Modelling Income Data Using Two Extensions of the Exponential Distribution

Enrique Calderín–Ojeda[a]
Francisco Azpitarte[b]
Emilio Gómez–Déniz[c]

[a] *Centre for Actuarial Studies, Department of Economics, The University of Melbourne, Australia.*
[b] *Melbourne Institute of Applied Economics and Social Research, The University of Melbourne, Australia and Brotherhood of St Laurence.*
[c] *Department of Quantitative Methods in Economics and TiDES Institute. University of Las Palmas de Gran Canaria, Spain.*

Abstract

In this paper we propose two extensions of the Exponential model to describe income distributions. The Exponential ArcTan (EAT) and the composite EAT–Lognormal models discussed in this paper preserve key properties of the Exponential model including its capacity to model distributions with zero incomes. This is an important feature as the presence of zeros conditions the modelling of income distributions as it rules out the possibility of using many parametric models commonly used in the literature. Many researchers opt for excluding the zeros from the analysis, however, this may not be a sensible approach especially when the number of zeros is large or if one is interested in accurately describing the lower part of the distribution. We apply the EAT and the EAT–Lognormal models to study the distribution of incomes in Australia for the period 2001–2012. We find that these models in general outperform the Gamma and Exponential models while preserving the capacity of the latter to model zeros.

Key Words: *Income distribution; Australia; Mixture model; Exponential distribution; Lognormal distribution; Zero Income.*

1 Introduction

The parametric analysis of income distributions has received considerably attention in the economics and econophysics literature. Following the pioneering work of Vilfredo Pareto (1897), many functional forms have been proposed in the literature to study income distributions [1]. The statistical performance of these models will depend on the features of the data and the capacity of the model to capture those characteristics. In particular, the choice of the parametric model is highly influenced by the presence of observations with zero incomes. Largely overlooked in the income distribution literature, the presence of zeros rules out the possibility of using models which have been proven to give a good fit to income data like the Lognormal, Gamma or GB2 as these models do not include the zero in their domains. Many researchers overcome this problem by excluding the observations with zero incomes from the analysis and assuming that the density at zero equals zero. This approach, however, is likely to be invalid, especially when the number of zeros is large and the analyst is interested in describing the bottom part of the income distribution for the study of poverty, inequality or the polarization of incomes.

An alternative approach is to analyse income distributions using parametric models that can accommodate the zeros when fitting the model to the empirical data. This a more sensible approach as the analyst ought to make the best use of the information from the data including those observations with zero income. The Exponential model is particularly suitable option as it has positive density at zero. In fact, As Banerjee *et al.* (2006) show using Australian data, the Exponential distribution gives a good description of most of the income distribution although it fails to capture some features of the upper part of the distribution.

This paper contributes to the existing the literature by proposing two simple extensions of the Exponential model to describe income distributions: the Exponential ArcTan (EAT) distribution which is achieved by using the methodology derived in Gómez–Déniz and Calderín–Ojeda (2015a) or in Gómez–Déniz and Calderín–Ojeda (2015b) and the composite EAT–Lognormal model which is derived following the procedure given in Calderín–

[1] This includes the Lognormal distribution, the Exponential law, as well as more complex models with more parameters such as the Singh–Maddala, the Gamma, and the Generalized Beta of the Second Kind. For a detailed discussion of these models and its application to the analysis of income distributions see Kleiber and Kotz (2003).

Ojeda and Kwok (2015). Thus this paper adds to the limited research on modelling distributions with null or negative values. This includes the so-called Dagum Type–II distribution proposed by Dagum (1977) which is a four-parameter model with positive density at zero. Clementi *et al.* (2009) propose the κ–generalized statistical distribution, a three-parameter model with positive density at zero, to analyse the income distribution in the U.S. They found that this model in general outperforms models like the Singh–Maddala and Dagum type I[2]. We illustrate the suitability of the new models using income data for Australia for the period from 2001 to 2012. We fit the models to the distributions of household disposable income which include a non–trivial number of zeros. Our empirical results show that the EAT and EAT–Lognormal provide in general a better fit to the data than the Gamma and the Exponential models. Importantly, this is achieved without significantly increasing the number of parameters of the model which makes these models particularly attractive to model income distributions in the presence of zeros.

The rest of the paper is organised as follows. In Section 2 we present the new models and their most relevant properties. Section 3 discusses the application of the new models to study changes in the distribution of household disposable income in Australia for the period 2001–2012. In the first part of this section we describe the data sources used for the analysis. We then present the main results derived from the empirical application and we discuss the advantages of the models introduced in this paper with respect to other parametric models widely used for the analysis of income distributions. Finally, Section 4 includes the conclusions and some issues for further research.

2 Parametric models

2.1 The Exponential ArcTan distribution

Let us initially consider the half–Cauchy distribution (see Jacob and Jayakumar (2012)) truncated at $\alpha > 0$ with probability density function (pdf) given

[2] Clementi *et al.* (2012) use the same model to study the distribution of wealth in the US. For a review of the parametric models that have been proposed for the analysis of wealth distributions see Clementi and Gallegati (2016, Ch.4).

by

$$f(y)=\frac{1}{\tan^{-1}\alpha}\frac{1}{1+y^2},\quad 0<y<\alpha. \tag{1}$$

Now, let $\bar{F}(x)$ be the survival function of a random variable X with support in $[a,b]$, where a and b can be finite or non–finite and consider also the transformation $y=\alpha\bar{F}(x)$. Then, the corresponding pdf of the random variable X obtained from (1) results

$$f(x;\alpha)=\frac{1}{\tan^{-1}\alpha}\frac{\alpha f(x)}{1+[\alpha\bar{F}(x)]^2}, \tag{2}$$

for $a\leq x\leq b$ and $\alpha>0$. The survival function of X, derived from (2) by integration, is provided by

$$\bar{F}(x;\alpha)=\frac{\tan^{-1}(\alpha\bar{F}(x))}{\tan^{-1}\alpha}. \tag{3}$$

