IMPACT OF OUTLIERS ON INEQUALITY MEASURES – A COMPARISON BETWEEN POLISH VOIVODESHIPS

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 12(18)

Katarzyna Ostasiewicz
Wrocław University of Economics

ISSN 1644-6739

Summary: It is known that outlying (large) incomes strongly influence the results of inequality measuring. Thus, there is a question how to deal with such observations. In this paper the rule of excluding observations based on $(Q_1 - 1.5Q; Q_3 + 1.5Q)$ interval is investigated, for data from household budget survey in 2011, for Polish voivodeships. It is shown that although including more observations obviously changes the values of inequality measures, the relative values of them are surprisingly quite stable, with the rank correlation coefficient never over 0.9.

Keywords: inequality, outliers, Gini index, Theil index, Atkinson indexes.

DOI: 10.15611/sps.2014.12.06

1. Introduction

Outlying (very high) observations greatly influence most of commonly used inequality indexes, such as Gini or Theil indexes. The question of how to deal with such observations is twofold.

First, is it justified to include such observations even if we have data concerning the whole population and we are sure that they are not artificial? In experimental sciences, such as physics, outlying observations are usually regarded as errors and as such – omitted. On the other hand, it is true that in societies there exist individuals that are much richer than all the others and they are "real", thus, cannot be treated as "errors". There are two arguments that might be raised against including huge incomes into calculations of given society's economic inequality. First, our concern about inequalities arises mainly with respect to social cohesion, quality of life of its members, etc. It is highly doubtful that a single very rich person could influence mutual relations within the whole society. Also, many effects of inequality result from comparisons which individuals make between themselves and the others. Still, they compare themselves rather with theirs neighbours than with Bill Gates. The second argument against including extremely high incomes into inequality calculations is that the richest people are usually very mobile and the place of e.g. paying taxes is often dictated only by chance or by

convenience. Thus, including such a person in one region in one year, while in other region in subsequent year, could impose and smear out the overall view of social changes.

Second, if we deal with samples, there arise one more questions. Having in two regions exactly the same structure of incomes (and thus exactly the same inequality measures for the whole populations), and having in both the same number of the same extreme observations, we can still have a significant probability of obtaining – based on the samples – quite different estimations of inequality measures. Therefore, in the case of dealing with samples, there is no way of avoiding the question of extreme data leading to improper conclusions.

Thus, there are many reasons to question including extreme data into calculations of inequality measures [Van Kerm 2007; Schluter 2012]. Moreover, it has been shown that inequality measures, unlike other quantities, e.g. poverty measures, are extremely sensitive to such observations [Cowell, Victoria-Feser 1996; Cowell, Flachaire 2007]. There is no one standard approach to this problem, and a few different solutions are applied. Some authors exclude extreme values basing on criterion that there is a "gap" between these observations and all the others. Others omit all observations that are outside some range, determined by positional measures of sample's variation [Neri et al. 2009; Dudek 2013]. There are also some more sophisticated methods, e.g. excluding these values that can significantly change the value of inequality with precise methodology of addressing this problem [Cowell, Flachaire 2007; Davidson, Flachaire 2007; Hlasny, Verme 2013].

The aim of this paper is to investigate the status of one of the most often used solution, that is, excluding observation that are outside the range of $(Q_1 - 1.5Q; Q_3 + 1.5Q)$. It is usually observed that such an approach typically excludes at most a few per cent of observations, which seems reasonable. On the other hand, the value 1.5 appearing in this form of interval is quite arbitral. The issue investigated here is as follows: how the slight changes of value 1.5 influence the results? As it seems obvious that unless this value coincide with the "gap" in ordered observations, decreasing 1.5 will decrease all possible measures of inequality while increasing it – increase these measures. That is, our concern here will be whether such changes may reverse relative values for different regions. As in most cases it is not the absolute value of inequality measures that matters but rather their relative differences – between various regions or in various moments – that seem reasonable

to investigate the impact of the excluding rule on the order of inequality measures. The data that will be used here to examine and illustrate the problem comes from household budget survey, carried out in 2011.

The paper is organized as follows. In the next section the inequality measures examined here are defined and methodology of the investigations is described in more details. The following section investigates relationship between examined rule of excluding observations and approach based on appearing of the "gap" within ordered observations. In section 4 we present analysis of order of Polish voivodeships (ordered according to the value of examined inequality measures) as dependent on the range of included observations. Final section gives a summary of the considerations.

2. Definitions and methods

In this section the measures of inequality that will be considered are described, as well as the method to be applied to investigate the applicability of the discussed rule of excluding outlying observations.

There are investigated here the values of the following inequality measures.

- Gini index which is the most frequently used in both scientific and policy-oriented elaborations. It is calculated as:

$$G = \frac{1}{2n^2\bar{x}}\sum_{i,j=1}^{n}\left|x_i - x_j\right|, \tag{1}$$

where x_i denotes individual observations, n – the number of them, and $\bar{x}$ – the average of x_i.

- Theil index, calculated by:

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{\bar{x}}\ln\frac{x_i}{\bar{x}}, \tag{2}$$

- Atkinson indexes which are a family of inequality measures differing in the value of one parameter, ε. The form of this index is as follows:

$$A_\varepsilon = 1T\frac{1}{\bar{x}}\left[\frac{1}{n}\sum_{i=1}^{n}x_i^{1-\varepsilon}\right]^{1/(1-\varepsilon)}, \tag{3}$$

with $\varepsilon > 0$. Parameter ε is called an "inequality aversion" parameter and it is known that the greater is the value of this parameter, the more sensitive is the Atkinson index to the inequalities at the bottom of income distribution.

