

A Service of



Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre

Lukasiewicz, Piotr; Karpio, Krzysztof; Orlowski, Arkadiusz

Working Paper

Two-part models of income distributions in Poland

Institute of Economic Research Working Papers, No. 64/2017

Provided in Cooperation with:

Institute of Economic Research (IER), Toruń (Poland)

Suggested Citation: Lukasiewicz, Piotr; Karpio, Krzysztof; Orlowski, Arkadiusz (2017): Two-part models of income distributions in Poland, Institute of Economic Research Working Papers, No. 64/2017, Institute of Economic Research (IER), Toruń

This Version is available at: https://hdl.handle.net/10419/219887

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.



https://creativecommons.org/licenses/by/3.0/

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.





Institute of Economic Research Working Papers

No. 64/2017

Two-part models of income distributions in Poland

Piotr Łukasiewicz, Krzysztof Karpio, Arkadiusz Orłowski

Article prepared and submitted for:

9th International Conference on Applied Economics Contemporary Issues in Economy, Institute of Economic Research, Polish Economic Society Branch in Toruń, Faculty of Economic Sciences and Management, Nicolaus Copernicus University, Toruń, Poland, 22-23 June 2017

Toruń, Poland 2017

© Copyright: Creative Commons Attribution 3.0 License

Piotr Łukasiewicz

piotr_lukasiewicz@sggw.pl

Department of Informatics, Warsaw University of Life Sciences (SGGW), Nowoursynowska 159, 02-776 Warsaw, Poland

Krzysztof Karpio

krzysztof_karpio@sggw.pl

Department of Informatics, Warsaw University of Life Sciences (SGGW), Nowoursynowska 159, 02-776 Warsaw, Poland

Arkadiusz Orłowski

arkadiusz_orlowski@sggw.pl

Department of Informatics, Warsaw University of Life Sciences (SGGW), Nowoursynowska 159, 02-776 Warsaw, Poland

Two-part models of income distributions in Poland

JEL Classification: C51; C52; D31

Keywords: income distribution; incomes models; Pareto model; power law

Abstract

Research background: studies of structures of incomes distributions have been performer for about 15 years. They indicate that there is no one model which describes the distributions in their whole range. This effect is explained by the existence of different mechanisms yielding to low-medium and high incomes. While more than 97% of the distributions can be described by models with two or three parameter, high incomes (about 3% or less) is in agreement with power law.

Purpose of the article: the aim of this paper is an analysis of the structure of distributions of households' incomes in Poland. By using various models we verify the hypothesis about two-part structure of those distributions.

Methodology/methods: the studies are based on the households' budgets microdata for years 2004-2012. The two-component models were used to describe the incomes distributions. The major parts of the distributions have been described by the two or three parametric models: lognormal, Dagum, and Singha-Madalla. The highest incomes were described by the Pareto model.

Findings: one has showed that two or three parametric models explain from about 95% to more than 99% of ranges of income distributions. The poorest agreement with data is for lognormal model, while the best agreement has been obtained for Dagum model. Regarding the highest incomes the Pareto model describe the data very well only for the selected years. For the remaining years the results are not so obvious. The tails of the income distributions seems to have more complex structure.

Introduction

The studies of structures of the income distributions have been performer for about 15 years. They indicate that there is no one model which describes the distributions in their whole ranges. This effect has been observed for the distributions of incomes in U.S., United Kingdom, Germany and Japan. In the majority of studies incomes are described the best by lognormal model with power law tail. Suoma (2001, pp. 463-470) studied Japanese income distributions for years 1887 – 1998. He showed that two-part model, lognormal with power law tail is the universal structure describing distributions of personal incomes in Japan. In the paper (Nirei & Souma, 2004, pp. 161-168) the authors continue researches and propose dynamic stochastic model explaining power low tails. Dragulescu & Yakovenko (2001, pp. 213-221) study the income distributions in United Kingdom (1994-1999) and in individual U.S. states (1998). They describe income distributions by two-part model: exponential and power law. Nirei & Souma (2007, pp. 440-459) studied income distributions in Japan and the U.S. for years from 1960 to 1999. They confirmed the hypothesis about two-part structure of income distributions. They described the left-central part of the distributions by the exponential model and the top 1% of incomes – by the power law. Clementi & Gallegati (2005, pp. 3-14) investigated income distributions of households in the U.S. (1980-2001), United Kingdom (1990-2001) and Germany (1990-2002). A lowmiddle income group was approximated by lognormal function and a high income group by power law function.