Besides, (2) and (3) are appropriate density and survival functions, respectively when the support of the parameter α is extended to $(-\infty,\infty)-\{0\}$, satisfying that $\bar{F}(x;\alpha)=\bar{F}(x;-\alpha)$. Additionally, by taking in (3) limit when α approaches to zero and applying L'Hospital's rule, it is simple to derive that the parent survival function, $\bar{F}(x)$, is obtained as a limiting case. In particular, when $\bar{F}(x)$ is replaced by the survival function of the exponential distribution, the Exponential ArcTan (EAT) distribution is obtained. The family of survival functions in (3) has been recently applied recently to the classical Pareto distribution showing an outstanding performance in three different scenarios, to model claim size data (Gómez–Déniz and Calderín–Ojeda (2015a)), to describe city size data (Gómez–Déniz and Calderín–Ojeda (2015b)) and finally to derive a parametric family of Lorenz curves (Gómez–Déniz (2015)).

The survival function and pdf of the EAT distribution are provided by

$$\bar{F}(x;\alpha)=\frac{\tan^{-1}(\alpha\, e^{-\theta\, x})}{\tan^{-1}\alpha},\quad x\geq 0\quad\text{and} \tag{4}$$

$$f(x;\alpha)=\frac{1}{\tan^{-1}\alpha}\frac{\alpha\,\theta\, e^{-\theta\, x}}{1+\alpha^2\, e^{-2\theta\, x}},\quad x\geq 0. \tag{5}$$

respectively, where $\theta > 0$ and $\alpha \in (-\infty, \infty) - \{0\}$.

Alternatively the approximation of the $\tan^{-1}$ function by means of second and third–order polynomials and simple rational functions (see Rajan et al. (2006) for details)

$$\tan^{-1} z \approx \frac{z\pi}{2(1+z)}, \quad z > 0, \tag{6}$$

can be used to approximate the survival function given in (3),

$$\bar{F}(x;\alpha) \approx \frac{(1+\alpha)\bar{F}(x)}{1+\alpha\bar{F}(x)}, \quad \text{for } \alpha > 0. \tag{7}$$

Note that the latter expression is related to the family of distributions proposed by Marshall and Olkin (1997). Then, the probabilistic family of distributions introduced in (3) is geometric–minimum stable (see Marshall and Olkin (1997)). In this particular, a simple interpretation of the EAT distribution is described as follows. Let us suppose that $\{X_i\}_{i=1}^{n}$ are independent and identically distributed random variables with cumulative distribution function $F(x)$, where n is random and it follows the probability mass function

$$\Pr(N = n) = \frac{1}{1+\alpha}\left(\frac{\alpha}{1+\alpha}\right)^{n-1}, \quad n = 1, 2, \ldots \tag{8}$$

(i.e. the geometric distribution); then it is easy to show that the marginal survival function of $X = \min\{X_1, X_2, \ldots, X_n\}$ is given by (7). Thus, the new distribution is approximately the maximum order statistics derived from a sample obtained from geometric distribution.

2.2 Basic properties

Some computations yield the Laplace transform of the EAT distribution in terms of the hypergeometric function, ${}_2F_1$. This results

$$\mathcal{L}_X(t) = E(e^{-tX}) = \frac{\alpha\theta}{(t+\theta)\tan^{-1}\alpha}\, {}_2F_1\left(\left\{1, \frac{t+\theta}{2\theta}\right\}; \frac{1}{2}\left(3 + \frac{t}{\theta}\right); -\alpha^2\right),$$

from which the moments of the distribution are obtained. In particular, the kth order moment about the origin of the EAT distribution is

$$E(X^k) = \frac{k!\,\alpha}{2^{k+1}\,\theta^k\,\tan^{-1}\alpha}\,\Phi(-\alpha^2, k+1, 1/2), \tag{9}$$

where $\Phi(z,s,a) = \sum_{j=0}^{\infty} \frac{z^j}{(a+j)^s}$ is the Lerch transcendent function. Additionally, the quantile function can be easily derived and is given by

$$x_u = -\frac{1}{\theta}\,\log\left[\frac{1}{\alpha}\tan\big((1-u)\,\tan^{-1}\alpha\big)\right], \quad 0 < u < 1. \tag{10}$$

Proposition 1. *The EAT distribution is log–concave.*

Proof. This is easily verified by computing

$$\frac{d^2}{dx^2}(-\log f(x;\alpha)) = \left(\frac{2\alpha\theta e^{\theta x}}{\alpha^2 + e^{2\theta x}}\right)^2 > 0.$$

□

As a consequence of this result we have that the EAT distribution is unimodal and its convolution with any unimodal density is again unimodal (see Ibragimov, 1956). Besides, as compared with the exponential distribution, the EAT model is more flexible since it allows for unimodality when $\alpha > 1$ and zeromodality when $\alpha \leq 1$. For the former case the modal value is

$$x_m = \frac{1}{\theta}\,\log\alpha, \quad \alpha > 0. \tag{11}$$

Another result derived from log-concavity property is that the hazard rate function

$$h(x;\alpha) = \frac{f(x;\alpha)}{\bar{F}(x;\alpha)} = \frac{\alpha\theta e^{-\theta x}}{(1+\alpha^2 e^{-2\theta x})\tan^{-1}(\alpha e^{-\theta x})},$$

is non–decreasing for all α. Furthermore, due to the log-concavity, the EAT distribution has an exponential tail, i.e., $f(x;\alpha) = o(\exp(-\mu x))$ for some $\mu > 0$ as $x \to \infty$.