The often used rule of excluding outlying observation, examined in this paper, is based on omitting those observations which lie outside the range defined as:

$$(Q_1 - 1.5Q; Q_3 + 1.5Q), \quad (4)$$

where Q_1 denotes the first quartile, Q_3 – the third quartile, while Q – interquartile range, that is: $Q = Q_3 - Q_1$.

It is usually noted that outside this range there are only at most a few per cents of observations.

The method of investigating the validity of this rule of excluding outlying observations while calculating inequality measures will be as follows. The value 1.5 in (4) is chosen more or less arbitrarily. Thus, let us consider more generalized case of excluding observations lying outside the range defined as:

$$(Q_1 - qQ; Q_3 + qQ), \quad (5)$$

with $q \geq 0$. We are especially interested in values of q close to 1.5 and we investigate sensitivity of the results while slightly changing this value.

The method of excluding observations based on (4) is also compared with another popular one, that is, excluding observations which follows some "gap" in the ascendant ordered observations. We expect all

Table 1. Abbreviations used in the text for the Polish voivodeships

Voivodeship	Abbreviation
Dolnośląskie	02
Kujawsko-pomorskie	04
Lubelskie	06
Lubuskie	08
Łódzkie	10
Małopolskie	12
Mazowieckie	14
Opolskie	16
Podkarpackie	18
Podlaskie	20
Pomorskie	22
Śląskie	24
Świętokrzyskie	26
Warmińsko-mazurskie	28
Wielkopolskie	30
Zachodniopomorskie	32

Source: own elaboration based on household budget survey for 2011.

inequality measures to increase with including more and more observations less and less close to the median value, that is – with increasing q in (5), which fixes the interval of included observations. The only possibility not to increase inequality measures would be that increasing q did not impose including more distant observations – that would be the case if there were a "gap" in observations coinciding with the value of q in vicinity of which we are investigating inequality measures. In the next section these two rules of excluding observations are compared.

The data examined here are results of household budget survey for 2011, for 16 Polish voivodeships. In what follows the standard symbols for voivodeships are applied, that is, 02 for Dolnośląskie, and so on, until 16 for Zachodniopomorskie, see Table 1.

The OECD equivalence scale for net income has been applied ("Oxford scale"), within which the first adult is assigned with the weight 1, each following adult with 0.7, and a child up to 14 years with the weight 0.5 [OECD 1982]. Inequality measures are calculated on individuals' level (not for households) (see e.g. [Jenkins 1991]).

3. "Gaps" in ordered observations vs. distance from the median value

In this section the values of inequality measures versus value of q (Eq. (5)) are presented. Also, appearing of the first "gap" in the data is investigated. As income is in fact a discrete quantity, some concrete definition of the "gap" has to be adopted. Our main concern here are intervals of the form (5), thus, the "gap" will be defined for the purpose of this paper in strict relation to it. That is, we will mark the appearance of the "gap" if for an interval of some width, expressed in terms of interquartile range, Q, there will be no observations. Four cases are investigated here: widths equal to $0.01Q$, $0.02Q$, $0.03Q$ and $0.05Q$.

Figures 1, 2 and 3 below present exemplary plots of inequality measures vs. q for Dolnośląskie voivodeship. Solid lines visualize values of given inequality measures (values on the left axes), while dotted lines – per cent of observations that correspond to the given q (values on the right axes). The values for $q = 0$ correspond to inequality indexes calculated for observation from the interval $(Q_1; Q_3)$. The per cent of observations included within such an interval is always 0.5, that is, the dashed lines start always from 0.5. With increasing q we include more and more observations into calculations, approaching the maximum values of inequality measures (and 100% of observations)

for q large enough. This value of q for which the maximum values of inequality measures are reached may be of course different for different voivodeships.

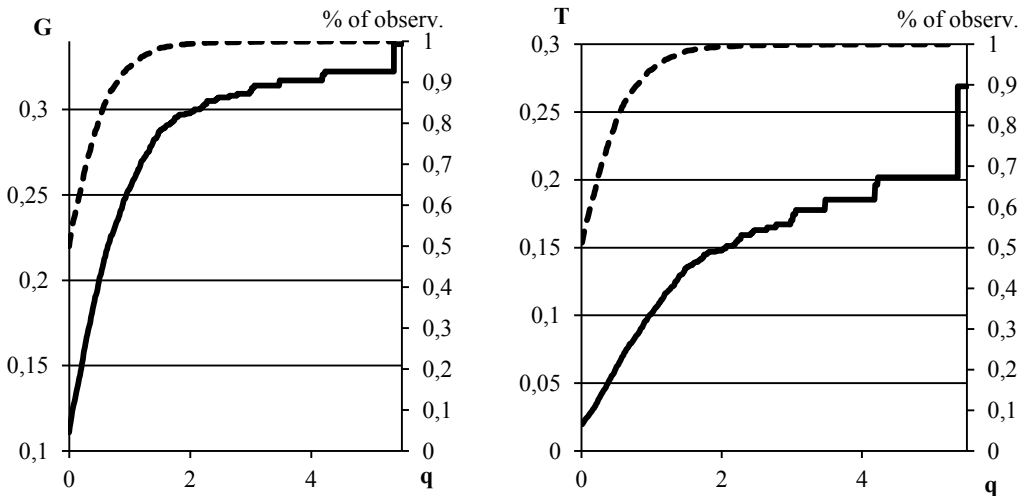

Figure 1. Gini and Theil indexes versus the range of included observations for Dolnośląskie voivodeship

Source: own calculations.

