, both data sets may also serve for each other as a point of reference. The results of the analysis - for the same country and period – may in fact vary strongly, depending on the data. A comparison of survey and tax data for the United States, conducted by Burkhauser et al. (2009), indicates that the results for the top 1% of the population are much lower for the survey data (high convergence was observed in the rest of the income distribution). In case of discrepancies, however, there is a natural question: which results should be considered more reliable? The answer to this question may result from the analysis of the characteristics of both data sets – what types of data are collected, what kind of information is generally ignored, etc. Also the structure of the

[1] However, not all individuals in the society are taxpayers.

data set, method of gathering the data may be of interest and provide some interesting, additional information.

Detailed analyses of the income of wealthy people in Poland were carried out only on the basis of the survey data so far (see Brzezinski 2010). Therefore, the subject of this paper is to compare the situation of wealthy individuals in Poland, resulting from the data collected in two independent sets. The first data set comes from the household budget survey, conducted annually by the Central Statistical Office. The second contains information on incomes, taken from the official register kept for tax purposes by the tax offices of Lower Silesia (one of the 16 Polish regions). In addition to comparison of the situation of the wealthiest part of the population, there will be analyzed the relationship between theoretical probabilities and empirical frequencies of significant digits in numbers denoting income in both sets. The purpose of this unusual analysis is to evaluate the accuracy and completeness of the collected data.

Structure of this paper is as follows. The second section presents a brief characterization of the two data sets and the definition of income. The third section describes the methods used to analyze and evaluate data. The fourth section presents the results of the empirical analyses. The fifth section briefly discusses the results. And last section concludes.

2. Data on income distribution in Poland

2.1. Household Budget Survey

The widest survey research on the economic situation of households in Poland is a Household Budget Survey (HBS), conducted yearly by the Central Statistical Office. Outcome of this study is very important for the assessment of standard of living in Poland and in determining the number of official, social policy indicators. In particular, HBS results are one of the important factors, taken into account while deciding on the level of the minimum wage. The HBS data is also used to assess the poverty extent, to calculate estimates of taxes paid, and to determine the level of social benefits. Data on consumption and prices of goods are also used to set weights for calculating the consumer price index (see Central Statistical Office 2011a).

To allow - at least partial – inference about changes in the financial situation of households, about half of the sample is surveyed in the consecutive year. The method of sample selection is rotating panel with partial replacement of the sample. This means that households that were surveyed for the first time in a given year are also asked to participate in the survey in the following year. Before 2000 the single household could be surveyed up to 4 years, but since 2000 the maximum period of the participation in the study has been reduced to only two years. This is a very significant limitation from the point of view of the possibility of observing changes over time.

Sampling scheme in the HBS is two-stage, based on a territorial units record TERYT. The units at the first stage are primary sampling units that are based on statistical regions defined in TERYT system. Before the sampling, primary sampling units are stratified by province and city size. As a result, number of primary sampling units from a given strata is proportional to the estimated number of dwellings in this strata. The secondary sampling unit is flat, drawn within the primary sampling units. The study includes all households that are living in the drawn flat (house). Due to such sampling scheme, the final data set is representative for both the Polish population and the population of each of the 16 provinces.

Due to the relatively large proportion of households refusing to participate in the survey[2], to minimize the impact of differences between the structure of the sample and the structure of the entire population, in the final data sets weights are assigned to individual households. These weights[3], which denote the inverse of probability of selecting a particular type of household, are estimated on the basis of year 2002 Census.

To maintain comparability with the tax data, further analysis will take into account data for the years 2006-2010, covering a period of rapid economic growth in Poland (2006-2007), and then stagnation, resulting from the financial crisis (2008-2010). The sample size in the analyzed period was approximately 37 000 households.

2.2. Tax records

Among the data sets collected by the public institutions, particularly useful in studies on distribution of income and wealth are data collected by tax offices. Fiscal data sets include not only information on income, but also amounts of taxes paid and contributions for social and health security. Given the diversity of tax forms, it also allows for the distinction of different social groups – particularly pensioners and self-employed. In addition, since 2007 some personal income tax forms include the question concerning the number of dependent children – this information is required to benefit from the tax exemption. And the number of children, though not always complete[4], allow for a partial analysis of income and public-law burdens in the context of the demographic situation of taxpayers.

[2] This proportion depends on the year and is about 50%. Instead of households refusing to participate in the survey are accepted household selected as a reserve within each primary sampling unit.

[3] Weights are not taken into account in the further analysis. As the average weights are systematically lower for the richest (the group of 0.1%, 1% and 10% the richest) than for the whole sample, their inclusion would result in the decrease in value of averages and respective quantiles. Published weights are, of course, justified in terms of the sampling scheme, but do not comply with distribution of income. As is clear from other studies (see, for example, Moore at al. 2000 for a broad discussion of survey data errors), non-response rates are much higher among the households with the highest income. It means that higher weights should be applied for this part of the sample.

[4] Some taxpayers have income too low to benefit from this exemption. The second problem results from the possibility (but not obligation) of joint taxation of spouses.

The main advantage of administrative data is – in terms of their use for research purposes – compulsory character of its collection. Units (individuals, companies, institutions) covered by a register are legally obliged to provide the required information. Refusal to provide information or providing false information is punishable. Administrative coercion does not entail, of course, the completeness and full credibility of the collected data. Some information is intentionally or unconsciously, falsified or concealed, what is usually the basis for criticism of this type of data. In the context of the analysis of a situation of high-income taxpayers, it should be noted that the potential incentives to underreport income (in the form of tax avoidance or tax evasion) increase with income and tax duty. The potential threat of being punished, however, causes that concealing or misreporting of information cannot simply result from reluctance to participate in the research, as is the case of surveys.

Information on aggregated income and taxes paid, estimated on the basis of tax returns, is published annually by the Ministry of Finance. These publications, however, only include income ranges corresponding to the currently applicable tax thresholds (see, e.g., Ministry of Finance 2011), so they are not useful in studies on income distribution. Therefore analyses presented in the next of this paper will be based on individual, anonymized data, in which one record corresponds to a single income tax return. The data set includes the tax returns filed between 2007 and 2011, reporting the income for the years 2006-2010. Data are collected in the panel form, which allows tracking changes in the situation of individual taxpayers in consecutive years.

The data set includes all the tax returns for personal income and capital gains, filed during that period by residents of the Lower Silesia province (one of 16 provinces of Poland). Therefore, the analyzed data set, although it covers more than 2.3 million taxpayers[5], cannot be regarded as representative for the population of Poland. So, direct comparison between results based on survey outcomes and tax records will be carried out for the part of the HBS data set, concerning the Lower Silesia province.