The following result establishes stochastic order $\preceq_{st}$ between random variables following the EAT distribution.

Proposition 2. *Let X and Y two random variables following the EAT distribution. Then, it is verified that,*

$$X \succeq_{st} Y : \bar{F}(x;\alpha) \preceq_{st} \bar{F}(y;\alpha), \quad \forall \alpha.$$

Proof. It is straightforward. □

Finally, next result establishes ordering with respect to the parameter α.

Proposition 3. *Let X a random variable following the EAT distribution with parameters $\alpha \neq 0$ and $\theta > 0$. Then, it is verified that,*

$$\begin{aligned} \alpha_1 < \alpha_2 : \bar{F}(x;\alpha_1) < \bar{F}(x;\alpha_2), \quad \forall x > 0, \\ \alpha_1 > \alpha_2 : \bar{F}(x;\alpha_1) > \bar{F}(x;\alpha_2), \quad \forall x > 0. \end{aligned}$$

Proof. After differentiating (4) with respect to α, it is simple to see that the sign of this derivative depends on the sign of

$$\Phi(\alpha) = e^{-\theta x}(1+\alpha^2)\tan^{-1}\alpha - (1+\alpha^2 e^{-2\theta x})\tan^{-1}(\alpha e^{-\theta x}).$$

Now, having into account that $\Phi(0) = 0$, $\Phi(\infty) = \infty$ and that

$$\Phi'(\alpha) = 2\alpha e^{-2\theta x}\left[e^{\theta x}\tan^{-1}\alpha - \tan^{-1}(\alpha e^{-\theta x})\right] > 0,$$

we conclude that $\Phi(\alpha) > 0$, for all α and the result hence. □

To end this section, we present the following result which is shown without proof.

Proposition 4. *The conditional survival function of the EAT distribution satisfies*

$$\Pr(X > x + t | X > t) \leq \Pr(X > x), \quad \forall \alpha,$$

where the equality only holds when α tends to zero, i.e. the memoryless (forgetfulness) property which characterizes the Exponential distribution.

2.3 Composite EAT–Lognormal distribution

In situations where the mode of the empirical distribution is greater than zero, composite models based on the methodology proposed in Calderín–Ojeda and Kwok (2015) can be simply obtained. In this sense, as the Lognormal distribution has been widely discussed to describe income data, the composite EAT–Lognormal model can straightforwardly derived by using this approach. Certainly, similar composite models based on the EAT distribution can be simply obtained if other distributions with a closed–form expression for the modal value are used as second component of the spliced model.

The key idea behind this procedure consists of using as first component of the continuous composite model, adequate truncation of the EAT up to the modal value (say x_m), estimated from the data and the second part of the distribution uses adequate truncation of the Lognormal distribution.

Then, the probability density function of the composite EAT–Lognormal via mode–matching is given by

$$f(x) = \begin{cases} r \dfrac{1}{\tan^{-1}\alpha - \tan^{-1}(\alpha\, e^{-\theta\, x_m})} \dfrac{\alpha\theta\, e^{-\theta\, x}}{1+(\alpha\, e^{-\theta\, x})^2}, & 0 < x \leq x_m, \\ (1-r)\, \dfrac{\left(1-\Phi\left(\dfrac{\log x_m - \mu}{\sigma}\right)\right)^{-1}}{\sqrt{2\pi}\, x\, \sigma}\, e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2}, & x_m \leq x < \infty, \end{cases} \tag{12}$$

with $\sigma > 0$, $\mu \in \mathbb{R}$, $0 \leq r \leq 1$ and $\alpha > 1$ to define a positive mode. Besides $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. The mixing weight r is given by

$$r = \frac{f_2(x_m)\, F_1(x_m)}{f_2(x_m)\, F_1(x_m) + f_1(x_m)\,(1 - F_2(x_m))}. \tag{13}$$

where f_1 and F_1 are the pdf and cdf of the EAT distribution and f_2 and F_2 are the pdf and cdf of the Lognormal distribution respectively. Now, the mode-matching condition leads to

$$\theta = e^{\sigma^2 - \mu} \log \alpha.$$

Then, by substituting the corresponding densities and distribution functions into (13), the mixing weight is now written as

$$\begin{aligned} r &= \frac{1}{\sqrt{2\pi}\, x_m\, \sigma}\, e^{-\frac{1}{2}\left(\frac{\log x_m - \mu}{\sigma}\right)^2} \left[\tan^{-1}\alpha - \tan^{-1}(\alpha\, e^{-\theta\, x_m})\right] \\ &\times \left[\frac{1}{\sqrt{2\pi}\, x_m\, \sigma}\, e^{-\frac{1}{2}\left(\frac{\log x_m - \mu}{\sigma}\right)^2} \left[\tan^{-1}\alpha - \tan^{-1}(\alpha\, e^{-\theta\, x_m})\right]\right. \\ &+ \left.\left(1 - \Phi\left(\frac{\log x_m - \mu}{\sigma}\right)\right) \frac{\alpha\theta\, e^{-\theta\, x_m}}{1 + (\alpha\, e^{-\theta\, x_m})^2}\right]^{-1}. \end{aligned}$$

The cdf of the composite EAT–Lognormal distribution is provided by

$$F(x) = \begin{cases} r\, \dfrac{\tan^{-1}\alpha - \tan^{-1}(\alpha\, e^{-\theta\, x})}{\tan^{-1}\alpha - \tan^{-1}(\alpha\, e^{-\theta\, x_m})}, & 0 < x \le x_m, \\ r + (1 - r)\, \dfrac{e^{-\mu + \frac{\sigma^2}{2}}\, \Gamma\left(\frac{1}{2}; \left(\frac{\log x - \mu + \sigma^2}{\sqrt{2}\sigma}\right)^2\right)}{2\sqrt{\pi}\left(1 - \Phi\left(\dfrac{\log x_m - \mu}{\sigma}\right)\right)}, & x_m \le x < \infty, \end{cases} \tag{14}$$

where $\Phi(\cdot)$ represents the cdf of the standard normal distribution and $\Gamma(\cdot;\cdot)$ is the incomplete gamma function. The quantile function can be simply derived numerically by inverting the incomplete gamma function.