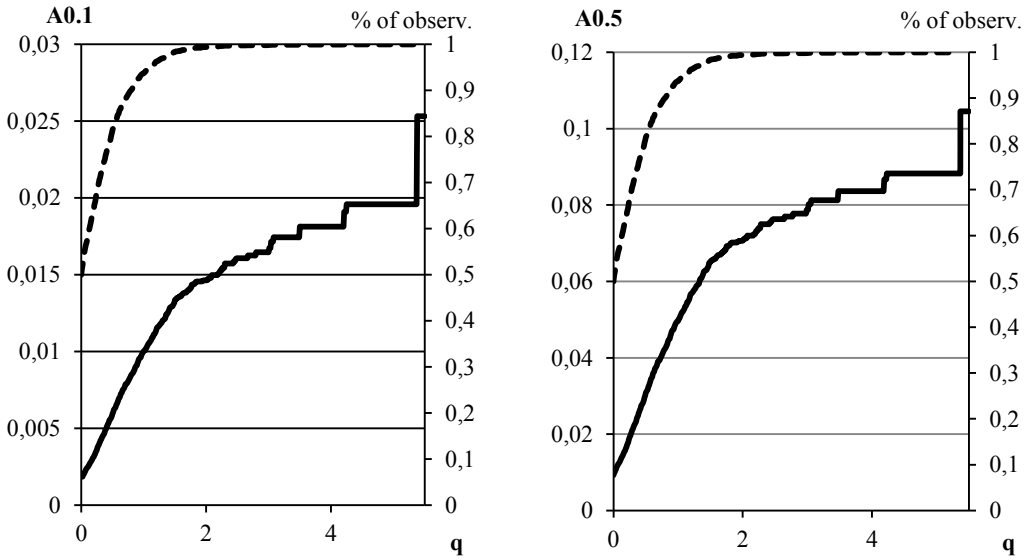

Figure 2. Atkinson indexes ($\varepsilon = 0.1$, $\varepsilon = 0.5$) versus the range of included observations for Dolnośląskie voivodeship

Source: own calculations.

At the very first glance one important feature of each of these plots is striking: after the range of quite smooth increase of inequality measures there appear some "steps" which correspond to the gaps in ordered observations. The observations appearing after these gaps have an overwhelming impact on the values of inequality measures. This is most visible in the case of Theil index: four observations (in fact, it is one household consisting of 4 persons, what may be checked within detailed data) cause the increase of this measure from 0.2 to 0.27, that is, 35% increase. As might be expected, the higher is the value of ε for Atkinson indexes, the less sensitive is the measure for extremely high observations. However, even for $\varepsilon = 1.5$ still a noticeable increase of Atkinson index may be observed, caused by outlying observations.

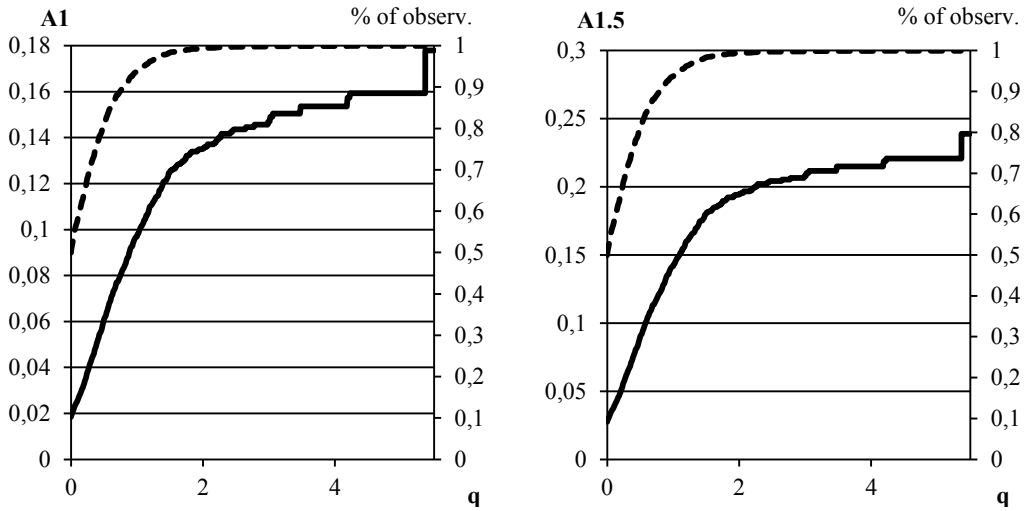

Figure 3. Atkinson indexes ($\varepsilon = 0.999$, $\varepsilon = 1.5$) versus the range of included observations for Dolnośląskie voivodeship

Source: own calculations.

It seems obvious that observations after such a wide empty range (in the vicinity of $q = 4$) may be called outlying ones. However, this wide empty range is proceeded by a few shorter empty ranges, some of them also quite visible on the plots. The question is, whether first appearance of the empty range (also called here a "gap") coincides with some specific value of q (presumably $q = 1.5$) in cases of all voivodeships?

