MODELLING INCOME DISTRIBUTION: THE CASE OF POLAND

Anna Sączewska-Piotrowska

University of Economics in Katowice
Faculty of Economics, Department of Labour Market Forecasting and Analysis
ul. 1 Maja 50, 40-287 Katowice, Poland
E-mail: anna.saczewska-piotrowska@ue.katowice.pl

Abstract: *Income distribution can be examined using different methods. One of the method is fitting theoretical distributions to empirical income data. The aim of the paper is to fit the best theoretical models to income data in Poland. The selected models are taken into account: Pareto, Lomax, lognormal, log-logistic, Singh-Maddala and Dagum. Probability density function and cumulative density functions of these selected models are presented. Empirical and theoretical basic income measures (mean and median) as well as inequality measures (Gini coefficient, inter-decile ratio, Palma ratio) are presented and compared. The analysis shows that mean and median income were systematically rising in observation period (the real income is analysed). The values of inequality measures were rising to 2005 and in subsequent years the values were decreasing. In 2000-2015 income distributions are mainly follow Singh-Maddala and log-logistic distribution. Income distribution were changed between 2000 and 2015: there are the visible changes in graphical presentation of empirical income data and in worse fitting theoretical distributions to data.*

Key words: *parametric estimation, income distribution, Gini coefficient, Palma ratio*

JEL codes: *C13, D31, D63*

1. Introduction

Income distribution can be described in different ways. One of the used methods is to fit theoretical distribution approximating the income data. The first theoretical model was proposed by Pareto (1896). However, Pareto model gives the poor fit to empirical data. In the subsequent years there were fitted another types of theoretical models, e.g. lognormal model (Champernowne, 1952; Bordley et al., 1996), Burr type III also called Dagum model (Dagum, 1977; McDonald and Xu, 1995; Majumder and Chakravarty, 1990), Burr type XII also called Singh-Maddala model (Singh and Maddala, 1976; McDonald and Xu, 1995). Before 1989 the good approximation in Poland gave the log-normal model (e.g. Vielrose, 1960), but this situation changed after transformation of Polish economy. In some previous studies basing on data after transformation (Kot, 2000; Ostasiewicz, 2013; Salamaga, 2016) log-normal model gave worse fit than Burr type III or Burr type XII models.

In recent years the authors still deal with the problem of fitting theoretical distribution to income data. For example, Huang and Oluyede (2014) proposed a new family including several known sub-models, such as Dagum and Fisk, and new sub-models, such as Kumaraswamy-Dagum and exponentiated Kumaraswamy-Fisk.

The aim of the paper is to fit the best theoretical models to income data in Poland. The selected models are taken into account: Pareto, Lomax, lognormal, log-logistic, Singh-Maddala and Dagum. Probability density functions and cumulative density functions of these selected models are presented. Empirical and theoretical basic income measures (mean and median) as well as inequality measures (Gini coefficient, inter-decile ratio, Palma ratio) are presented and compared.

2. Methodology and Data

The analysis of income distribution was based on the data from Social Diagnosis project (Council for Social Monitoring, 2015). Generally, Social Diagnosis project is based on panel research. The first sample was taken in 2000. The next sample took place three years later and since then measurement has been repeated every two years (eight waves in 2000-2015). The household was the study unit. Table 1 contains information on the number of households surveyed in subsequent waves of panel.

Tab. 1 Number on households in database of Social Diagnosis project

Year	2000	2003	2005	2007	2009	2011	2013	2015
Wave	I	II	III	IV	V	VI	VII	VIII
Number of households	3005	3962	3881	5532	12380	12359	12343	11738

Source: Own calculations based on data from Council for Social Monitoring (2015)

The basic variable is net income per household in Poland in March/June in subsequent waves of panel. In order to take account the differences in a household's size and its composition an equivalised income was calculated by dividing the household's income by its equivalent size. There was used the modified OECD (Organization for Economic Co-operation and Development) equivalence scale. This scale assigns 1 to the first adult of the household, 0.5 to each subsequent adult aged 14 or more and 0.3 to children (each person under 14). The D9/D1 ratio is the ratio of the upper bound value of the ninth decile to the upper bound value of the first decile (OECD, 2017). This measure ranges from 1 to infinity. The higher values of the D9/D1 ratio, the higher income inequality.