2.3. Definition and sources of income

In HBS incomes from many sources are recorded – including, among others, loans, insurance payments and income in-kind. In the next of this paper, however, only income from sources subject to personal income tax will be taken into account[6] – due to the possibility of direct comparison of data from this source with data from tax records. The following income sources will be considered: income from employment (group 901), income from self-employment (group 902), social security benefits

[5] Due to migrations, deaths and registration of new taxpayers, the number of taxpayers who file the tax form in any given year is lower – about 1.8 million.

[6] Other sources of income recorded in the HBS (such as alimony, social assistance benefits) are not significant in the case of wealthy individuals.

(especially pensions, group 905) and revenue from the sale of products and agricultural services (group 911111). Revenue from the latter group is only partially subject to the personal income tax. But there is no possibility of separation of taxed and untaxed income, based on available data. For this reason, all income from this group is included.

Income from these sources, declared by the respondents, is the gross income – before personal income tax, but after deduction of the compulsory social and health security (if applicable). It is recorded in one month of the year in which the household is surveyed (each month about one twelfth of the whole sample is surveyed). Therefore, when making direct comparisons between the survey and tax data, declared income will be multiplied by 12.

In the case of tax data, income is directly related to the type of tax return filed. The study will take into account the following types of tax returns, characterized by the symbol of the form on which the return is filed.

- Form PIT36, covering income from self-employment and certain types of agricultural production, taxed at the standard, progressive tax scale.
- Form PIT36L, covering income from self-employment and certain types of agricultural production, taxed at a flat rate.
- Form PIT37, filed by taxpayers not running their own businesses, covering income mainly from hired work.
- Form PIT40, including income from hired work, if tax return is filed by employer on behalf of the taxpayer (upon the request of the taxpayer).
- Form PIT40A, including income from pensions, if tax return is filed by Social Insurance Institution on behalf of the taxpayer. If a taxpayer get an additional income (or file the tax return himself for some other reasons), forms PIT36 or PIT37 are used.
- Form PIT38, including capital gains. Income from interest on deposits and from some investment funds is not declared, since the tax on income from these sources is paid automatically by a bank or fund management company. This means underestimation of the actual income from this source.

Basing on the type of tax form filed, in the next of the paper three categories of income will be considered. The first will include income from employment and social security – forms PIT37, PIT40 and PIT40A (equivalent to groups 901 and 905 in HBS). The second category will include income from self-employment and certain types of agricultural production – forms PIT36 and PIT36L (corresponding to groups of 902 and 911111 in HBS). The third category – capital gains, reported on form PIT38 – are not recorded in HBS. Taxpayer receiving income from capital has to file form PIT38. Other forms, however, are filed alternatively, depending on the source of income. However, form PIT36 has priority over PIT37. It means that in the case of income from both the self-

employment and some other source, the taxpayer files form PIT36, placing in it the income from various sources. This can lead to some overestimation of income from self-employment.

Forms PIT36 and PIT37 may be filed by the taxpayer individually or jointly with a spouse or dependent child (in the case of single parents). Joint taxation is, however, facultative and the taxpayer may, but need not use it. This means that the choice of form of taxation is not uniquely determined by the taxpayer's family situation - as already mentioned, the data collected by the tax authorities allow only a partial identification of the family situation. Therefore, data from tax returns will be given for each individual taxpayer – even if the joint taxation has been chosen, most of the information (including income) is declared for each taxpayer separately.

Such definition of the basic unit is inconsistent with the definition adopted for the HBS data, where the basic unit is the household. But it is not possible to find an unambiguous solution to this problem, because in the HBS income is not recorded separately for each individual in the household. Any division of household income among individuals (based on household composition) would require estimates of the number of people earning income subject to personal income tax. It also needs assumptions about intra-household income distribution, which usually depends on income level and is very asymmetric in the case of households with high incomes. Therefore, the original structure will be maintained for both data sets, and the characterized differences will be taken into account in the analysis.

3. Methods

To assess the situation of the wealthiest part of the population on the basis of data from both presented data sets, there will be used both traditional methods, usually applied in this type of analysis (see Brzezinski 2010), and some less common. To the first of these groups belong estimates of quantiles of the income distribution, the average income in groups of individuals (households) with the highest income and the income shares of the top *p*%, expressed by the formula:

$$IS(\mathbf{x}, p) = \frac{\sum_{i=\lceil n \cdot p \rceil}^{n} x_i}{\sum_{i=1}^{n} x_i} \tag{1}$$

where $\lceil n \cdot p \rceil$ denotes rounding up to the integer closest to $n \cdot p$, n – number of individuals (households) and p – quantile of the income distribution ($0 \le p \le 1$) for $\mathbf{x} = (x_1, x_2, ..., x_n)$ being an ordered vector of non-negative values $x_1 \le x_2 \le ... \le x_n$, representing the distribution of income.

Typical measures that are also popular in the analyses of affluence presented in the literature are indexes of mobility, expressed as the probability of leaving (by the person or household) the group of top income earners (depending on the definition of this group) in a given period (see, for example, Saez and Veall 2005).

Besides these popular measures, the analysis presented in this paper will be expanded on the analysis of the curves, based on point Zenga index and on analysis of goodness-of-fit of empirical frequency distributions of the significant digits (of the numbers denoting income) and the Benford's distribution.

The first one of these analyses refers to the idea that underlies Zenga's indexes and describes the relationship (and its changes) between groups of people (households), whose income is below and above a specified level. The measure, which is a generalization of the quotient of the quantiles of the income distribution, holds the information not only on the situation at a given point (quantile) of the income distribution, but on the entire groups - the lower and upper. This is especially important for the analysis of wealth, because it takes into account the effects of very few, but very high incomes. A point measure of a relative income change, denoting changes in the distribution of income in a given period of time will be defined as (see Kosny 2011):

$$RIC(\mathbf{x^0}, \mathbf{x}^1, p^0, p^1) = \frac{\bar{M}(\mathbf{x^1}, p^1)}{\overset{+}{M}(\mathbf{x^1}, p^1)} - \frac{\bar{M}(\mathbf{x}^0, p^0)}{\overset{+}{M}(\mathbf{x}^0, p^0)} \tag{2}$$

where $\mathbf{x^0}$ and $\mathbf{x}^1$ denote distributions of income at the beginning and the end of the period respectively. The line separating two groups – the poorer and the richer – is set by p^0 and p^1 for these two moments. The lower and the upper mean will be defined as $\bar{M}(\mathbf{x}, p) = \frac{\sum_{i=1}^{\lfloor n \cdot p \rfloor} x_i}{\lfloor n \cdot p \rfloor}$ and $\overset{+}{M}(\mathbf{x}, p) = \frac{\sum_{i=\lfloor n \cdot p \rfloor + 1}^{n} x_i}{n - \lfloor n \cdot p \rfloor}$ where $\lfloor n \cdot p \rfloor$ denotes rounding down to the integer closest to $n \cdot p$.

The values of the point index given by (2) range between -1 and 1. They reflect changes in the average income of the lower group with respect to the upper group. For a given *p*, *RIC* indicates the change (expressed in percentage points) in the share of the average income of the lower group (100% · *p* of the population) in relation to the average income of the upper group (100% · (1-*p*) of this population).