Figure 4 illustrates the values of q for which the first appearance of "gap" may be observed. "Gaps" are defined, as mentioned above,

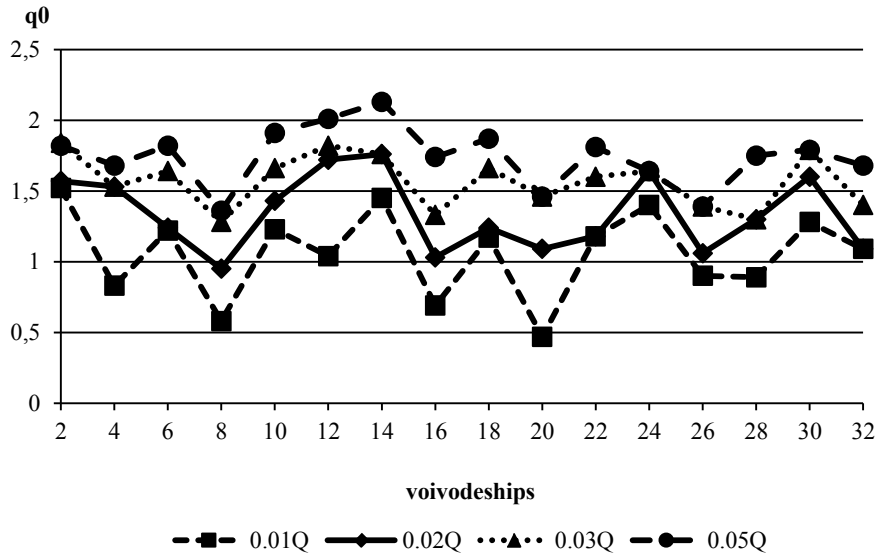

Figure 4. First appearance of "gaps" of different widths for all voivodeships

Source: own calculations.

fourfold: no observations within the range of the width $0.01Q$, $0.02Q$, $0.03Q$ and $0.05Q$.

It may be observed that there is no consistency in the value of q for which the first appearance of the gap may be observed, for either of the three definitions of the gap. The least variability (slightly less than 12%) is for the gap defined as an empty interval of the width $0.05Q$, and the average first occurrence of such defined gap is for $q = 1.74$. However, for $0.03Q$ variability is almost the same, slightly more than 12%, while $q = 1.57$, which is close to the frequently used value of 1.5.

This observation may serve as a (probably weak) justification of the choice of the value 1.5 as a default value for all cases.

4. Comparison of inequality measures among voivodeships while changing the range of included observations

As it is not surprising that including more observations into account – that is, increasing q – increases the values of all inequality measures, there arises a question of their relative increase for different regions.

Figure 5 presents Gini index for all 16 voivodeships versus value of q, in the range $q \in \langle 0; 2.5 \rangle$. It may be noticed that although some curves preserve their relative position in the whole range of q, there are also

some intersections. As it is not clearly visible while plotting absolute values of Gini index, Figure 6 presents the ranks of values of Gini index for all voivodeships, plotted against *q* within the same range as in Figure

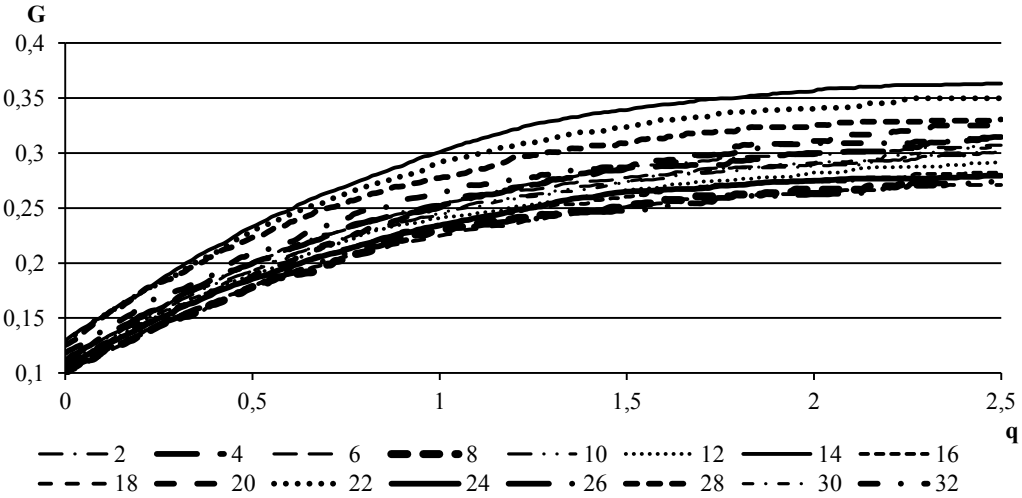

Figure 5. Gini index vs. the range of included observations for Polish voivodeships

Source: own calculations.