The most popular measure of income inequality is Gini coefficient defined as the relationship of cumulative shares of the population arranged according to the level of equivalised disposable income, to the cumulative share of the equivalised total disposable income received by them. In alternative approach Gini coefficient is defined as half of the relative mean absolute difference which can be expressed by the formula (Sen, 1997):

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

where x_i is income of household i and there are n households, μ is the mean income. The Gini coefficient ranges between 0 (perfect equality) to 1 (perfect inequality). It is popular to express the Gini coefficient in percentages.

The Palma ratio of inequality was proposed by Alex Cobham and Andy Sumner, on the basis of the Palma proposition: an observation by Jose Gabriel Palma that currently changes in income or consumption inequality are almost exclusively due to changes in the share of the richest 10 per cent and poorest 40 per cent, because the 'middle' group between the richest and poorest always capture approximately 50 per cent of gross national income (Cobham and Sumner, 2013).

The analysis of empirical income distribution often consists of fitting the theoretical models. In this paper there were fitted known models to the datasets, by maximum likelihood, from 2000 to 2015. The distributions are fitted in two versions. For example, two-parameter (2P) and three-parameter (3P) lognormal model were fitted. Some of the considered models have two parameters: Pareto type I, Lomax (Pareto type II with location parameter $\mu = 0$), lognormal (2P) and log-logistic (or Fisk) (2P) distributions. The other four models have three parameters: lognormal (3P), log-logistic (3P), Singh-Maddala (or Burr type XII or Burr) (3P) and Dagum (or Burr type III or inverse Burr) distributions. The remaining models have four parameters: Singh-Maddala (4P) and Dagum (4P). Table 2 shows the probability density functions (PDF) and cumulative distribution functions (CDF) of the models considered.

Based on the best fitted distributions the theoretical basic characteristics and inequality measures were calculated. Formulas are presented in literature, e.g. Kleiber and Kotz (2003).

To assess whether the data follow the assumed distribution a graphical tool (quantile-quantile plot) there was used. Quantile-quantile plot shows quantiles of the empirical data vs. the theoretical quantiles. The plots with approximately straight lines suggest that the income distribution follow a given distribution.

The goodness of fit was tested by the Anderson-Darling test. The null hypothesis is: the data follow the specified distribution. The test hypothesis is rejected if the Anderson-Darling statistic is greater than a critical value (e.g. value of 2.5018 at $\alpha = 0.05$). Based on results of Anderson-Darling test there were made rankings of income distributions (for each year) and the best fits were chosen.

All calculations and plots are made in EasyFit and R software.

Tab. 2 Probability density functions and cumulative distribution functions

Distribution	PDF	CDF
Pareto	$\frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}$	$1-\left(\frac{\beta}{x}\right)^{\alpha}$
Lomax	$\frac{\alpha\beta^{\alpha}}{(x+\beta)^{\alpha+1}}$	$1-\left(\frac{\beta}{x+\beta}\right)^{\alpha}$
Lognormal (2P)	$\frac{\exp\left[-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^{2}\right]}{x\sigma\sqrt{2\pi}}$	$\Phi\left(\frac{\ln x-\mu}{\sigma}\right)$
Lognormal (3P)	$\frac{\exp\left[-\frac{1}{2}\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)^{2}\right]}{(x-\gamma)\sigma\sqrt{2\pi}}$	$\Phi\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)$
Log-logistic (2P)	$\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\left[1+\left(\frac{x}{\beta}\right)^{\alpha}\right]^{-2}$	$\left[1+\left(\frac{\beta}{x}\right)^{\alpha}\right]^{-1}$
Log-logistic (3P)	$\frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left[1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]^{-2}$	$\left[1+\left(\frac{\beta}{x-\gamma}\right)^{\alpha}\right]^{-1}$
Singh-Maddala (3P)	$\frac{\alpha k\left(\frac{x}{\beta}\right)^{\alpha-1}}{\beta\left[1+\left(\frac{x}{\beta}\right)^{\alpha}\right]^{k+1}}$	$1-\left[1+\left(\frac{x}{\beta}\right)^{\alpha}\right]^{-k}$
Singh-Maddala (4P)	$\frac{\alpha k\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\beta\left[1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]^{k+1}}$	$1-\left[1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]^{-k}$
Dagum (3P)	$\frac{\alpha k\left(\frac{x}{\beta}\right)^{\alpha k-1}}{\beta\left[1+\left(\frac{x}{\beta}\right)^{\alpha}\right]^{k+1}}$	$\left[1+\left(\frac{x}{\beta}\right)^{-\alpha}\right]^{-k}$
Dagum (4P)	$\frac{\alpha k\left(\frac{x-\gamma}{\beta}\right)^{\alpha k-1}}{\beta\left[1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]^{k+1}}$	$\left[1+\left(\frac{x-\gamma}{\beta}\right)^{-\alpha}\right]^{-k}$