2.4 Lorenz curve and inequality measures

The Lorenz curve is a powerful tool to measure the distribution of wealth in a society. Following Gastwirth (1971) and the original proposal by Pietra (1915), given a distribution function $F(x)$ with support in the subset of the positive real numbers and with finite expectation μ, the Lorenz curve is defined as

$$L_{F(x)}(p) = \frac{1}{\mu} \int_0^p F^{-1}(x)\, dx, \quad 0 \le p \le 1, \tag{15}$$

where $F^{-1}(x) = \sup\{y : F(y) \le x\}$.

A well–known characterization of the Lorenz curve is that if $L(p)$ is defined and continuous in the interval $[0, 1]$ with second derivative $L''(p)$, the function $L(p)$ is a Lorenz curve if and only if

$$L(0) = 0, \quad L(1) = 1, \quad L'(0^+) \ge 0 \text{ for } p \in (0, 1), \quad L''(p) \ge 0. \tag{16}$$

By inverting $1-\bar{F}(x;\alpha)$ from (4) we get that the Lorenz curve associated to the EAT distribution results

$$L_{F(x;\alpha)}(p) = \frac{1}{\mu}\int_0^p \log\left\{\frac{\tan[(1-t)\tan^{-1}\alpha]}{\alpha}\right\} dt, \tag{17}$$

where μ is obtained from (9). Unfortunately due to the difficulty to solve this integral a closed–form expression of (17) was not achieved.

Alternatively, we can use (7) to get an approximation of the exact Lorenz curve defined in (17), curve given by

$$L_{F(x;\alpha)}(p) = \frac{1}{\theta}\left[(1-p)\log(1-p) + \left(p+\frac{1}{\alpha}\right)\log(1+p\alpha)\right], \quad 0 \leq p \leq 1,(18)$$

which is a genuine Lorenz curve (satisfying the properties given in (17)) and which reduces to the exponential Lorenz curve (see Gastwirth (1971)) when α tends to zero. More effort is necessary to find the Lorenz curve associated to the cdf given in (14).

Now, the corresponding Gini and Pietra indices can be computed straightforwardly . The Gini index is defined as

$$G = 1 - 2\int_0^1 L_\alpha(p)\, dp.$$

Then the Gini index associated to approximation (18) is

$$G = \frac{1}{\alpha^2\theta}[\alpha(1+(1+\theta)\alpha) - (1+\alpha)^2\log(1+\alpha)].$$

Less known but no less interesting, the Pietra index measures the proportion of total income that needs to be reallocated across the population to achieve perfect equality in income. This proportion is given by

$$P = \max_{0\leq p\leq 1}[p - L(p)] = \frac{1}{2\mu}E|X-\mu|$$

and corresponds to the maximal vertical deviation between the Lorenz curve and the egalitarian line (see Pietra (1915) and Frosini (2012)).

Differentiating $p - L_{F(x;\alpha)}(p)$ we find that the Pietra index is attained at $p_0 = \frac{1-e^{-\theta}}{1+\alpha e^{-\theta}}$. Then, the Pietra index is $p_0 - L_{F(x;\alpha)}(p_0)$.

3 Empirical Application with Australian Data

3.1 Data Sources

For the empirical application we use data from the Household Income and Labour Dynamics in Australia (HILDA) survey for the period 2001–2012. This is a nationally representative survey with detailed information about the income of families in Australia which makes it particularly suitable to study changes in the distribution of income in this country. The HILDA survey began in 2001 with a sample of 7,682 households containing 19,914 people. Subsequent waves of HILDA have collected information from members of the original sample and from other new members of their households related to them[3]. Following the income distribution literature, we take the individual as the unit of analysis and assume that individuals' income is equal to the disposable income of the household. This is defined as the net income available to household members and is given by the sum of market income plus government transfers minus personal income taxes. Market income comprises all private incomes in the form of wages and salaries, business and investment income, private pensions, private transfers, and any windfall income received by any household member. Government payments include the value of all public transfers provided by the Australian government, including pensions, parenting payments, scholarships, mobility and carer allowances, and other government benefits. The sum of these income components is reduced by personal income tax payments made by household members during the financial year.

3.2 Results

We use the HILDA data on household incomes to investigate the suitability of the parametric models to describe income distributions and their changes over time. Table 1 below shows the descriptive statistics of the distributions of household disposable for the period 2001–2012. This was a period of strong economic growth in which Australia outperformed most high-income economies. There was a rapid growth in disposable incomes as reflected by the large increase in the mean and median values of these two variables. Importantly for our analysis, there are many families with zero income values where a number of observations with zero disposable income was below 50

[3] For a detailed description of the HILDA sample see Wooden and Watson (2007).

throughout the period considered. In this regard, by 2001 more than 40 individuals reported zero disposable incomes and this group accounted for 0.22 per cent of the whole sample. The proportion of observations with zero market income steadily declined over the period 2001–2007 reflecting the strong economic growth and the increase in market opportunities over that period. The rate of zeros rose in 2008 and 2009 probably due to the increase in unemployment in the aftermath of the Global Financial Crisis. Finally, the proportion of zeros decreased again in the period 2011–2012. In addition to this, histograms for the household disposable income variable for the years 2003, 2007, 2009 and 2012 are displayed in Figure 1. In all years there is positive mass at zero. Furthermore, as it is typical with income distributions, the empirical distributions are unimodal and positively skewed.