The second analysis is related to the observations made by Newcomb and Benford (see Hill 1995). Their observations show that in some data sets the empirical frequencies for the first significant

digit are not uniformly distributed. Moreover these frequencies – for completely different data sets – follow the same rule that can be written in the form of the following equation:

$$P(D = d) = \log_{10}\left(1 + d^{-1}\right) \tag{3}$$

where D denotes the most significant digit and $d \in \{1,2,...,9\}$. Hill (1995) proposes a formal analysis of the conditions that must be met so as the frequency distribution of digits at the most significant position (of numbers in a given set) is consistent with the distribution (3). Despite this, there is no clear, empirically useful rule indicating data sets that should comply with Benford's law. But analyses that are aimed at assessing the reliability of various types of data by studying the goodness-of-fit of empirical frequencies and the theoretical distribution are primarily done for financial data (on income, taxation and the volume of transactions – see, for example, Nigrini 1996, Durtschi et al. 2004, Cleary and Thibodeau 2005, Cho and Gaines 2007). Postulates formulated by Hill (1995) are – at least in theory – met for this type of data: individual observations are randomly drawn from the randomly chosen distributions (reflecting different sources of income[7]) and data are scale and base invariant (for each source of income).

Of course, the empirical analysis of the goodness-of-fit of empirical distributions for the two analyzed data sets and the theoretical (reference) distribution is not able to give an unambiguous answer to the question about the reliability of data. However, it enables assessment of the data collection process – to what extent has it been distorted by the systematic errors? In this sense, the greater the distance from the distribution given by (3) shall indicate a lower quality of the collected data[8].

In the next of the paper will be used an extended version of Benford's distribution, which allows assessment of the frequency of digits, not just at first, but also at the following significant positions. This distribution is then given by (see Hill 1995):

$$P(D_1 = d_1,..., D_k = d_k) = \log_{10}\left[1 + \left(\sum_{i=1}^{k} d_i \cdot 10^{k-i}\right)^{-1}\right] \tag{4}$$

where d_k denotes digit at the *k*-th most significant position, $d_1 \in \{1,2,...,9\}$, $d_j \in \{0,1,2,...,9\}$ for $j = 2,\ldots, k$.

The goodness-of-fit of the empirical distributions and the theoretical distribution will be conducted on the basis of the modified chi-square statistics, given by the following formula:

---

[7] In this case, the source of income should be interpreted more broadly than just as income from work, self-employment, etc. The sources of income in this sense may be, for example, companies.

[8] As will be shown in the empirical part, the differences between the sets are relatively constant. And modifications of the original data sets – removing parts of the data according to source of income or level of income – cause a significant decrease in the goodness-of-fit to the theoretical distribution.

$$DI = \sum_{i=1}^{9} \frac{(p_i - w_i)^2}{p_i} \tag{5}$$

where p_i denotes the theoretical probability of digit i and w_i the corresponding empirical frequency. In the case of analysis of the distribution for the second and consecutive significant digits, i takes values from 0 to 9. The modification with respect to the original chi-square statistics is enforced by the sample size. Large samples, together with huge differences in the size of both samples, would results in estimates that are – in fact – not comparable among the data sets. Unfortunately, the drawback of such a modified measure is the lack of knowledge of its distribution, and thus the inability to assess the statistical significance of observed differences and goodness-of-fit of empirical distributions with the Benford's distribution[9]. Therefore, attention will be focused primarily on assessing the direction of dependency – in which case discrepancies are larger and in which smaller.

4. Empirical evidence

4.1. Quality of data

Both analyzed data sets are burdened with systematic errors. In HBS, as for other surveys, the basic problem is nonresponse. This issue is of particular importance in the context of research on wealth. In the case of households with very high incomes, very important factor is the desire to protect the privacy and reluctance to share with others the details of both income and consumption. Also need to fill in a quite complex questionnaire on the structure of consumption, which is a labor intensive activity, is discouraging – the amount of money offered for the participation in the survey is often not a sufficient incentive for high income households. Besides nonresponse rates, depending on the household income[10], deliberate concealment of information has to be classified as non-random distortions of the survey results. In some cases – for example if income is derived from illegal sources – the respondent may misreport the income because of a fear that the data collected will be provided to the tax authorities (tax evasion, resulting from the underreporting the income from work, is not usually considered very negative in Poland). In other situations the problem results mostly from social norms – spending on prostitution and drugs is usually zero in HBS data. The third source of systematic errors, important from the point of view of this analysis, is the lack of knowledge of respondents regarding the definition of certain categories or amounts of their actual spending in these areas. The values of personal income tax, declared by the respondents, are an example of this type of errors, as distribution of personal income tax in the sample is completely different that from the entire population. Finally, the fourth source of error in assessing the actual situation of households is the lack

[9] Due to the sample size – especially in the case of data from tax records – standard goodness-of-fit tests indicate significant differences.

[10] The Central Statistical Office publishes only the overall percentage of households that refused to participate in the study (see Central Statistical Office 2011b).

of some (income) categories in recorded information. In the context of the most affluent households, capital gains are the most important such category.

Besides non-random errors, all surveys involve also random errors. As the group of the most affluent people is relatively small, samples that are said to be representative for the entire population may not adequately reflect this subgroup. Certain drawbacks are also included in the sampling scheme itself (not actual data in the TERYT system) and in a process of collecting the information (reporting the income and expenditure for a single month, filling the questionnaire by respondents).

Other problems are associated with the data from tax records. In this case, the most important source of discrepancies of the reported data with the actual income is tax evasion (e.g., failure to file a tax return at all) or tax avoidance (formally legal, but unintended by authority, reduction of tax duty – e.g. transferring the income to tax havens). Although both phenomena relate to taxpayers at all income levels, in the case of lower-income taxpayers the basic problem is the income from unreported employment. In the case of taxpayers with the highest incomes there are many legal (in a formal sense) possibilities to reduce the taxable income by transferring it to the sources (or areas) with lower tax rates.

The problem, inevitably related to tax data, is also their selective character. They do not include - despite the efforts of the administration, aiming to maximize the tax base - all sources of income. This means not only underestimation of the actual income, but also incompleteness of records of taxpayers held by tax offices. Untaxed sources of income, however, have relatively greater importance for the less wealthy taxpayers.

Additionally, in the case of taxes paid by the institution (e.g. bank, for gains from investment), the amount of tax paid is not directly attributable to the taxpayer. In this way, the income from this source – which could be important in the case of the most affluent – is not reflected in the tax returns.

The above list does not exhaust all the factors that negatively affect the completeness and reliability of the data in both data sets. But the discussed problems are the most important with regard to the group of the most affluent taxpayers and their households (a common feature of both HBS and tax records is that in both cases the actual values of income are understated). Because of mentioned shortcomings, the very precise comparison of results for both data sets is not justified – it is much more important to assess whether the differences can be considered large or small, and whether the direction of discrepancies is stable over time.