Source: Based on Fisk (1961), Singh-Maddala (1976), Dagum (1977), Kleiber and Kotz (2003)

3. Results and Discussion

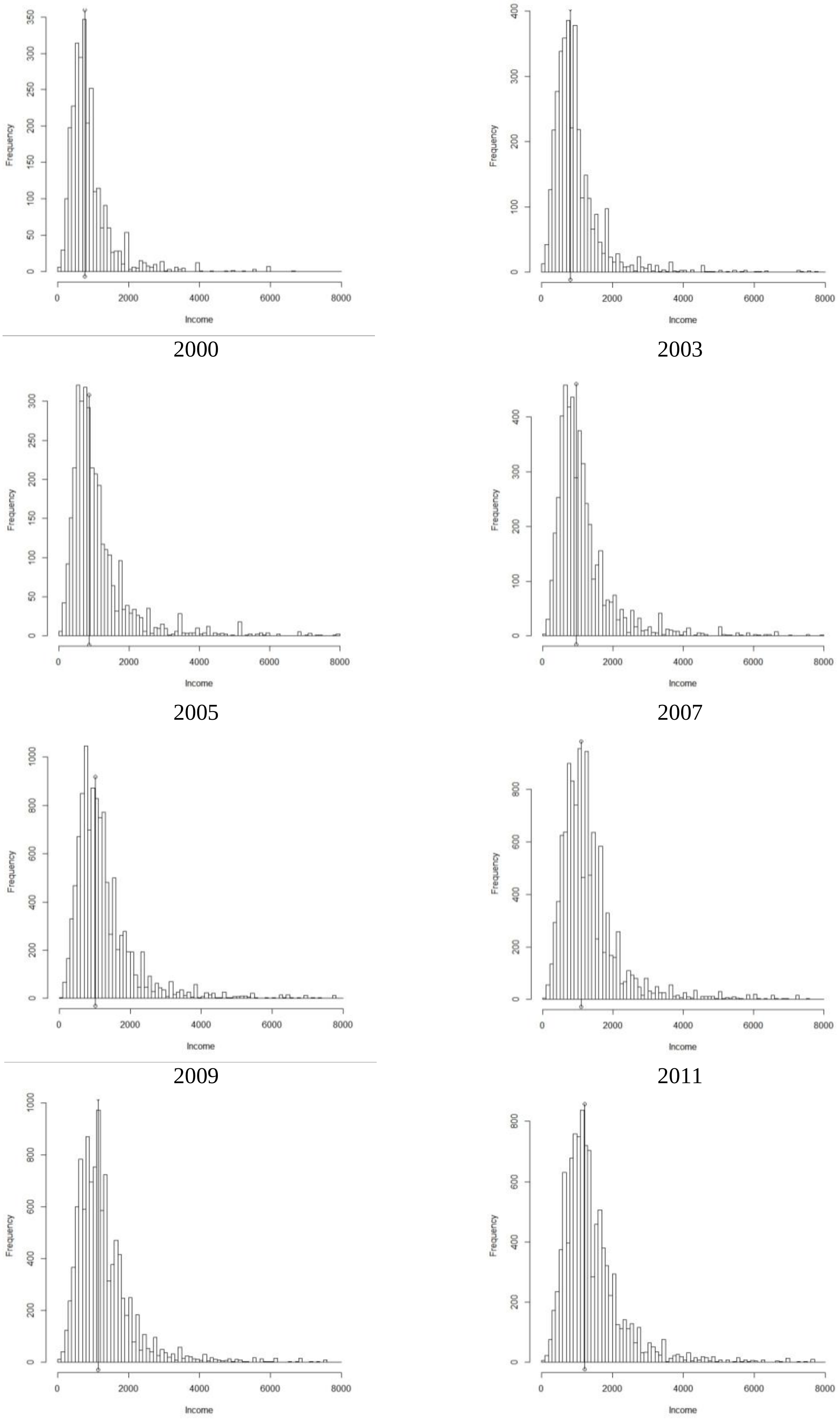
The basic characteristics of income are presented in table 3. From 2000 to 2015 mean and median income were increasing. All of inequality measures reached the highest values in 2005. It should be noted that in 2015 the households are characterized by the best income situation – the highest mean and median of income and simultaneously the lowest values of all inequality measures.