Table 1. Descriptive statistics of household disposable income data in Australia, 2001–2012

Year	Observations	Mean	Median	Zeros	% of zeros
2001	19,859	52,826	47,381	44	0.22
2002	18,269	54,193	48,271	29	0.16
2003	17,602	56,357	49,606	24	0.14
2004	17,160	58,162	51,995	26	0.15
2005	17,437	62,614	56,616	27	0.15
2006	17,407	67,872	60,065	15	0.09
2007	17,211	74,302	65,001	15	0.09
2008	17,077	78,137	69,891	24	0.14
2009	17,583	84,705	77,730	47	0.27
2010	17,821	85,828	75,751	33	0.19
2011	23,365	89,697	79,051	37	0.16
2012	23,154	92,971	82,992	20	0.09

Notes: Mean, and median values expressed in Australian dollars

Source: Authors' calculations based on HILDA data

Our main goal is to assess the suitability of the EAT and the composite EAT–Lognormal models to describe distributions of income. In addition, we compare the results for the EAT and the EAT–Lognormal (EATLG) models with those of the Gamma distribution, a model which has been shown to provide a good fit of household incomes (Drăgulescu *et al.* 2001). The parameter estimates of these models for the distribution of household disposable income in Australia for various years are reported in Table 2. The parameters were estimated by the method of maximum likelihood (ML) by directly maximizing the likelihood surface, except for the composite EAT–Lognormal model whose parameters were estimated numerically via a segment-wise maximization by using the function "mle"/"mle2" in *R*. Note that only for the sake of comparison and for the purpose of applying the method of maximum likelihood estimation to the whole sample when using the Gamma distribution, the zeros in the sample have been replaced by ones. Also shown in the table are the standard errors of the ML estimates derived by inverting the Fisher's information function.

Our results show that all models capture the large increase in the mean and median incomes experienced in Australia over some of the years during the period 2001–2012. In order to compare the parameter θ with the mean incomes presented in Table 1, we have denoted $T = 1/\theta$ (e.g. the mean of the exponential distribution) to express this parameter in units of Australian dollars. Additionally the parametrization $g(x) = (T^{\alpha}\Gamma(\alpha))^{-1}\, x^{\alpha-1}\, \exp(-T/x)$ with x, T, $\alpha > 0$ has been chosen for the Gamma distribution. Thus, for instance, the estimated value of the parameter T of the EAT model steadily increased between 2003 and 2012 which is consistent with the increase in the average income documented above. Similarly, we find that the estimate of the parameter μ of the composite EAT–Lognormal increased over the period under analysis reflecting the rise in the mean and median incomes of families in Australia over the last decade. Additionally, the value of the parameter α for the EAT distribution is greater than one, across the years considered, indicating the fact that its density is unimodal what is consistent with the histograms displayed in Figure 1. To illustrate the impact of excluding the observations with zero incomes from the analysis, Table 3 reports the parameter estimates for the four models without taking into account the zeros. The omission of zeros clearly influences the estimation results, especially in the case of the Gamma model which appears to be more sensitive to the exclusion of zeros than the other models. Thus, for both the Gamma and the EAT models, the value of the parameter $\hat{T}$ drops when zeros are excluded. However, the fall is substantially larger for the Gamma model for which the estimated coefficient declines by more than 1000 units, whereas the size of the change for the EAT model is below 100 units in most years. In contrast, the value of the parameter $\hat{\alpha}$ of the Gamma and the EAT models increases when zeros are dropped. The impact is larger for the Gamma model as the increase in the estimated coefficient is more than twice that of the EAT model. The composite EAT–Lognormal is more robust to the exclusion of zeros as reflected by the small variation in the estimated coefficients relative to the other models.

Table 2. Parameter estimates (above) and standard errors (below) for the distribution of household disposable income in Australia

	Gamma		EAT		EATLG		
Year	$\hat{T}$	$\hat{\alpha}$	$\hat{T}$	$\hat{\alpha}$	$\hat{\alpha}$	$\hat{\mu}$	$\hat{\sigma}$
2003	22735	2.479	26577	5.599	27.45	10.75	0.604
	252.6	0.025	200.6	0.000	1.116	0.007	0.006
2007	32765	2.268	37512	4.620	24.51	11.00	0.642
	369.1	0.023	286.2	0.000	0.981	0.007	0.006
2009	39599	2.139	40271	5.680	19.15	11.23	0.559
	443.6	0.021	304.5	0.000	0.604	0.006	0.006
2012	39985	2.325	45382	5.129	24.39	11.25	0.617
	388.6	0.020	298.5	0.000	0.819	0.006	0.005

Source: Authors' calculations based on HILDA data

Table 3. Parameter estimates (above) and standard errors (below) for the distribution of household disposable income excluding zeros in Australia