4.2. Situation of the highest income groups

As already mentioned in Section 2, a set of tax data is not representative for the population of Poland. It covers a number of nearly 2 million taxpayers filing each year a tax return in the tax offices

in Lower Silesia. Therefore, besides the results for this data set, Table 1 presents parameters calculated on the basis of data from HBS, calculated both for the whole Poland, and for the Lower Silesia province. The values of all quantiles in this table take account of inflation and are expressed in U.S. dollars (an average annual exchange rate from 2006 has been used).

Table 1. Sample size and quantiles of the income distribution

Quantile	Data set	Year				
		2006	**2007**	**2008**	**2009**	**2010**
0.25	Tax records	2 387	2 453	2 719	2 851	2 895
	HBS – Lower Silesia	4 447	4 552	5 257	5 566	5 736
	HBS – Poland	4 640	4 838	5 400	5 566	5 736
0.50	Tax records	3 953	4 122	4 576	4 777	4 890
	HBS – Lower Silesia	7 079	7 531	8 659	9 044	9 177
	HBS – Poland	7 349	7 761	8 641	8 870	9 177
0.75	Tax records	6 852	7 326	8 186	8 397	8 568
	HBS – Lower Silesia	10 750	11 243	12 961	13 914	13 799
	HBS – Poland	11 178	11 901	13 177	13 566	13 833
0.90	Tax records	11 565	12 297	13 547	13 900	14 153
	HBS – Lower Silesia	15 471	15 880	18 177	19 657	20 215
	HBS – Poland	16 342	17 480	19 041	19 480	20 205
0.99	Tax records	39 161	43 117	46 134	45 031	45 637
	HBS – Lower Silesia	33 964	31 445	37 963	40 594	38 125
	HBS – Poland	38 601	40 166	41 916	42 544	44 654
0.999	Tax records	**169 863**	**206 057**	**204 199**	**176 385**	**175 709**
	HBS – Lower Silesia	**83 752**	**79 881**	**65 240**	**93 360**	**144 389**
	HBS – Poland	**92 488**	**107 328**	**101 697**	**98 058**	**102 556**
Sample size	Tax records	1 772 975	1 807 136	1 833 325	1 824 117	1 823 372
	HBS – Lower Silesia	2789	2797	2824	2829	2831
	HBS – Poland	36 068	36 150	36 332	36 235	36 278

The differences between Poland and the Lower Silesia province in the HBS data are not significant below the quantile of order 0.90, but the values for the Lower Silesia are usually slightly lower. Above this quantile differences increase (except quantile of order 0.999 in 2010 – but it can be assumed that this is due to the lack of representativeness, resulting from too small sample size).

Much bigger differences exist between tax data and data from HBS. The average values of ratios of quantiles are presented in Table 2.

Table 2. Average relation between tax record quantiles and HBS quantiles for Lower Silesia

Quantile	0.25	0.50	0.75	0.90	0.99	0.999
Average share	52%	54%	63%	73%	121%	217%

Differences in the bottom of the distribution are related to, inter alia, the unit adopted in both data sets, which is person (taxpayer) in case of tax data, and household for HBS. Due to the lack of information on the number of individuals achieving income subject to personal income tax in HBS data, the unification of units is not possible. However, lower values for the tax data in the lower part of the income distribution indicate the potential compatibility of both sets. Assuming an average number of income earners per household[11] equal to 2, the share for the first and second quartile, amounting to approximately 50%, can be interpreted as the equal distribution of income within the household. The increase in the ratio for the higher quantiles (of order 0.75 and 0.9) suggests – assuming the same interpretation – the increase in asymmetry in the intra-household income distribution. It is reasonable that the probability of achieving a comparable income for all individuals earning income in the household decreases with household income.

In the highest areas of income distribution the relationship between results for both data sets is, however, completely reversed. Incomes, according to the tax records, are on average more than twice higher than in HBS data (as already mentioned, data from the HBS for Lower Silesia in 2010 seem to be random, as in the case of Poland so significant change with respect to the year 2009 does not exist). The use of any multiplier that reflects the relationship household – individual person (the taxpayer) would even widen the observed differences between tax data and HBS data[12]. Therefore, the only explanation for these differences is very significant underestimation of the income from the highest areas of income distribution in the HBS data.

It is also worth noting that regardless of the economic slowdown that occurred in Poland after 2008, real incomes of people with lower incomes (up to the quantile of order 0.90) grew in real terms over the period under analysis. The financial crisis has affected the situation of the richest, whose incomes have fallen in real terms between 2007 and 2010.

[11] In the analyzed period, the average number of people over 18 years per household was about 2.3, but not every adult person derives income subject to personal income tax.

[12] Even if this value is slightly below 1, what would mean that in the highest income households asymmetry in intra-household is typically very large (very high incomes are usually achieved by only one person in the household, irrespectively of household's size).

Quantile measures, presented in Table 1 indicate, however, only income thresholds, and neglect inequality within each group. So, results presented in Table 3 – averages and standard deviations (figures in U.S. dollars, after adjusting for inflation) – supplement earlier observations.

Table 3. Averages and standard deviations of incomes in the upper part of the income distribution

Quantile	Measure	Data set	Year				
			2006	**2007**	**2008**	**2009**	**2010**
0.90-0.99	Mean	Tax records	**17 935**	**19 296**	**20 969**	**21 224**	**21 689**
		HBS – Lower Silesia	**20 025**	**20 701**	**23 223**	**25 318**	**25 401**
	Standard dev.	Tax records	6 307	6 994	7 397	7 103	7 251
		HBS – Lower Silesia	4 074	3 835	4 356	5 007	4 355
0.99-0.999	Mean	Tax records	**67 110**	**77 305**	**80 176**	**73 878**	**74 232**
		HBS – Lower Silesia	**45 688**	**43 960**	**48 667**	**54 422**	**50 845**
	Standard dev.	Tax records	29 542	36 266	35 690	29 571	29 425
		HBS – Lower Silesia	11 616	12 232	7 793	13 852	19 376
0.999-1.00	Mean	Tax records	**412 421**	**524 407**	**514 425**	**386 392**	**396 745**
		HBS – Lower Silesia	**143 478**	**115 797**	**81 079**	**120 601**	**247 377**
	Standard dev.	Tax records	797 573*	771 934*	814 744*	557 035*	593 261*
		HBS – Lower Silesia	57 188*	45 280*	20 299*	6 543*	79 588*

* Results not reliable due to the sample size (n = 2)

Taking into account all observations, the average incomes in the highest group are on average almost four times higher for tax records than for the HBS data (this ratio varies between 1.60 in 2010 and 6.34 in 2008). The differences between data sets have increased also for the two other groups. For all analyzed groups and periods, tax records are also characterized by significantly higher values of standard deviation. This means that the survey data only partially reflect the variance among the highest incomes. And the variance in this area of income distribution largely affects the measures of inequality and polarization, estimated for the entire distribution of income.