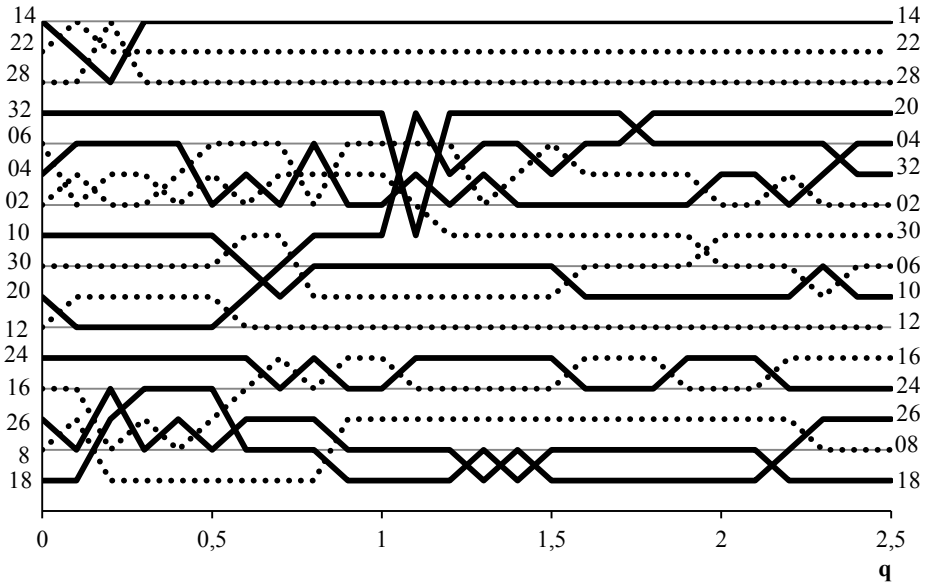

Figure 6. Ranks of Gini index values for Polish voivodeships vs. the range of included observations

Source: own calculations.

Table 2. Values of rank correlation coefficient for Polish voivodeships' orderings produced by adopting different values of cutoff $(Q_1 - qQ; Q_3 + qQ)$

q	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5
0	1.000	0.982	0.968	0.976	0.968	0.971	0.982	0.976	0.985	0.976	0.976	0.906	0.950	0.938	0.941	0.944	0.929	0.929	0.912	0.915	0.906	0.906	0.906	0.903	0.984	0.984
0.1		1.000	0.965	0.974	0.971	0.968	0.979	0.962	0.968	0.968	0.968	0.894	0.947	0.932	0.932	0.944	0.924	0.924	0.903	0.906	0.903	0.903	0.900	0.897	0.983	0.983
0.2			1.000	0.979	0.982	0.968	0.965	0.944	0.962	0.926	0.926	0.856	0.906	0.897	0.894	0.903	0.879	0.879	0.859	0.868	0.865	0.865	0.853	0.862	0.977	0.977
0.3				1.000	0.994	0.988	0.976	0.959	0.974	0.944	0.944	0.871	0.921	0.921	0.909	0.926	0.906	0.906	0.885	0.891	0.888	0.888	0.871	0.874	0.979	0.979
0.4					1.000	0.988	0.979	0.956	0.968	0.938	0.938	0.871	0.924	0.915	0.909	0.926	0.900	0.900	0.879	0.888	0.885	0.885	0.871	0.876	0.979	0.979
0.5						1.000	0.982	0.971	0.962	0.956	0.956	0.876	0.932	0.915	0.915	0.938	0.912	0.912	0.891	0.897	0.882	0.882	0.876	0.868	0.977	0.977
0.6							1.000	0.991	0.979	0.974	0.974	0.909	0.959	0.938	0.944	0.956	0.941	0.941	0.924	0.926	0.921	0.921	0.924	0.921	0.987	0.987
0.7								1.000	0.976	0.982	0.982	0.918	0.962	0.941	0.950	0.959	0.956	0.956	0.941	0.938	0.926	0.926	0.938	0.926	0.988	0.988
0.8									1.000	0.976	0.976	0.935	0.968	0.968	0.965	0.965	0.956	0.956	0.944	0.947	0.941	0.941	0.935	0.938	0.991	0.991
0.9										1.000	1.000	0.947	0.985	0.968	0.976	0.982	0.974	0.974	0.962	0.959	0.941	0.941	0.953	0.935	0.989	0.989
1											1.000	0.947	0.985	0.968	0.976	0.982	0.974	0.974	0.962	0.959	0.941	0.941	0.953	0.935	0.989	0.989
1.1												1.000	0.968	0.965	0.971	0.965	0.962	0.962	0.974	0.976	0.965	0.965	0.968	0.956	0.995	0.995
1.2													1.000	0.988	0.997	0.997	0.988	0.988	0.982	0.985	0.974	0.974	0.979	0.968	0.994	0.994
1.3														1.000	0.994	0.991	0.991	0.991	0.988	0.991	0.988	0.988	0.979	0.974	0.995	0.995
1.4															1.000	0.994	0.991	0.991	0.988	0.991	0.982	0.982	0.985	0.976	0.995	0.995
1.5																1.000	0.991	0.991	0.985	0.988	0.976	0.976	0.976	0.962	0.993	0.993
1.6																	1.000	1.000	0.997	0.994	0.988	0.988	0.991	0.976	0.996	0.996
1.7																		1.000	0.997	0.994	0.988	0.988	0.991	0.976	0.996	0.996
1.8																			1.000	0.997	0.991	0.991	0.994	0.979	0.997	0.997
1.9																				1.000	0.994	0.994	0.991	0.976	0.996	0.996
2																					1.000	1.000	0.991	0.985	0.998	0.998
2.1																						1.000	0.991	0.985	0.998	0.998
2.2																							1.000	0.991	0.998	0.998
2.3																								1.000	0.999	0.999
2.4																									1.000	1.000
2.5																										1.000

Source: own calculations.