Tab. 3 Characteristics of income*, Poland, 2000-2015

Year	Mean	Median	Gini (%)	D9/D1	Palma ratio
2000	932.17	760.00	33.45	4.129	1.332
2003	999.06	801.70	35.14	4.637	1.262
2005	1147.49	856.81	37.61	4.834	1.629
2007	1178.80	927.76	35.51	4.480	1.490
2009	1305.49	1034.31	35.24	4.242	1.474
2011	1369.52	1091.51	34.69	4.074	1.395
2013	1372.41	1147.87	33.79	4.063	1.357
2015	1484.55	1216.34	32.01	3.850	1.168

* real income (2000 prices); income is adjusted according to OECD modified equivalence scale
Source: Based on data from Council for Social Monitoring (2015)

Fig. 1 Histograms of income, Poland, 2000-2015



2000

2003

2005

2007

2009

2011

2013

2015

Source: Based on data from Council for Social Monitoring (2015)

The visualisation of income distributions (figure 2) allows to stay that income distributions are characterized by right-skewness (this is a characteristic of all income distributions). It should be noted that shapes of distributions were changing in 2000-2015 and the peaks at the beginning of the period were more clear than at the end of analysed period. Based on figure 2 it can be stated that more households receive income close to median marked on the picture as the vertical lines completed by circles. It is also visible that the median of income is higher from year to year.

In the next step there were fitted the theoretical models to the empirical income data. Tab. 4 shows the best fitted distributions to data in subsequent years of analysed period. The table also contains results of Anderson-Darling test.

Tab. 4 Best fitted models to income data, 2000-2015

Year	Distribution	Parameters	Anderson-Darling test* Statistic	 Reject?
2000	Burr	k=0.81416, α=3.3608, β=694.46	1.3146	No
2003	Burr 4P	k=0.59265, α=4.6165, β=834.31, γ=-194.68	1.0684	No
2005	Log-logistic 3P	α=2.653, β=856.85, γ=29.437	2.9231	Yes
2007	Burr	k=0.75949, α=3.2684, β=822.2	0.7154	No
2009	Burr	k=0.80785, α=3.2276, β=943.2	3.9895	Yes
2011	Burr	k=0.85299, α=3.1941, β=1024.7	4.9974	Yes
2013	Log-logistic 3P	α=3.0838, β=1119.0, γ=2.2724	5.2333	Yes
2015	Burr	k=0.88842, α=3.374, β=1175.2	3.7784	Yes