Changes in the relative situation of different groups, being a consequence of the observed changes in wealth, are presented in Figure 1. Presented values of *RIC* indexes denote the change (in percentage points) in the share of an average income for the group of people with incomes lower than a given quantile in the average income of a group with higher incomes. In each case the reference point was the preceding year. The results for the quantiles of order 0.999-1.00 for HBS data have not been presented due to the sample size.

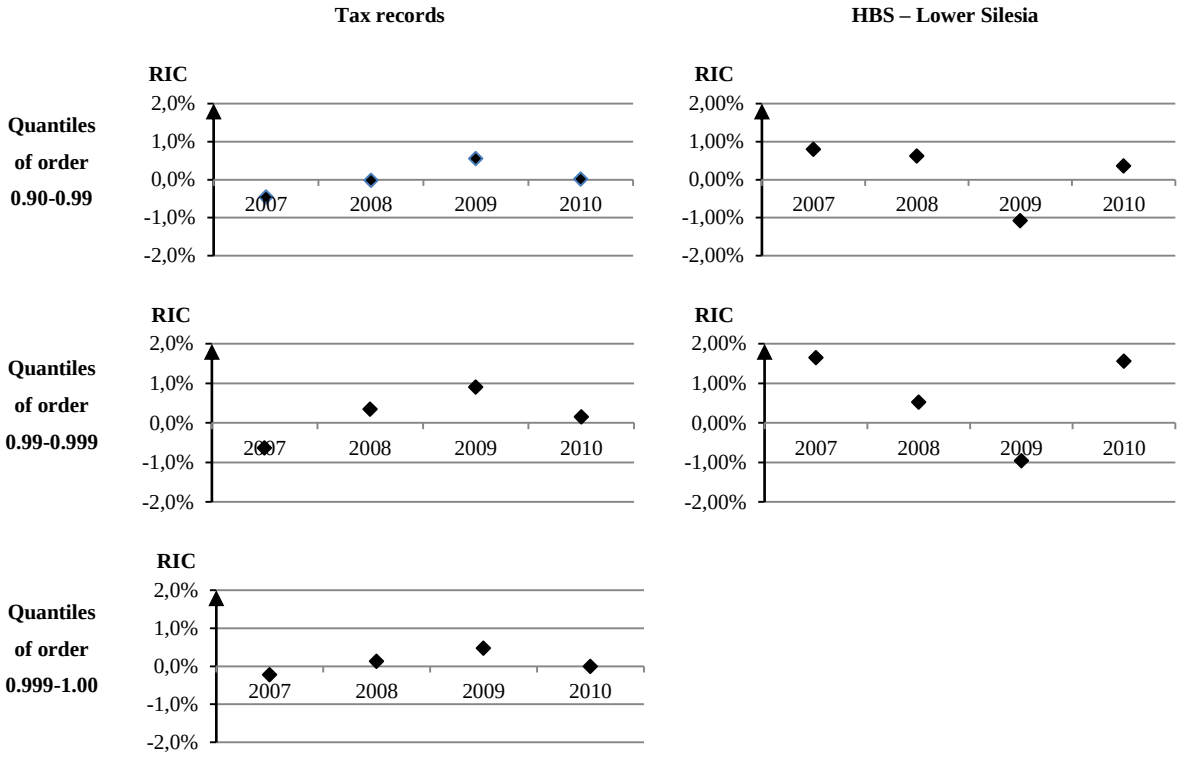

Figure 1. Relative income changes

As results from presented graphs, differences between both data sets include not only the absolute value of the income of wealthy individuals. While tax data represent an improvement in relative situation of more affluent with respect to the poorer in 2007 (negative values indicating reduction of the income share of the poor in relation to the income of more wealthy individuals) and a decline (or no change) in subsequent years, whereas HBS data indicate the relative improvement in the situation of wealthy households in 2009 and a deterioration in the remaining periods. Due to the fact that the financial crisis in the first place affected the income of wealthy individuals, the relative improvement in the situation of the poor in the early years of the crisis seems to be a more likely scenario.

The analyses presented so far were aimed at describing the income situation (relative and absolute) of particular groups. But from the standpoint of society as a whole, particularly important is the share of income derived by individuals (households) in the group in the total income of the population. Analyses of this type, often presented in the literature, are designed to answer the question of the concentration of income. The data presented in Table 4 represent the share of income of the group in the aggregate income of entire population.

Table 4. Income shares

Quantile	Data set	Year				
		2006	**2007**	**2008**	**2009**	**2010**
0.90-0.99	Tax records	25%	25%	25%	26%	26%
	HBS – Lower Silesia	21%	21%	21%	21%	21%
0.99-0.999	Tax records	9%	10%	10%	9%	9%
	HBS – Lower Silesia	5%	4%	4%	5%	4%
0.999-1.00	Tax records	6%	8%	7%	5%	5%
	HBS – Lower Silesia	1% *	1% *	1% *	1% *	2% *

* Results not reliable due to the sample size (n = 2)

Particularly large gaps between the analyzed data sets can be seen in the case of highest income taxpayers, whose share in total income of the population according to the HBS is significantly underestimated. Also in the case of a group with incomes between the quantile of order 0.99 and quantile of order 0.999, the share is two times higher for tax data. Such a significant underestimation of income in the highest income groups denotes a misstatement of the actual ability of these groups to accumulate income and to influence the economic growth.

Understatement of income in groups of individuals (households) with the highest income is the result of several factors, characterized in general at the beginning of this section. One of possible reasons is lack of 'capital gains' category in HBS, while the importance of this source of income increases with income. Analysis of the percentage of income derived from particular sources is presented in Table 5.

As follows from the presented data, the share of income from capital increases with an increase in total income. But the amount of capital gains – not included in the HBS data – does not fully explain the discrepancy between the values observed for the highest income group in both analyzed data sets. It is also worth noting that the importance of capital gains significantly reduced during the financial crisis – the share of capital income decreased approximately 3 times between 2006 and the 2009. This is because the greater risk aversion and the transfer of assets to more secure investments, and due to the losses caused by declines in the stock market.