	Gamma		EAT		EATLG		
Year	$\hat{T}$	$\hat{\alpha}$	$\hat{T}$	$\hat{\alpha}$	$\hat{\alpha}$	$\hat{\mu}$	$\hat{\sigma}$
2003	21436	2.633	26508	5.652	28.38	10.75	0.607
	237.65	0.026	199.9	0.000	1.181	0.007	0.006
2007	31658	2.349	37451	4.646	25.07	11.00	0.643
	356.8	0.024	285.6	0.000	1.021	0.007	0.006
2009	35683	2.380	40056	5.788	19.97	11.22	0.562
	398.0	0.024	302.5	0.000	0.648	0.007	0.006
2012	38574	2.412	45308	5.158	24.79	11.25	0.619
	374.3	0.021	297.9	0.000	0.840	0.006	0.005

Source: Authors' calculations based on HILDA data

We analyse the validation of these three models using both theoretical and practical approaches. Theoretical validation is assessed by means of Kullback–Leibler divergence, which is consistent with an information–criterion based approach. Two criteria have been considered: the negative log–likelihood (NLL) and the Hannan–Quinn information–criterion (HQIC) defined as twice the NLL plus twice $(k+1)$ $\log(\log(n))$, where k is the number of estimated parameters and n refers to the sample size. Model selection was also assessed from a practical perspective using the Kolmogorov–Smirnov (KS) and the Crámer–von Mises (CvM) goodness–of–fit measures to quantify the distance between the empirical distribution function (EDF) constructed from the data and the ones generated from the fitted models. Let $\hat{F}$ denote the cumulative distribution function of the fitted model, the original data by $x_1, ..., x_N$ and the ordered data in increasing magnitude by $x_{(1)}, ..., x_{(N)}$. Then the expressions of the KS and CvM statistics are given by:

- Kolmogorov–Smirnov (KS) test statistics: $D = max(D^+, D^-)$, where

$$D^+ = \max_{1\leq j\leq N} \left| \frac{j}{N} - \hat{F}(x_{(j)}) \right|, \; D^- = \max_{1\leq j\leq N} \left| \hat{F}(x_{(j)}) - \frac{j-1}{N} \right|.$$

- Crámer–von Mises (CvM) test statistics:

$$W^2 = \sum_{j=1}^{N} \left[\hat{F}(x_{(j)}) - \frac{2j-1}{2N} \right]^2 + \frac{1}{12N}.$$

Results on the goodness–of–fit of the three parametric models to the distribution of disposable income in Australia with and without zeros are presented in Tables 4 and 5, respectively. Note that for all measures a smaller value indicates a better fit to the data. Further we use the KS and CvM statistics to test the null hypothesis that the observed income data come from the parametric models considered.

We find that the EAT model provides in general the best overall fit of the three models when zeros are considered for the analysis, especially for the years 2009 and 2012. In terms of the NLL, the EATLG model, present a similar performance for the years 2003 and 2007. Interestingly, both the composite and the EAT models significantly outperform the Gamma distribution for all years considered. Thus, these two simple extensions of the Exponential model provide a better fit to the empirical data while preserving the capacity of the Exponential to capture the presence of zeros in the data, since these models have a positive probability density at zero income. In this sense the Exponential distribution has also been fitted to household disposable income data for the year 2012, the estimate mean is $\hat{T} = 92971$ with NLL value of 288037.

On the other hand, when the zeros are excluded from the analysis, we find that the fit of the EAT distribution improves the Gamma for the years 2009 and 2012. The EAT model also significantly improves the goodness–of–fit of the Gamma model, especially when this is measured with the CvM statistic for the years considered. Our results suggest that there exists enough statistical evidence to reject the null hypothesis that the data come from any of the models considered[4]. However, it is relevant to mention that these tests reject the Gamma distribution earlier than the EAT distribution for almost all the years examined.

We also study the goodness–of–fit the three parametric models graphically in Figures 1–3. Figures 1 and 2 compare the estimated density and survival functions for each model with the empirical functions derived from the data disposable incomes for the years 2003, 2007, 2009, and 2012. Inspection of the empirical histograms and the density functions reveals that while the probability of finding zeros in the data is clearly positive in all years, the only models with positive density at zero are the EAT model and the EAT–Lognormal models whereas the Gamma model has zero density at this point. No significant differences were found regarding the survival functions. Indeed, for all years, the survival functions of the three models cross multiple times along the income distribution. Figure 3 shows the log–log plots for each of the models. In these graphs we have plotted the logarithm of the incomes against the log of the ranked position in the sample. Income values reported in the HILDA survey are top–censored which explains the shape of the empirical function at the top end of the distribution. Inspection of the plots reveals that, while all models provide similar fit to the data at the bottom and middle parts of the distribution, the composite EAT–Lognormal model clearly outperforms the other models when fitting the upper tail of the distribution. For these log–log plots only sample values greater than one have been considered. Further analysis using income data adjusted to take differences in household size into account using the modified OECD scales, shows that the EAT–Lognormal model performs even better when top incomes are smoothed taking into account differences in family size[5].

[4] We do not report the p–values of the KS and CvM statistics which take value zero in all the years considered. They were computed via Monte Carlo methods using a simulation size of 10000 repetitions.

[5] These figures are not reported here for the sake of space but are available from the authors upon request.