*critical value of 2.5018 at α = 0.05

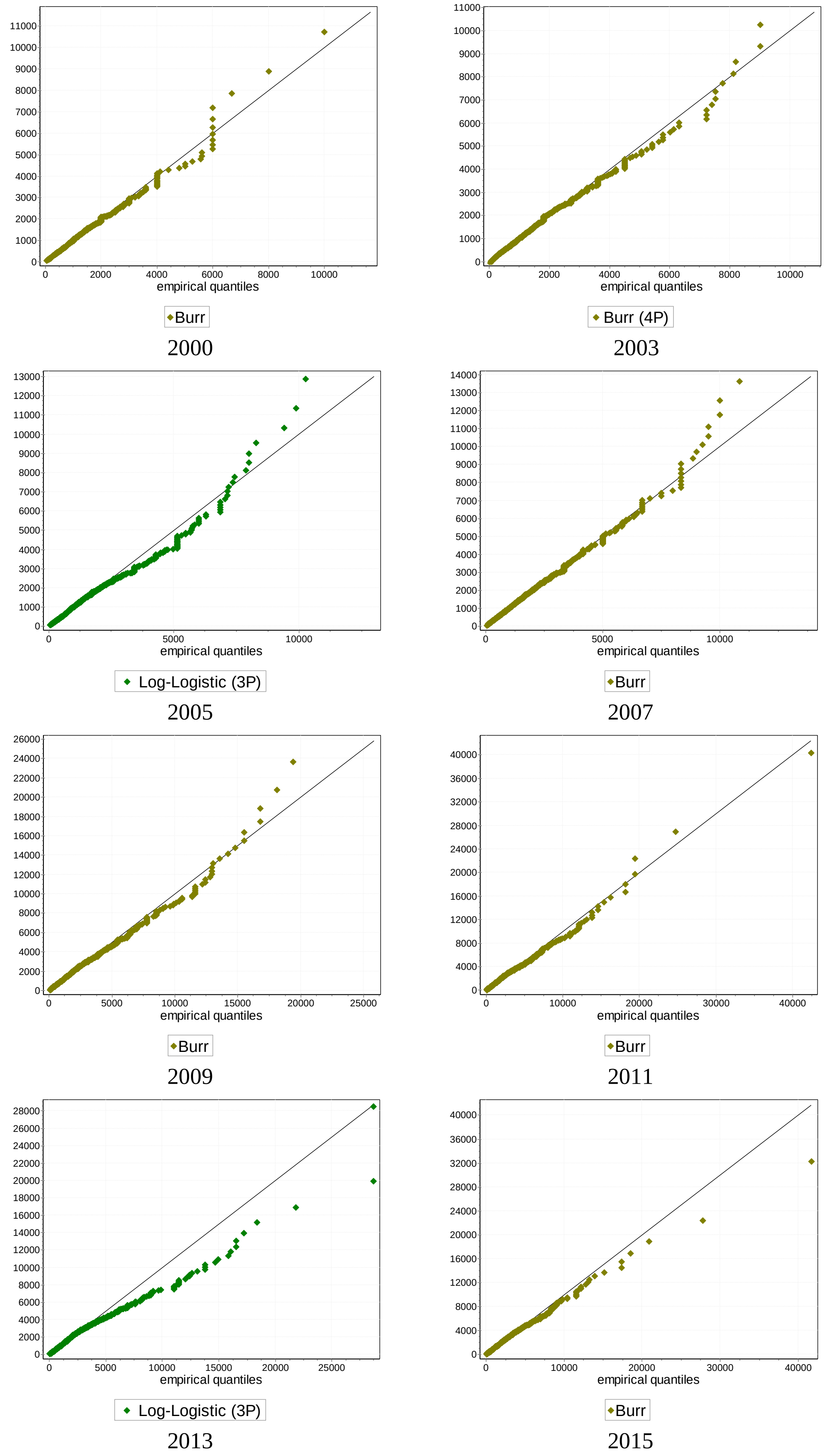
Source: Based on data from Council for Social Monitoring (2015)

Results of the goodness of fit tests suggest that the Singh-Maddala (six times) and log-logistic (two times) models in Poland give the best fit to data for all years under the study. In 2000, 2003 and 2007 the theoretical models were better fitted to empirical data than in other years (according to Anderson-Darling test not to reject null hypothesis). In 2005 and in 2009-2015 the null hypothesis should be rejected, which means that the data does not follow the theoretical distributions.

Additionally, there were used quantile-quantile plots (figure 2), which show that at the beginning of the analysed period (years 2000 and 2003) the theoretical distributions were better fitted to income data than in the subsequent years. In these years the plots are characterized by relatively straight lines and it leads to conclusion that the income distribution follow a given distribution.

Goodness of approximation of income distribution was also investigated by comparing empirical and estimated characteristics (table 4). Empirical and theoretical values of D9/D1 ratio, Gini coefficient and Palma ratio show the systematic decrease in income inequality in Poland after 2005. The causes of decreasing inequality may be various kinds. One of the reasons is undoubtedly accession of Poland to European Union (EU) and connected with this event migration of part of unemployed persons and inflow of money to farmers. Detailed analysis of causes is beyond the scope of this paper. Returning to comparisons, it can be concluded that the theoretical models tend to underestimate the mean income and Gini coefficient. The models also usually overestimate the D9/D1 and the Palma ratios. It should be noted that the Palma ratio is overestimated in all analysed years except the first year of the study. In other studies the fitted theoretical models underestimate or overestimate (depending on measure) the inequality measures. On the one hand, Salamaga (2016) analysing income distribution in Poland indicates that Singh-Maddala model overestimates the mean income and income inequality measures (analysis of income distributions for men and women in the Malopolska voivodship). On the other hand, Ostasiewicz (2013) points out the tendency to underestimation of Singh-Maddala model (analysis of income distribution for big city in Poland).