The motivation to take the studies presented in this paper was small number of similar studies for the new countries of European Union. In this paper we conducted such a study of income distributions in Poland. This subject was investigated by Jagielski & Kutner (2010, pp. 615-618). They analyzed incomes from Household's Budgets Survey but the highest incomes data were extracted from rank of the 100 richest Poles. They showed that Polish income distributions may have three-part structure.

The aim of this paper is analysis of the structure of the income distributions. We verify the hypothesis about two-part structure of the income distributions. In the first step we fit well-known models of incomes: lognormal, Dagum, and Singh-Maddala to the whole income distributions. As expected none of the above models describe data for the highest incomes. In the second step we fit power law (Pareto model) to the tails of the income distributions.

Data and data selection

Data from the Household Budget Survey (HBS) project from 2004 to 2012 have been used in this work. One selected subset of data containing microdata

about available monthly incomes in the households. Household's available income is a sum of household's gross incomes from various sources reduced by all income taxes as well as by social security and health insurance taxes.

The zero or negative incomes (about 0.6% - 0.7%) have been removed from data. Income of each household has been recalculated into the annual income in thousands PLN and expressed as the income per person. The number of data records varied from about 32,000 to 37,000 depending on year. One constructed empirical cumulative distributions based on the detailed data to evaluate power models and to present the results. The empirical cumulative distribution is defined:

$$F_{emp}(x_i) = \frac{k_i}{N}$$

where data x_i , i = 1, ..., N are sorted ascending, k_i is rank of income x_i .

Methodology

We take into account three commonly used models: lognormal, Dagum, and Singh-Maddala. Probability density function (pdf) of lognormal distribution is

$$f_{LN}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right),\tag{1}$$

where x > 0, while the σ parameter fulfills the condition $\sigma > 0$. The μ and σ parameters are interpreted as mean value and standard deviation of incomes logarithms respectively. Cumulative density function (cdf) of lognormal distribution can be expressed by the cdf Φ of the standard normal distribution

$$F_{LN}(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right). \tag{2}$$

Lognormal distribution has been often used to describe distributions of wages and incomes. Pdf of Dagum distribution (Dagum, 2008, pp. 3-25) is described by the equation:

$$f_D(x) = \frac{\alpha\beta\delta}{x^{\beta+1}(1+\alpha x^{-\beta})^{\delta+1}},\tag{3}$$

where x > 0, while the parameters fulfill the conditions: $\alpha > 1$ and β , $\delta > 0$. The pdf of Singh-Manddala distribution (Singh & Manddala, 1976, pp. 963-970) can be expressed by the formula:

$$f_{SM}(x) = \frac{\alpha\beta\delta}{x^{-\beta+1}(1+\alpha x^{\beta})^{\delta+1}},$$
(4)

where x > 0 and α , β , $\delta > 0$, $\beta \delta > 1$. Cumulative density functions of the above models are described by the equations:

$$F_D(x) = (1 + \alpha x^{-\beta})^{-\delta} \tag{5}$$

$$F_{SM}(x) = 1 - (1 + \alpha x^{\beta})^{-\delta}$$
 (6)

Studies performed in various countries show that models (3) and (4) exhibit high conformance with empirical distributions of incomes (Bandourian, *et al.*, 2002, p. 47; Dagum, & Lemmi, 1988, pp. 123-157; Kleiber, 1996, pp. 265-268). They are universal, they may describe zero- as well as one-modal distributions (see Łukasiewicz, *et al.*, 2012, pp. B82-B85). Dagum model is used very often in studies of incomes (see i.e. Łukasiewicz, & Orłowski, 2003, pp. 122-130; Łukasiewicz, & Orłowski, 2004, pp. 146-151; Quintano, & D'Agostino, 2006, pp. 525-546).