Table 5. Income sources

Data set	Quantile	Source of income	Year				
			2006	2007	2008	2009	2010
Tax records	0.90-0.99	Own business	28.1%	27.6%	27.4%	27.9%	27.9%
		Hired work, social security	70.3%	70.8%	71.9%	71.6%	70.9%
		Capital	1.6%	1.6%	0.7%	0.5%	1.2%
	0.99-0.999	Own business	64.9%	67.9%	70.2%	66.2%	64.3%
		Hired work, social security	31.9%	28.4%	28.6%	32.7%	33.9%
		Capital	3.2%	3.7%	1.2%	1.1%	1.9%
	0.999-1.00	Own business	75.1%	78.3%	82.9%	86.3%	84.1%
		Hired work, social security	6.9%	4.3%	4.5%	6.8%	7.9%
		Capital	17.9%	17.4%	12.6%	6.9%	7.9%
HBS – Lower Silesia	0.90-0.99	Own business	23.3%	24.0%	20.6%	21.1%	21.2%
		Hired work, social security	76.7%	76.0%	79.4%	78.9%	78.8%
	0.99-0.999	Own business	38.5%	39.9%	36.7%	34.8%	30.5%
		Hired work, social security	61.5%	60.1%	63.3%	65.2%	69.5%
	0.999-1.00	Own business	70.5% *	75.4% *	51.6% *	77.0% *	25.0% *
		Hired work, social security	29.5% *	24.6% *	48.4% *	23.0% *	75.0% *

* Results not reliable due to the sample size (n = 2)

In addition to capital income, with the increase in total income, also income share from self-employment is growing. In the case of the highest income group this is the main source of income. The contribution of income from this source to the total income, however, is systematically lower in the case of HBS data. To some extent this may be due to reporting some part of income from labor or social security benefits using the form PIT36 – in the case of revenue from several sources. However, almost two times lower share of income from self-employment, resulting from the HBS data for the group between quantiles of order 0.99 and 0.999 suggests serious underestimation.

A very important element in assessing the situation in the upper part of the income distribution is the stability of the composition of each group. Estimates of mobility of the high income earners, defined as a probability of leaving the group in the given period of time, are presented in Table 6.

Tax data are gathered in the form of panel and allow analysis of the taxpayers' situation throughout the period 2006-2010. Panels in the HBS data are only two-year and cover about half of the sample. So, this gives the opportunity to analyze the situation in the next year, but does not allow estimating the probability of leaving the group for the entire period.

Table 6. Mobility of the top income earners

Data set	Group definition - quantiles	Probability of leaving the group in the period				
		2006-2007	2007-2008	2008-2009	2009-2010	2006-2010
Tax records	0.90-1.00	22%	22%	22%	20%	39%
	0.99-1.00	31%	31%	33%	30%	53%
	0.999-1.00	42%	42%	44%	42%	63%
HBS – Lower Silesia	0.90-1.00	35%	36%	37%	42%	-
	0.99-1.00	56%	55%	48%	58%	-

The probabilities presented in Table 6 increase with income, but are relatively stable over time (for annual periods), especially for tax data. Estimates of mobility on the basis of both data sets, however, differ significantly – much higher values were obtained in the case of HBS data. Taking into account differences in research units – taxpayers for tax records and households, in which usually more than one person derives income, for HBS – greater stability should be observed in case of HBS data (even taking into account the already mentioned impact of intra-household income distribution). This may suggest that among households with the high income, which are to be surveyed in the second year, nonresponse rates are lower for household which experienced decline in their income (and left the upper group) – payment associated with participation in the study may be relatively more important for them. For tax data, some impact on estimates of mobility can have the difference between the size of the entire analyzed population of taxpayers (approx. 2.3 million) and the number of taxpayers who file a tax return each year (about 1.8 million). This difference is – to some extent – due to migration between provinces. If the high-income taxpayer migrates from the area of Lower Silesia region because of the worsening his economic situation, the case would not be included in the analysis. However, such situations are not frequent.

4.3. Compliance with significant-digit law

Analysis presented in the previous section indicated the existence of significant discrepancies in the assessment of the income of individuals (households) with the highest income. Some factors which can negatively affect the reliability of the data in both sets have been discussed. Analysis of goodness-of-fit of the empirical distributions of significant digits (in numbers that denote income in both sets) may suggest in which data set the negative, non-random factors had stronger impact on results. Therefore, analysis of the data from HBS will be carried out for the whole set, not just a part concerning Lower Silesia province.

Empirical distributions of frequencies of the most significant digits are presented in Figures 2 and 3 for the tax and HBS data respectively. Despite the discrete character of these distributions, lines

connecting the points were drawn to facilitate the analysis of both charts. Theoretical distribution is marked with bold line.

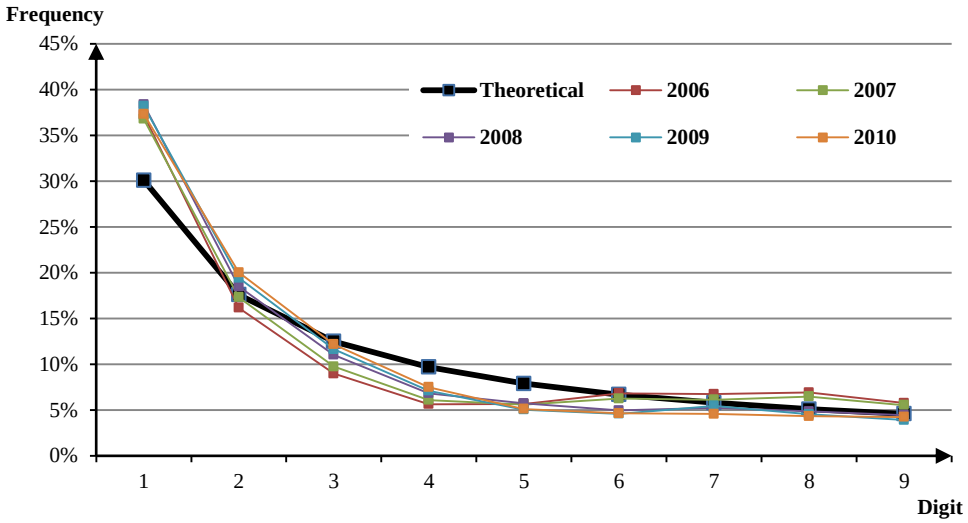

Figure 2. Comparison of frequencies for the theoretical distribution and distribution for tax records

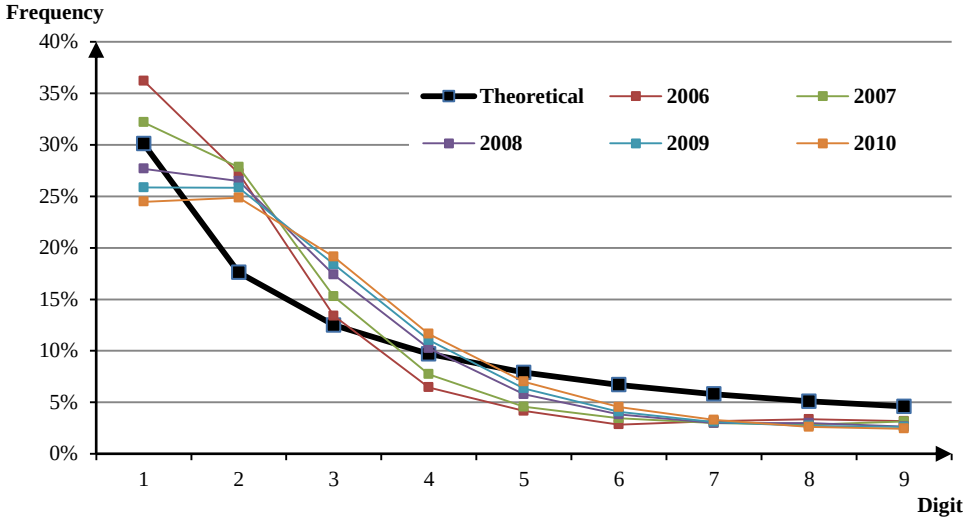

Figure 3. Comparison of frequencies for the theoretical distribution and distribution for HBS data

Higher goodness-of-fit of empirical distributions for tax data and the theoretical (reference) distribution, observed in the Figures, is confirmed by the distance measures presented in Table 7. Fit of the data from tax returns is systematically higher in analyzed period.