Table 4. Model validation measures for models of household disposable income in Australia

Year	Measure	Gamma	EAT	EATLG
2003	NLL	206929	206810	206810
	HQIC	413872	413634	413640
	KS	0.038	0.039	0.038
	CvM	7.335	2.882	9.381
2007	NLL	207593	207544	207544
	HQIC	415200	415102	415107
	KS 5	0.039	0.037	0.041
	CvM	8.050	3.880	10.166
2009	NLL	214705	213853	214350
	HQIC	429424	427719	428719
	KS	0.064	0.021	0.047
	CvM	24.360	1.718	10.921
2012	NLL	284279	284014	284251
	HQIC	568572	568041	568521
	KS	0.035	0.033	0.041
	CvM	10.728	3.410	12.277

Source: Authors' calculations based on HILDA data

Table 5. Model validation measures for models of household disposable income excluding zeros in Australia

Year	Measure	Gamma	EAT	EATLG
2003	NLL	206308	206516	206484
	HQIC	412625	413045	412987
	KS	0.035	0.040	0.037
	CvM	4.614	2.978	9.606
2007	NLL	207231	207358	207338
	HQIC	414472	414730	414694
	KS	0.038	0.038	0.042
	CvM	6.460	3.938	10.342
2009	NLL	213583	213255	213705
	HQIC	427175	426524	427429
	KS	0.051	0.022	0.046
	CvM	14.340	1.723	11.168
2012	NLL	283774	283759	283971
	HQIC	567557	567532	567961
	KS	0.033	0.034	0.042
	CvM	8.055	3.481	12.481

Source: Authors' calculations based on HILDA data

To further illustrate the value of the new parametric families for modelling income variables with large number of zeros, we have fitted the Exponential and EAT models to the distribution of household market income. Households with no market income typically include families whose members are permanently out of the labour force, like those in retirement or those with disabilities, and also jobless families whose members of working age are out the labour market due to unemployment. For these households public transfer constitute the main income source as they tend to rely on publics transfers such as the age or disability pensions or unemployment benefits. The proportion of observations with zero market income is much higher than in the case of household disposable income. Thus the proportion of zeros for the year 2012 is 6.93%. In the presence of such a large number of zeros, estimation of the Gamma model using information from the whole sample is not feasible.

The estimate mean of the exponential distribution is $\hat{T} = 51152.4$ and with NLL value of 273683 where as for the EAT distribution we have obtained $\hat{T} = 42889.1$ and $\hat{\alpha} = 1.169$ and NLL equals to 273455. Note that for the latter model, $\hat{\alpha}$ has reduced its value considerably, as compare to household disposable income data, approximating the shape of the density curve to the exponential one. Of course, other spliced models derived from the EAT distribution that mimic some features of the data (e.g bimodality), would certainly improve the fit to the data.

4 Conclusions and further research

This paper contributes to existing literature by proposing two simple extensions of the Exponential distribution to describe income distributions: the two-parameter Exponential ArcTan (EAT) distribution and the composite EAT–Lognormal model. The new models preserve key properties of the Exponential model including its capacity to model distributions with zero incomes. The existence of observations with zero incomes poses an important challenge for the analyst as it precludes the possibility of fitting widely used models such as the Lognormal, Gamma or the GB2 which do not include the zero in their domains. When working with income variables with zeros, many analysts opt for excluding the zeros from the analysis. The extensions of the Exponential model proposed in this paper constitute a more sensible approach as it allows to make the best use of the data using information on zeros without increasing considerably the number of parameters.

The suitability of the new models was evaluated using income data for Australia for the period from 2001 to 2012 which include an small number of zeros. Our empirical results show that the EAT and composite EAT–Lognormal provide in general a better fit to the data than the Exponential and Gamma models. The use of parametric models with positive density at zero is especially important when the number of zeros is large. A future extension of this paper will consider the use of the EAT and other mixture models to describe the distribution of the individual components of family income such as market income, business income or investment income where the proportion of zeros is particularly large due to the large number of families that do not earn income from those sources.

Acknowledgements

This paper uses confidentialized unit record file data from the HILDA Survey. The HILDA Survey Project was initiated and is funded by the Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied

Economic and Social Research. This research was supported by the Australian Research Council Centre of Excellence for Children and Families over the Life Course (project number CE140100027). The Centre is administered by the Institute for Social Science Research at The University of Queensland, with nodes at The University of Western Australia, The University of Melbourne and The University of Sydney. The authors would like to express their gratitude to three anonymous referees for their relevant and useful comments. Enrique and Emilio are funded by the grant ECO2013-47092 (Ministerio de Economía y Competitividad, Spain). Francisco gratefully acknowledges financial support from the Ministerio de Economía y Competitividad (ECO2011-23460, ECO2013-46516-C4-2-R, ECO2014-52616-R).

References

Banerjee, A; Yakovenko, V.M. and Di Matteo, T. (2006) A study of the personal income distribution in Australia. *Physica A: Statistical Mechanics and its Applications*, 370, 54–59.

Calderín–Ojeda and E.; Kwok, C. F. (2015) Modeling claims data composite Stoppa models. *Scandinavian Actuarial Journal*.
DOI: 10.1080/03461238.2015.1034763.

Clementi, F., Gallegati, M. and Kaniadakis, G. (2009) A generalized statistical mechanics approach to income analysis. *Journal of Statistical Mechanics: Theory and Experiments*, P02037, arXiv:0902.0075.

Clementi, F., Gallegati, M. and Kaniadakis, G. (2012) A generalized statistical model for the size distribution of wealth. *Journal of Statistical Mechanics: Theory and Experiments*, P12006, arXiv:1209.4787v2.

Clementi, F. and Gallegati, M. (2016) The Distribution of Income and Wealth: Parametric Modeling with the κ-Generalized Family, Springer-Verlag Italia, Milan.

Dagum, C. (1977). A new model of personal income distribution: Specification and estimation". *Economie Appliquée*, 30:413-436.

Frosini, B.V. (2012). Approximation and decomposition of Gini, Pietra–Index and Theil inequality measures. *Empirical Economics*, 43, 175–197.

Gastwirth, J.L. (1971). A general definition of the Lorenz curve. *Econometrica*, 39, 1037–1039.