The Pareto model Type I has been used to describe the highest incomes (tails of the distributions). The model, known also as a power law, contains one parameter and the functions pdf and cdf are of the forms:

$$f_P(x) = \alpha x_m^{\alpha} x^{-\alpha - 1}, \qquad (7)$$

$$F_P(x) = 1 - x_m^{\alpha} x^{-\alpha}, \tag{8}$$

where $x \ge x_m$ and parameter $\alpha > 0$. A limit value of income is indicated by x_m . The pdf and cdf are equal to 0 for $x < x_m$.

All the models have been evaluated by means of the nonlinear least square method utilizing Levenberg-Marquardt algorithm. Thus, the coefficients of the models are estimated by a minimization of the function:

$$SSE(\theta) = \sum_{i=1}^{n} (y_i - f(x_i; \theta))^2, \qquad (9)$$

where θ is a vector of the model's parameters.

The models (1) - (3) were evaluated based on the grouped data. In the case of the model (7) obtained results were unstable because of the small number of counts for the highest incomes. The Pareto model's parameter α has been evaluated based on the cumulative data using the function (8).

The limit values of x_m were evaluated for each model and year after estimating the functions (1) - (3). The x_m was determined as the income above which the model's residuals start rising.

Analysis and results

The lognormal, Dagum, and Singh-Maddala models of incomes were fitted to the empirical distributions. The results are listed in Table 1. The columns contain: parameters' estimators, coefficients of determination $R^2 = 1 - SSE$, limits x_m and values of the theoretical cdf: $F(x_m)$. The latter is a percentage of the income distribution (percentage of households) explained by the model.

Table 1. Estimations of the lognormal (*LN*), Dagum (*D*), and Singh-Maddala (*SM*) models. There are standard errors of the parameters' estimators in the brackets. The symbols μ , σ are the parameters of the lognormal model

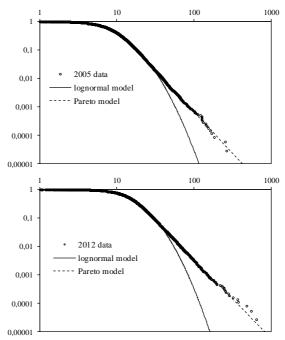
Year	Model	α (μ)	β (σ)	δ	R^2	χ_m	$F(x_m)$
2004	LN	2.149 (0.004)	0.612 (0.003)		0.995	31.59	0.983
	D	838.0 (76.1)	2.995 (0.030)	0.787 (0.016)	0.998	59.04	0.997
	SM	0.003 (<0.0001)	2.502 (0.016)	1.421 (0.041)	0.998	36.16	0.989
2005	LN	2.166 (0.003)	0.608 (0.002)		0.995	29.83	0.979
	D	1285.0 (112.8)	3.124 (0.029)	0.735 (0.014)	0.996	72.26	0.999
	SM	0.0028 (<0.0001)	2.531 (0.015)	1.378 (0.036)	0.997	34.88	0.986
	LN	2.250 (0.002)	0.593 (0.002)		0.996	33.10	0.981
2006	D	1220.0 (113.4)	3.045 (0.031)	0.810 (0.016)	0.997	102.40	> 0.999
	SM	0.0020 (<0.0001)	2.609 (0.015)	1.327 (0.034)	0.998	42.55	0,991
2007	LN	2.349 (0.002)	0.562 (0.002)		0.996	27.47	0.956
	D	1884.0 (109.3)	3.132 (0.019)	0.851 (0.010)	0.998	42.11	0.987
	SM	0.0011 (<0.0001)	2.784 (0.010)	1.241 (0.019)	0.998	32.35	0.972
2008	LN	2.479 (0.002)	0.556 (0.001)		0.996	32.36	0.962
	D	2060 (73.0)	3.050 (0.011)	0.927 (0.007)	0.999	103.10	0.999
	SM	0.0007 (<0.0001)	2.860 (0.007)	1.155 (0.010)	0.999	54.93	0.992
	LN	2.562 (0.002)	0.559 (0.002)		0.995	33.35	0.954
2009	D	2649.0 (168.3)	3.041 (0.019)	0.916 (0.012)	0.998	95.81	0.998
	SM	0.0006 (<0.0001)	2.776 (0.010)	1.298 (0.020)	0.999	37.59	0.968
	LN	2.616 (0.001)	0.559 (0.001)		0.996	34.64	0.949
2010	D	2611.0 (96.0)	2.995 (0.011)	0.955 (0.007)	0.999	73.30	0.995
	SM	0.0005 (<0.0001)	2.831 (0.006)	1.186 (0.011)	0.999	57.74	0.989