Table 7. Distances (*DI*) between empirical and theoretical distributions

Data set	Year				
	2006	**2007**	**2008**	**2009**	**2010**
Tax records	0.062	0.047	0.045	0.050	0.046
HBS - Poland	0.075	0.076	0.075	0.082	0.084

Differences in the goodness-of-fit may partly result from differences in sample size. But the most important seem to be the pattern of changes in empirical distributions for the HBS data, which remains unchanged in the period under analysis. Frequencies of digits 1 and 2 were systematically decreasing, while frequencies of digits 3-6 – increasing. Such a change, associated with overall increase in the average level of income[13], can suggest that the HBS data cover only a certain part of the entire income distribution – a similar effect, but at a much greater extent, can be obtained by calculating the empirical distributions for the part of the sample – see Table 8 (data for 2010) for details.

Table 8. Distributions for selected subsets

	Digit								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Theoretical distribution	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%
Empirical distribution tax records: upper 10% of the sample	19,5%	3,5%	1,3%	0,7%	29,6%	18,6%	12,2%	8,5%	6,1%
Empirical distribution tax records: form PIT40A	56,1%	11,3%	2,5%	1,0%	1,8%	4,0%	8,8%	8,2%	6,4%

Presented results concern only the distribution of the first significant digit. However, the analysis of distributions of next significant digits may be even more useful in evaluation of the quality of the data. They allow, inter alia, determining whether the declared amounts are not rounded. If they are rounded, it means that the data are not recorded in full conformity with the facts, but that reported data are transformed in a certain way. Empirical distributions for the first three significant digits in both analyzed sets are presented in Table 9 (data refer to the year 2010, but the results are similar for other years).

[13] In nominal values, this increase was much higher than presented in this analysis.

Table 9. Comparison of distributions for three most significant digits

	Position	Digit										DI
		0	1	2	3	4	5	6	7	8	9	
Theoretical distribution	**1**		30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	**5.8%**	5.1%	4.6%	
	2	12.0%	11.4%	10.9%	10.4%	10.0%	9.7%	9.3%	9.0%	8.8%	8.5%	
	3	10.2%	10.1%	10.1%	10.1%	10.0%	10.0%	9.9%	9.9%	9.9%	9.8%	
Empirical distribution - tax records	**1**		37.3%	20.0%	12.2%	7.5%	5.1%	4.7%	4.6%	4.3%	4.3%	**0.046**
	2	11.0%	10.9%	12.3%	10.8%	10.2%	9.9%	8.9%	9.6%	8.2%	8.2%	**0.003**
	3	10.7%	10.4%	9.6%	9.7%	9.7%	10.9%	10.6%	9.7%	9.6%	9.2%	**0.002**
Empirical distribution - HBS	**1**		26.2%	24.0%	17.8%	11.1%	6.9%	4.8%	3.5%	3.0%	2.9%	**0.084**
	2	**14.1%**	9.8%	10.7%	9.6%	9.5%	**11.7%**	9.1%	8.6%	9.1%	7.9%	**0.012**
	3	**36.3%**	6.5%	7.3%	6.6%	6.4%	**10.3%**	7.1%	6.2%	7.5%	5.7%	**0.763**

The data presented in Table 9 clearly show that – as expected – the incomes reported in tax returns are not rounded. Empirical distributions observed for second and third significant digit fit the theoretical distribution, as the tax offices require the very precise information on income. The situation is completely different in the case of data from the HBS. A very significant increase in the frequency of the digit 0 (to a lesser extent, also for the digit 5) on the third significant position suggests that the declared income is not given exactly – even though the reporting of income in HBS is carried out relatively precisely, by source of income[14].

5. Discussion

Assessment of situation the wealthy individuals (households) with respect to the absolute changes in the income situation, is similar for both data sets. In the analyzed period, real income grew in subsequent years for all groups below the quantile of order 0.99. Differences between data sets appeared only in the case of the highest percentile, for which the real value of income declined in different periods in both data sets.

Different results were obtained in the assessment of the relative situation of the wealthy people. Analysis of the relative income change – the relationship between income of wealthy and non-wealthy individuals – made for the tax data, indicated the improvement in the relative situation of the wealthy individuals only in 2007 (with respect to 2006). A similar analysis, carried out for the HBS data, showed the improvement of the situation of wealthy individuals in 2009 (with respect to 2008).

[14] A similar analysis for another data set (coming from the largest, independent study of quality of life in Poland – Social Diagnosis) has shown even larger scale of rounding. Huge deviations from the theoretical distribution have already occurred for the second significant digit. In this study, however, households declare their income from the previous month as a single value.

This difference is so significant that the actual situation in these two periods was very different because of the financial crisis.

The most important differences between data sets concerned not the direction of change, but estimates of the absolute level of wealth. Given the differences in the definition of units for which data were collected in both sets, it can be assumed that the estimates are similar for less wealthy people (households). In the case of the highest incomes – above the quantile of order 0.99 – the differences are significant and they cannot be justified otherwise than by systematic errors in the HBS data set. Significant differences were also observed with respect to the estimates of mobility. Results for tax data show an approximately two times lower probability of leaving the group of wealthy individuals than analogous estimates for the HBS data.

These findings lead to ask the question about the reliability of the results obtained on the basis of data from both sources. Given the disadvantages of both data sets, characterized briefly in Section 4.1, the results indicate, however, the higher reliability of tax data in the assessment of the situation of top-income earners. Systematic errors, associated with the collection of the data on income, in the case of both sets indicate underestimation of the real values of income. So, if estimates of averages and quantiles, calculated for the tax data, are higher despite the smaller (or at least no bigger) unit (household may not consist of less than one person), the HBS data concerning the upper tail of the income distribution is certainly not reliable. Differences in the values of the parameters are very large for group of people with the highest incomes, so using data from the HBS to describe the situation of the richest is burdened with a very high risk to draw completely wrong conclusions – both in terms of assessing the level of wealth of this group, as well as changes in this level.