Fig. 2 Quantile-quantile plots, Poland, 2000-2015



2000

2003

2005

2007

2009

2011

2013

2015

Source: Based on data from Council for Social Monitoring (2015)

Tab. 5 Characteristics of the theoretical distributions, Poland, 2000-2015

Year	Distribution	Mean	Median	Gini (%)	D9/D1	Palma
2000	Burr	931.85	758.12	33.37	4.106	1.323
2003	Burr (4P)	1001.80	797.01	35.02	4.436	1.535
2005	Log-logistic (3P)	1125.00	886.29	37.69	4.931	1.752
2007	Burr	1186.70	929.07	35.93	4.461	1.520
2009	Burr	1282.80	1041.50	34.97	4.402	1.501
2011	Burr	1360.50	1099.90	34.18	4.300	1.431
2013	Log-logistic (3P)	1341.10	1121.20	32.43	4.145	1.398
2015	Burr	1475.80	1234.80	31.56	3.893	1.248

Source: Based on data from Council for Social Monitoring (2015)

4. Conclusions

In the conducted analysis there were fitted income distributions in Poland. The analysis covered the years 2000-2015. The analysis shows that mean and median income were systematically rising in observation period (the real income is analysed). The values of inequality measures were rising to 2005 and in subsequent years the values were decreasing. In 2000-2015 income distributions are mainly follow Singh-Maddala and log-logistic distribution. Income distributions changed between 2000 and 2015: there are the visible changes in graphical presentation of empirical income data and in worse fitting theoretical distributions to data. The results clearly show that income situation of Polish households is still changing and the process of transformation is not finished yet. This suggests that systematic study of the income distribution is needed.

References

Bordley R.F., McDonald J.B., Mantrala A. (1996): *Something new, something old: parametric models for the size distribution of income*. "Journal of Income Distribution", Vol. 6, pp. 91-103.

Champernowne D.G. (1952): *The graduation of income distributions*. "Econometrica", Vol. 20, pp. 591-615.

Cobham A., Sumner A. (2013): *Is it all about the tails? The Palma measure of income inequality*, CGD Working Paper, No. 343, CGD, Washington DC.

Council for Social Monitoring (2015): *Integrated database*. Council for Social Monitoring.

Dagum C. (1977): *A new model of personal income distribution: specification and estimation*. "Economie Appliquée", Vol. 30, pp. 413-437.

Fisk P.R. (1961): *The graduation of income distribution*. "Econometrica", Vol. 29, pp. 171-184.

Huang S., Oluyede B.O. (2014): *Exponentiated Kumaraswamy-Dagum distribution with applications to income and lifetime data*. "Journal of Statistical Distributions and Applications", Vol. 1 pp. 1-8.

Kleiber C., Kotz S. (2003): *Statistical size distributions in economics and actuarial sciences*, John Wiley & Sons.

Kot S.M. (2000): *Ekonometryczne modele dobrobytu*, Wydawnictwo Naukowe PWN.

Majumder A., Chakravarty S.R. (1990): *Distribution of personal income: development of a new model and its application to U.S. income data*. "Journal of Applied Econometrics", Vol. 5, pp. 189-196.

McDonald J.B., Xu Y.J. (1995): *A generalization of the beta distribution with applications*. "Journal of Econometrics", Vol. 66, pp. 133-152.

OECD (2017): *Income inequality (indicator)*. Organization for Economic Co-operation and Development.

Ostasiewicz K. (2013): *Adekwatność wybranych rozkładów teoretycznych dochodów w zależności od metody aproksymacji*. "Przegląd Statystyczny", Vol. 4, pp. 499-521.

Pareto V. (1896): *La courbe de la répartition de la richesse*. Reprinted in Busoni G. (ed.) (1965): *OEeuvres complètes de Vilfredo Pareto, Tome 3: Écrits sur la courbe de la répartition de la richesse*, Librairie Droz. English translation in: "Rivista di Politica Economica", Vol. 87/1997, pp. 647-700.

Salamaga M. (2016): *Modelowanie rozkładów dochodów kobiet i mężczyzn w województwie Małopolskim*. "Wiadomości Statystyczne", Vol. 8, pp. 32-44.

Sen A.K. (1997): *On economic inequality*. Enlarged edition with a substantial annexe J. Foster, A. Sen: *On* *economic inequality after a quarter century*. Clarendon Press.

Singh S.K., Maddala G.S. (1976): *A function of size distribution of income*. "Econometrica", Vol. 44, pp. 963-973.

Vielrose E. (1960): *Rozkład dochodów według wielkości*, Polskie Wydawnictwo Gospodarcze.