Table 1 (cont.)

Year	Model	$\alpha(\mu)$	$\beta(\sigma)$	δ	R^2	x_m	$F(x_m)$
2011	LN	2.660 (0.002)	0.555 (0.002)		0.994	36.75	0.952
	D	5603.0 (261.5)	3.176 (0.013)	0.850 (0.008)	0.999	80.34	0.996
	SM	0.0004 (<0.0001)	2.836 (0.009)	1.217 (0.016)	0.999	42.98	0.969
2012	LN	2.707 (0.003)	0.557 (0.002)		0.992	40.27	0.959
	D	7823.0 (567.3)	3.222 (0.022)	0.814 (0.010)	0.998	47.70	0.976
	SM	0.0004 (<0.0001)	2.753 (0.012)	1.386 (0.026)	0.998	40.27	0.959

Source: own calculation based on the HBS microdata

Figure 1. Complementary cumulative density functions of the lognormal (solid line) and Pareto (dashed line) models for years: 2005 and 2012 in log-log scale. The horizontal axis: annual income in thousands PLN, the vertical axis: percentage of the households



Source: own preparation based on the HBS data

All the evaluated models describe empirical data very well. They are characterized by the high coefficients of determination and very small errors of their parameters. The values of R^2 are similar to each other for all the models. The smallest values of R^2 are observed for the lognormal model, which describes the smallest part of the incomes distributions: from 94.9% to 98.3%, depending on year. On the other hand, lognormal model doesn't explain from 1.7% (2004) to 5.1% (2010) of incomes, Singh-Maddala: from 0.8% (2008) to 4.1% (2012), and Dagum: from 0.1% (2006) to 2.4% (2012). Models' functions of incomes are plotted in Fig. 1 for years: 2007, 2009 and 2012. There are complementary cumulative density functions in the figure. They are given by the equations:

$$\overline{F}_{emp}(x_i) = 1 - F_{emp}(x_i)$$
 and $\overline{F}(x) = 1 - F(x)$.

In order to emphasize differences between the empirical and theoretical distributions the plots are on the log-log scale.

In the next step, tails of the empirical distributions have been approximated by the Pareto model. The tails have been defined in this work as incomes satisfying the inequality $x \ge x_m$. Further on, we will take into account and discus two cases: when the limit values x_m have been determined for: (i) lognormal model and (ii) Dagum model. For those two models one obtained the minimum and the maximum values of x_m respectively (see Table 1). In the case of Singh-Maddala model values of x_m are slightly bigger than in (i) and values of estimation parameters for Pareto model are similar to those in (i). Because of that, these results are omitted in this paper. Results of the estimations of Pareto model are presented in Table 2. The plots of the Pareto functions are in Fig. 2.

In the case (i) the quality of the Pareto model's fits is very high. We also observe very small errors of the α parameter. In the first year (2004) the Pareto exponent has value of 3.04, and in the following years has values 2.65 \div 2.93. The power law exponent is very stable in time, it's changes are small and values around 2.80. For comparison, Clementi & Gallegati (2005, pp. 3-14) obtained the Pareto exponents: $1.10 \div 3.34$ for the U.S. (1980–2001); $3.47 \div 5.76$ for the UK (1991–2001) and $1.63 \div 2.14$ for Germany (1990– 2002). In the case (ii) the results are more dispersed than in the case (i). Dagum model does not explain only below 0.5% of incomes for most of the year. The range of incomes which could be regarded as tail is very narrow. Such a good agreement of this model with income data is emphasized in the empirical studies (Bandourian, et al., 2002, p. 47; Kleiber, 1996, pp. 265–268). However, analysis of the residuals and SSE shows that Dagum model overestimates income distributions in their central part for most of the analyzed years. In other words, Dagum model fits well in the tail of the distribution at the expense of quality of the fit in the middle part.