5, $q \in \langle 0; 2.5 \rangle$, with the step Δq – 0.1 (the lowest rank corresponds to the lowest value of inequality index). Here, each change of relative ranks is clearly visible. (Every second line is dotted, just for the sake of clarity.)

It may be observed that some relative ranks (of some pairs of voivodeships) are quite stable, e.g. 10 is in the whole range more unequal than 12. On the other hand, there are also many inversions of ranks. Although it might be expected that for larger values of q, presumably $q \geq 1.5$, there would appear more sudden changes (due to outlying observations), there is only one such a case, for the voivodeship 08. The vicinity of $q = 1.5$ seems not to reveal any special features. Thus, comparing levels of inequality among Polish voivodeships one has to be conscious, how the choice of particular cutoff, q, may influence the final results. Table 2 below shows the values of rank correlation coefficient between pairs of values of q. Despite some changes in order seen in Figure 6, rank correlation coefficients are in most of cases greater than 0.9: of course, they are the greatest near the diagonal (which means small differences in the cutoff value q), decreasing in the far from diagonal "corner" of the table (which corresponds to the quite different ranges of included observations). We can see that the choice of cutoff $q = 1.4$ (instead of 1.5) would lead to the order which rank correlation with the order produced by $q = 1.5$ would be equal to 0.994, while order for $q = 1.6$ and order for $q = 1.5$ would give rank correlation coefficient equal to 0.991.

Figures 7–11 present analogous to Figure 6 orderings of Polish voivodeships for Theil index and four Atkinson indexes, for different values of cutoffs, ranging from $q = 0$ to $q = 2.5$, with the step $\Delta q = 0.1$.

Comparing Figures 6–11 one may detect significant similarities. First, the orders produced by different indexes are quite alike, but that is another question concerning similarities/differences between inequality measures, which we do not address here. Second, the intersections between lines appear in all plots in more or less the same number, what leads to similar values of rank correlation coefficients – they are not given here for indexes other than Gini index, however, their values are quite similar values as those in Table 2. Moreover, it can be noticed that most of voivodeships preserve in principle their ranks, oscillating about some given value up to two ranks. The only voivodeship that changes its relative position with respect to other voivodeships is Podlaskie (number 20). One should probably take this fact into account while performing some comparison analysis.

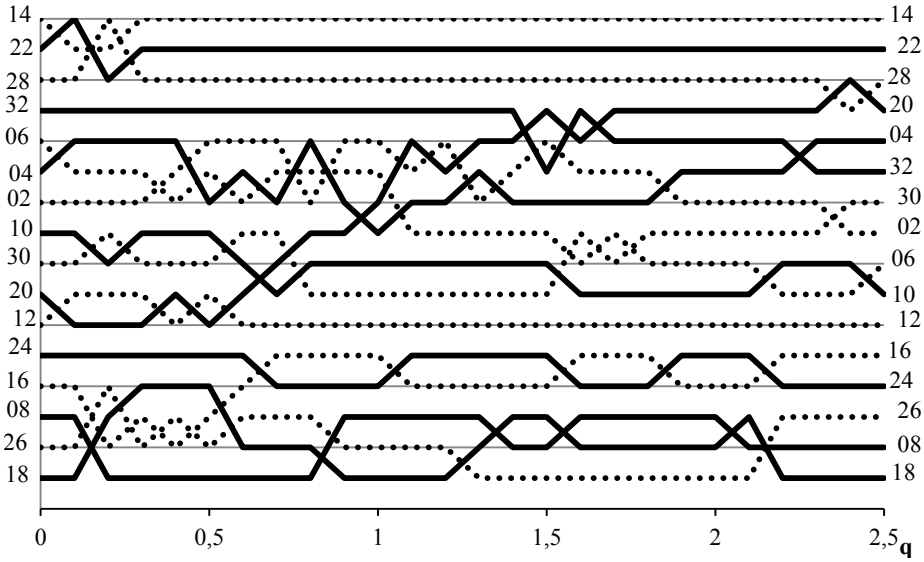

Figure 7. Ranks of Theil index values for Polish voivodeships vs. the range of included observations

Source: own calculations.

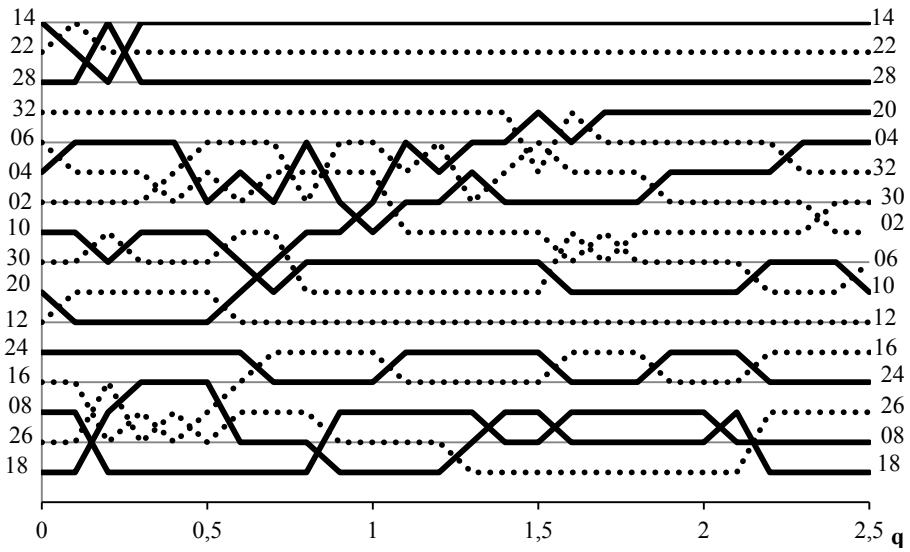

Figure 8. Ranks of Atkinson index ($\varepsilon = 0.1$) values for Polish voivodeships vs. the range of included observations

Source: own calculations.