It is worth to note that despite the large discrepancy between the results obtained in analyzing results for both data sets, estimates of various parameters are relatively stable over time[15] for each data set separately. This means that the underestimation of the highest income has a systematic character and is not just a consequence of random errors in a particular edition of the survey.

The most controversial, in author's opinion, part of the analysis is evaluation of goodness-of-fit of empirical frequency distributions of significant digits for both data sets. Although the characteristics of both data sets would suggest that these distributions should be consistent with Benford's distribution, there is no conclusive empirical rule that could allow for formal and explicit testing of this hypothesis. But as expected, the empirical distributions proved to be similar to the theoretical distribution. Formulated by Hill (1995) postulate of sampling from the randomly drawn distributions can be – in the context of income – interpreted as recording income, earned from different sources (various employers). It justifies a very good fit of empirical distributions in the case

[15] For the full sample (covering households from all provinces in Poland) in case of the HBS data set.

of PIT36 and PIT37[16], and very large discrepancies for the income declared on form PIT40A – where all income is paid by the Social Insurance Institution.

However, the empirical frequencies fit the reference distribution consistently better for the tax data. In addition, the direction of changes in the empirical distribution for HBS data – the decrease in frequencies of digits 1 and 2 and the increase for digits 3-6 – may suggest systematic errors in the sample selection (not necessarily at the stage of setting the sampling scheme – such errors can be caused by the systematic refusal of certain groups to participate in the survey).

An important part of the analysis is the observation of goodness-of-fit of empirical frequency distributions and the theoretical distribution for the consecutive significant digits. To distinguish between data sets, analysis of the frequency distributions for the second and third significant digit proved to be sufficient. In case of HBS data, significant distortion from the theoretical (reference) distribution has been observed for the third digit. The immediate conclusion from this observation is that respondents round the income reported in the survey. This seemingly insignificant behavior, however, mean that the respondents do not record actual values of their income. And in this situation it is difficult to determine whether modifications are limited only to rounding or whether respondents significantly change the values.

6. Conclusions

Studies of affluence are limited by the data availability – even more than studies on inequality and other aspects of income distribution. In some situations, the relevant data are not attainable at all, but sometimes similar data can be derived from several independent sources, collected within the unrelated databases. In the latter case, assessment of the situation requires a decision, which set should be considered more reliable. Indications in the process of making such a decision may be, in addition to the results of basic analyses (based on the compatibility of the results of observations with theoretical considerations), some additional information regarding the structure of the data, the methods of data collection, etc.

In the case of an analysis of the situation of wealthy people, tax data (although burdened with certain defects) are considered more reliable than survey data (see, e.g., Moore et al., 2000). This assessment was essentially confirmed in this analysis. The basic conclusion that flows from the analysis is the limited usefulness of the data from household budget surveys (usually applied to the analysis of the income distribution in Poland) to assess the situation of the top-income earners. At the same time, it can be expected that for the part of population with lower incomes, survey data – covering a much wider range of income sources – will allow a more reliable assessment of the situation.

[16] The results for these forms are not presented in the paper. Upon request, they can be made available by the author.

Summing up, it should be emphasized that both data sets covered a relatively short period of time (only 5 years). Though the overall relationship between them appears to be stable over time, the period is far too short to formulate completely unambiguous conclusions.

References

Atkinson, A.B., and A. Brandolini (2001). "Promise and Pitfalls in the Use of 'Secondary' Data-Sets: Income Inequality in OECD Countries as a Case Study", Journal of Economic Literature, 39(3), 771-799

Atkinson, A.B., T. Piketty, and E. Saez (2011). "Top Incomes in the Long Run of History", Journal of Economic Literature, 49(1), 3–71.

Brzeziński, M. (2010). "Income Affluence in Poland", Social Indicators Research, 99, 285–299.

Burkhauser, R.V., F. Shuaizhang, S.P. Jenkins, and J. Larrimore (2009). "Recent Trends in Top Income Shares in the USA: Reconciling Estimates from March CPS and IRS Tax Return Data", Center for Economic Studies Working Paper, CES 09-26.

Central Statistical Office (2011a). "Household Budget Survey in 2010", Warszawa

Central Statistical Office (2011b). "Methodology of Household Budget Survey", Warszawa (in Polish)

Cho, K.T.M., and B.J. Gaines (2007). "Breaking the (Benford) Law Statistical Fraud Detection in Campaign Finance", The American Statistician, 61(3), 218-223.

Cleary, R., and J.C. Thibodeau (2005). "Applying Digital Analysis Using Benfor's Law to Detect Frauds", Auditing: A Journal of Practice & Theory, 24(1), 77-81.

Diekmann, A. (2007). "Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data", Journal of Applied Statistics, 34(3), 321-329.

Durtschi, C., W. Hillison, and C. Pacini (2004). "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data", Journal of Forensic Accounting, 5, 17-34.

Esteban, J., and D. Ray (2011). "Linking Conflict to Inequality and Polarization", American Economic Review, 101, 1345–74.

Ferrer-i-Carbonell, A. (2005). "Income and well-being: an empirical analysis of the comparison income effect", Journal of Public Economics, 89, 997– 1019.

Hill, T. (1995). "A Statistical Derivation of the Significant-Digit Law", Statistical Science, 10 (4), 354-363.

Human Development Reports (1990-2011), United Nations Development Programme, New York.

Kośny, M. (2011). "*Relative* income changes and an identification of growth pattern", ECINEQ Working Paper Series, 230.

Leigh, A. (2009). "Top incomes". In: W. Salverda, B. Nolan and T. Smeeding (eds.), "The Oxford handbook of economic inequality", Oxford: Oxford University Press.

Moore, J.C., L.L. Stinson, and E.J. Welniak (2000). "Income measurement error in surveys: A review", Journal of Official Statistics, 16, 331–361.

Ministry of Finance (2011), "Information Concerning the Personal Income Tax in 2010", Warszawa (in Polish).

Nigrini, M.J. (1996). "A Taxpayer Compliance Application of Beneford's Law", Journal of the American Taxation Association, 18, 72-91.

Osberg, L. (2009). "Measuring Economic Security in Insecure Times: New Perspectives, New Events, and the Index of Economic Well-Being", Centre for the Study of Living Standards Research Report 2009-12.

Piketty, T., and E. Saez (2003). "Income Inequality in the United States, 1913-1998", Quarterly Journal of Economics, 118(1), 1-39.

Saez, E., and M.R. Veall (2005). "The Evolution of High Incomes in Northern America: Lessons from Canadian Evidence", The American Economic Review, 95(3), 831-849.

Sierminska, E., A. Brandolini, and T.M. Smeeding (2006). "The Luxembourg Wealth Study – A cross-country comparable database for household wealth research", Journal of Economic Inequality, 4(3), 375‑383.

Wheary, J., T.M. Shapiro, and T. Draut (2007). "By a Thread: The New Experience of America's Middle Class", Demos: A Network for Ideas & Action.