Table 2. Estimations of the Pareto model. Values of limits x_m have been set for lognormal (*LN*) and Dagum (*D*) models. Values s_α indicate standard errors of parameters' estimators α

Year	Model	x_m	α	Sα	R^2
2004	LN	31.59	3.036	0.008	0.996
2004	D	59.04	3.205	0.043	0.983
2005	LN	29.83	2.887	0.006	0.996
2003	D	72.26	2.542	0.046	0.984
2006	LN	33.10	2.865	0.004	0.998
2006	D	102.40	2.252	0.060	0.986
2007	LN	27.47	2.669	0.003	0.998
2007	D	42.11	2.567	0.085	0.994
2000	LN	32.36	2.649	0.007	0.989
2008	D	103.10	1.756	0.050	0.960
2009	LN	33.35	2.928	0.003	0.998
2009	D	95.81	2.438	0.040	0.979
2010	LN	34.64	2.732	0.008	0.983
2010	D	73.30	1.902	0.026	0.963
2011	LN	36.75	2.911	0.005	0.994
2011	D	80.34	2.747	0.036	0.969
2012	LN	40.27	2.761	0.005	0.996
2012	D	47.70	2.714	0.007	0.994

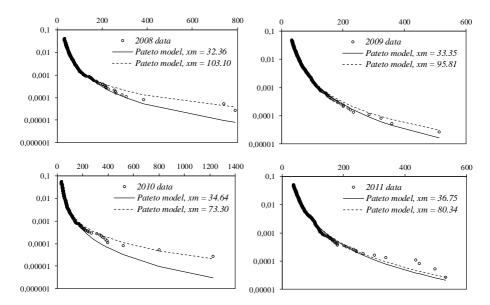
Source: own calculations based on the HBS data

This effect is also visible in Table 1: Singh-Maddala model explains smaller parts of income distributions than Dagum model, whereas is characterized by the same or even greater goodness of fits. In the case (ii) the models of tails (Table 2) are characterized by lower qualities of fits and greater errors than in the case (i). Pareto exponent is $1.76 \div 3.21$. The results are more dispersed than in the case (i).

The lognormal models with power law tails are stable through the years: σ parameter is about 0.56 since 2007 and the Pareto exponent has values close to 2.8. The μ parameter increases with average income. The biggest discrepancies are observed in right tail-ends of the distributions. The tails are described by Pareto model very well for the majority of years. There are some deviations of the model from empirical data in the far ends of the distributions for years: 2008, 2010, 2011 (see Fig. 2). However, errors of the data points in

those regions of the distributions are relatively large. Those differences seems to be assigned to the statistical fluctuations. Further detailed studied should provide knowledge about the size of such fluctuations and their influence on the model's parameters.

Figure 2. Tails of income distributions and Pareto model fits for 2008, 2009, 2010 and 2011 in semi-log scale. Values of limits x_m have been set for lognormal model (solid line) and Dagum (dashed line). The horizontal axis: annual income in thousands PLN, the vertical axis: percentage of the households



Source: own preparation based on the HBS data

Summary

The income distributions of Polish households in 2004 - 2012 were studied in this paper. One evaluated an agreement with data of three models of incomes: lognormal, Dagum, and Singh-Maddala. The lowest goodness-of-fits were observed for lognormal model, while the best fits were for Dagum. None of the analyzed models were able to describe the distributions in their whole ranges with sufficiently high precision. The biggest discrepancies were observed in right tail-ends of the distributions.