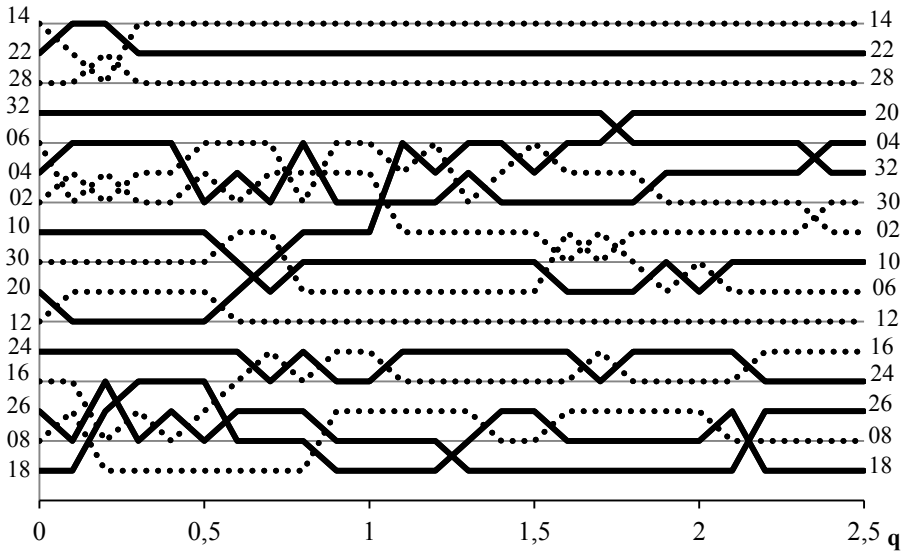

Figure 9. Ranks of Atkinson index ($\varepsilon = 0.5$) values for Polish voivodeships vs. the range of included observations

Source: own calculations.

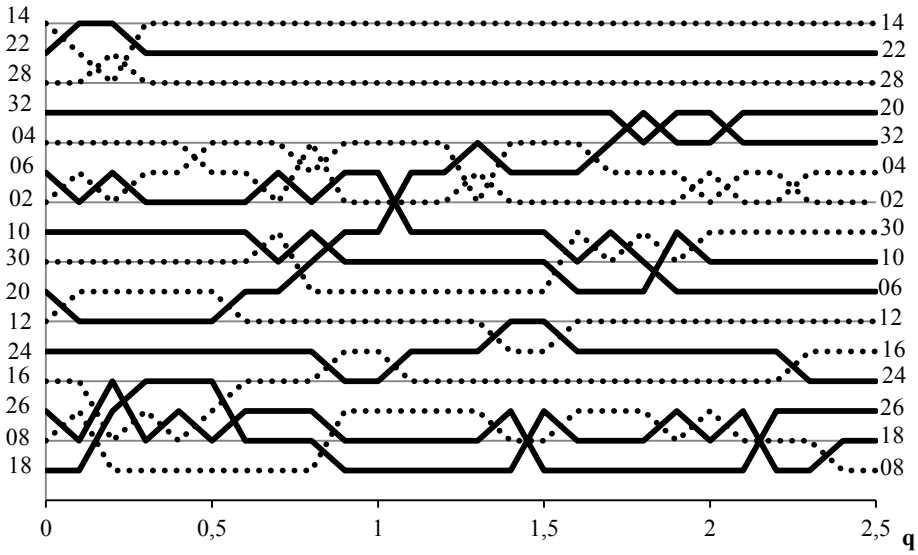

Figure 10. Ranks of Atkinson index ($\varepsilon = 0.999$) values for Polish voivodeships vs. the range of included observations

Source: own calculations.

As for the value $q = 1.5$ we cannot observe any special behaviour of plots in vicinity of this value. Thus, at least with respect to this issue, there are no reasons to choose $q = 1.5$ rather than $q = 1$ or $q = 2$.

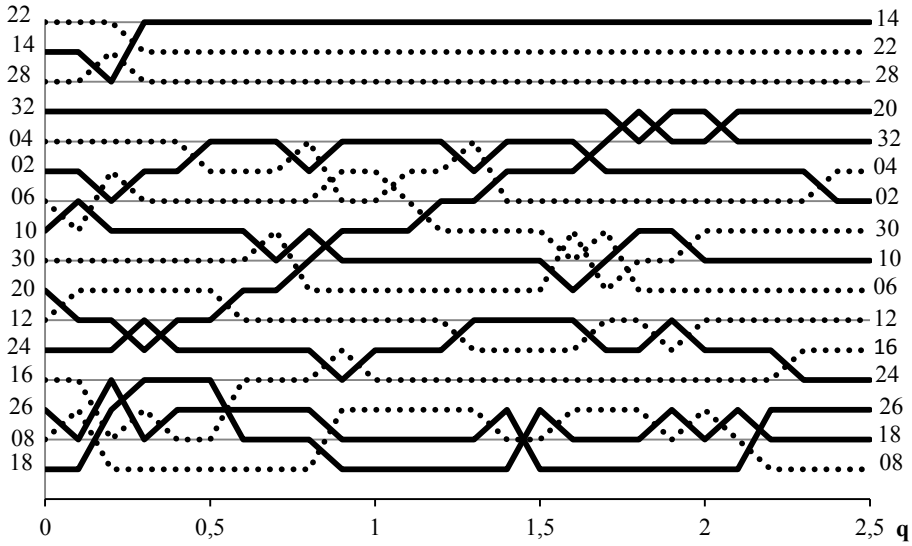

Figure 11. Ranks of Atkinson index ($\varepsilon = 1.5$) values for Polish voivodeships vs. the range of included observations

Source: own calculations.

Even extending analysis to q ranging from 0 to 7 does not change the picture significantly. The rank correlation coefficient between orders produced by different q's within this extended range are never lower than 0.9. 0.9 may seem not so much, but taking into regard that, for example, rank correlation coefficient between orders of Gini indexes given for the same set of countries by UN and EU sources is equal to approximately 0.7 for some years, it seems that outlying observations are not the main source of problems while attempting to order properly different regions.

5. Summary

We have investigated here a certain rule of excluding outlying observations, while calculating inequality measures. The examined rule consists of excluding those incomes which natural logarithms lie outside

the interval $(Q_1 - 1.5Q; Q_3 + 1.5Q)$ of incomes' logarithms. Our purpose was to investigate whether the choice of particular value, 1.5, is in any way special or significant. To this aim we have examined the values of a few inequality indexes while taking into account incomes from the range of $(Q_1 - qQ; Q_3 + qQ)$, where q was varying from 0 to 2.5. We saw that changing value of q changes the order of different voivodeships only slightly, with the only exception of Podlaskie voivodeship. On the other hand, as for not relative but absolute values of inequality measures, these values obviously change with changing q, and it is not clear which value to choose to avoid both underestimating inequality by excluding too many observations and overestimating by including nonrepresentative ones. It can be noticed, however, that value close to 1.5 corresponds to appearing of the first "gap" in ordered logarithms of observations, with "gap" defined as an empty interval of the width of $0.03Q$. Such a definition of a gap is also an arbitrary one, but it may be observed that, at least for the data we deal here with, it corresponds to the relatively low variance of the position of the first appearing.

References

Cowell F.A., Flachaire E., *Income distribution and inequality measurement: The problem of extreme values*, "Journal of Econometrics" 2007, Vol. 141, pp. 1044–1072.

Cowell F.A., Victoria-Feser M.-P., *Robustness properties of inequality measures*, "Econometrica: Journal of the Econometric Society" 1996, pp. 77–101.

Davidson R., Flachaire E., *Asymptotic and bootstrap inference for inequality and poverty measures*, "Journal of Econometrics" 2007, Vol. 141, pp. 141–166.

Dudek H., *Equivalence scales for Poland – new evidence using complete demand systems approach*, Research Papers of Wrocław University of Economics No. 308, Wrocław 2013, pp. 128–143.

Hlasny V., Verme P., *Top incomes and the measurement of inequality in Egypt*, ECINEQ Working Paper Series No. 303, 2013.

Jenkins S., *The measurement of income inequality*, "Economic Inequality and Poverty: International Perspectives" 1991, pp. 3–38.

Neri L., Gagliardi F., Ciampalini G., Verma V., Betti G., *Outliers at upper end of income distribution (EU-SILC 2007)*, DMQ Working Paper No. 86, November 2009.

OECD, *The OECD List of Social Indicators*, Paris, 1982.

Schluter C., *On the problem of inference for inequality measures for heavy–tailed distributions*, "The Econometrics Journal" 2012, Vol. 15, No. 1, pp. 125–153.

Van Kerm P., *Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC*, IRISS Working Paper Series 2007-01, CEPS/INSTEAD, 2007.

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY

Nr 12 (18)

WPŁYW OBSERWACJI ODSTAJĄCYCH NA MIARY NIERÓWNOŚCI – PORÓWNANIE POMIĘDZY POLSKIMI WOJEWÓDZTWAMI

Streszczenie: Wiadomo, że odstające (o dużych wartościach) dochody silnie wpływają na pomiary nierówności. Powstaje zatem kwestia, jak postępować z takimi obserwacjami. W niniejszej pracy dyskutowana jest reguła wykluczania obserwacji odstających oparta na przedziale $(Q_1 - 1.5Q; Q_3 + 1.5Q)$, w odniesieniu do danych z Badań Budżetów Gospodarstw Domowych za rok 2011 dla wszystkich polskich województw. Pokazane jest, iż – choć włączanie coraz większej liczby coraz bardziej odstających obserwacji w sposób oczywisty zwiększa miary nierówności – porządek wartości tych miar (przy porównywaniu różnych województw) jest zaskakująco stabilny, ze współczynnikiem korelacji rang wynoszącym zawsze co najmniej 0,9.

Słowa kluczowe: nierówności, współczynnik Giniego, współczynnik Theila, współczynniki Atkinsona, obserwacje odstające.