Dagum model describes almost the whole range of the income distributions in the majority of years, unexplained tails are short. That's why the tails cannot be described with sufficiently high precision by Pareto model. The Power law exponents are not stable in time and have relatively big errors. At the same time Dagum model overestimates data in the middle of income distributions. In the case of Singh-Maddala model the results were similar to those obtained using lognormal model.

The lognormal models with power law tails are stable through all years. This two-part model will be used in future studies. The discrepancies between the model and data will be further investigated. Incomes of individuals in the households will be analyzed using this model.

References

- Bandourian, R., McDonald, J. & Turley, R. S. (2002). A Comparison of Parametric Models of Income Distribution Across Countries and Over Time. Luxembourg Income Study Working Paper No. 305. DOI: http://dx.doi.org/10.2139/ssrn.324900.
- Clementi, F., & Gallegati, M. (2005). Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States. In A. Chatterjee, S. Yarlagadda, & B. K. Chakrabarti (Eds.) *Econophysics of Wealth Distributions*. Springer-Verlag. DOI: http://dx.doi.org/10.1007/88-470-0389-X 1.
- Dagum, C. (2008). A New Model of Personal Income Distribution: Specification and Estimation. In D. Chotikapanich (Ed.) *Modeling Income Distributions and Lorenz Curves*. Springer. DOI: http://dx.doi.org/10.1007/978-0-387-72796-7 1.
- Dagum, C., & Lemmi, A. (1989). A Contribution to the Analysis of Income Distribution and Income Inequality and a Case Study: Italy. In D. J. Slottjee (Ed.), *Research on Economic Inequality*, 1. Greenwich CT: JAI Press.
- Dragulescu, A. A., & Yakovenko, V. M. (2001). Exponential and Power-Law Probability Distributions of Wealth and Income in the United Kingdom and the United States. *Physica A*, 299. DOI: http://dx.doi.org/10.1016/S0378-4371(01)00298-9.
- Jagielski, M., & Kutner R. (2010) Study of Households' Income in Poland by Using the Statistical Physics Approach. *Acta Physica Polonica A*, 117(4). DOI: http://dx.doi.org/10.12693/APhysPolA.117.615.
- Kleiber, C. (1996) Dagum vs. Singh-Maddala Income Distributions. *Economics Letter*, 53(3). 265-268 DOI: http://dx.doi.org/10.1016/S0165-1765(96)00937-8.
- Łukasiewicz, P., & Orłowski A. J. (2004). Probabilistic models of income distributions. *Physica A*, 344(1-2). DOI: http://dx.doi.org/10.1016/j.physa.2004.06.106.
- Łukasiewicz, P., & Orłowski, A. J. (2003) Probabilistic models of income distributions of Polish households. *Quantitative Methods in Economic Research III*.
- Łukasiewicz, P., Karpio, K., & Orłowski, A. J. (2012). The models of personal incomes in USA. *Acta Physica Polonica A*, 121(2B). DOI: http://dx.doi.org/10.12693/APhysPolA.121.B-82.
- Nirei, M., & Souma, W. (2004). Income distribution and stochastic multiplicative process with reset events. In M. Gallegati, A. P. Kirman, & M. Marsili (Eds.) *The Complex Dynamics of Economic Interaction*. Springer, Berlin, Heidelberg. DOI: http://dx.doi.org/10.1007/978-3-642-17045-4 9.

- Nirei, M., & Souma, W. (2007). A two factor model of income distribution dynamics. *Review of Income and Wealth*, 53(3). DOI: http://dx.doi.org/10.1111/j.1475-4991.2007.00242.x.
- Quintano, C., D'Agostino, A. (2006). Studying inequality in income distribution of single-person households in four developed countries. *Review of Income and Wealth*, 52(4). DOI: http://dx.doi.org/10.1111/j.1475-4991.2006.00206.x.
- Singh, S. K., & Manddala, G. S. (1976). A Function for Size Distribution of Incomes. *Econometrica*, 44(5). DOI: http://dx.doi.org/10.2307/1911538.
- Suoma, W. (2001). Universal Structure of the Personal Income Distribution. Fractals, 09(04). DOI: http://dx.doi.org/10.1142/S0218348X01000816.