Gómez–Déniz, E. and Calderín–Ojeda, E. (2015a). Modelling insurance data with the Pareto ArcTan distribution. *Astin Bulletin*, 45, 3, 639–660.

Gómez–Déniz, E. and Calderín–Ojeda, E. (2015b). On the use of the Pareto ArcTan distribution for describing city size in Australia and New Zealand. *Physica A: Statistical Mechanics and its Applications*, 436, 821–832.

Gómez–Déniz, E. (2015). A family of arctan Lorenz curves. *Empirical Economics*. (in press). DOI: 10.1007/s00181-015-1031-y

Ibragimov, J.A. (1956). On the composition of unimodal distributions (Russian). *Teorija verojatnostej*, 1, 283-288.

Jacob, E. and Jayakumar, K. (2012). On half–Cauchy distribution and process. *International Journal of Statistika and Mathematika*, 3, 2, 77–81.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken, NJ: John Wiley & Sons.

Marshall, A.W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, 84, 3, 641–652.

Pareto, V. (1897) *Cours d'Éconimie Politique*. Laussanne.

Pietra, G. (1915). Delle relazioni tra gli indici di variabilitÃ . *Atti Regio Istituto Veneto*, 74, II, 775–792.

Rajan, S., Wang, S., Inkol, R. and Joyal, A. (2006). Efficient approximations for the arctangent function. *IEEE Signal Processing Magazine*, 23, 3, 108–111.

Rolski, T.; Schmidli, H.; Schmidt, V. and Teugel, J. (1999). Stochastic processes for insurance and finance. John Wiley & Sons.

Wooden, M. and Watson, N. (2007). The HILDA Survey and its Contribution to Economic and Social Research (So Far). *Economic Record*, 83, 261, 208–231.

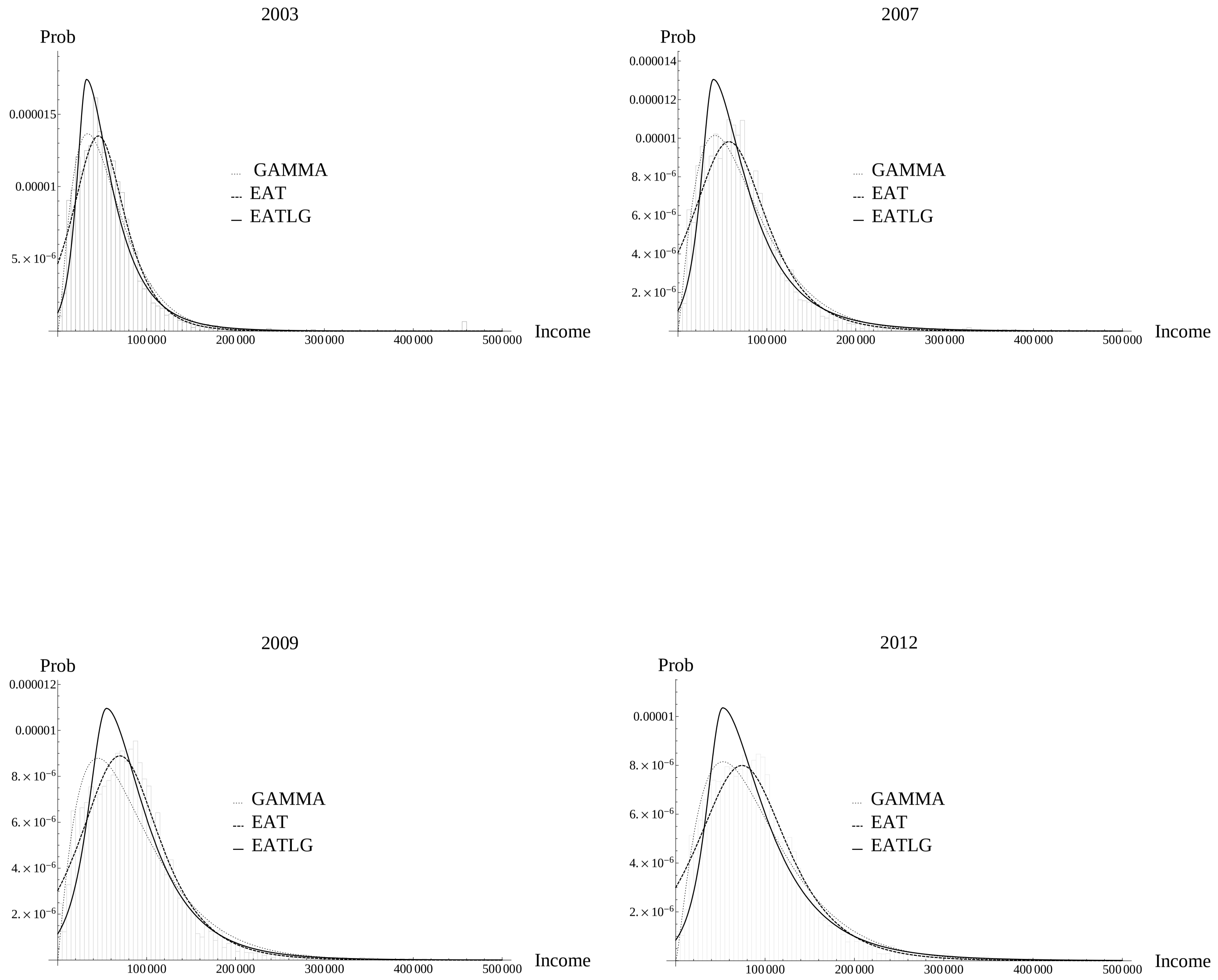

Figure 1: Histograms and density functions for Australian household disposable income data

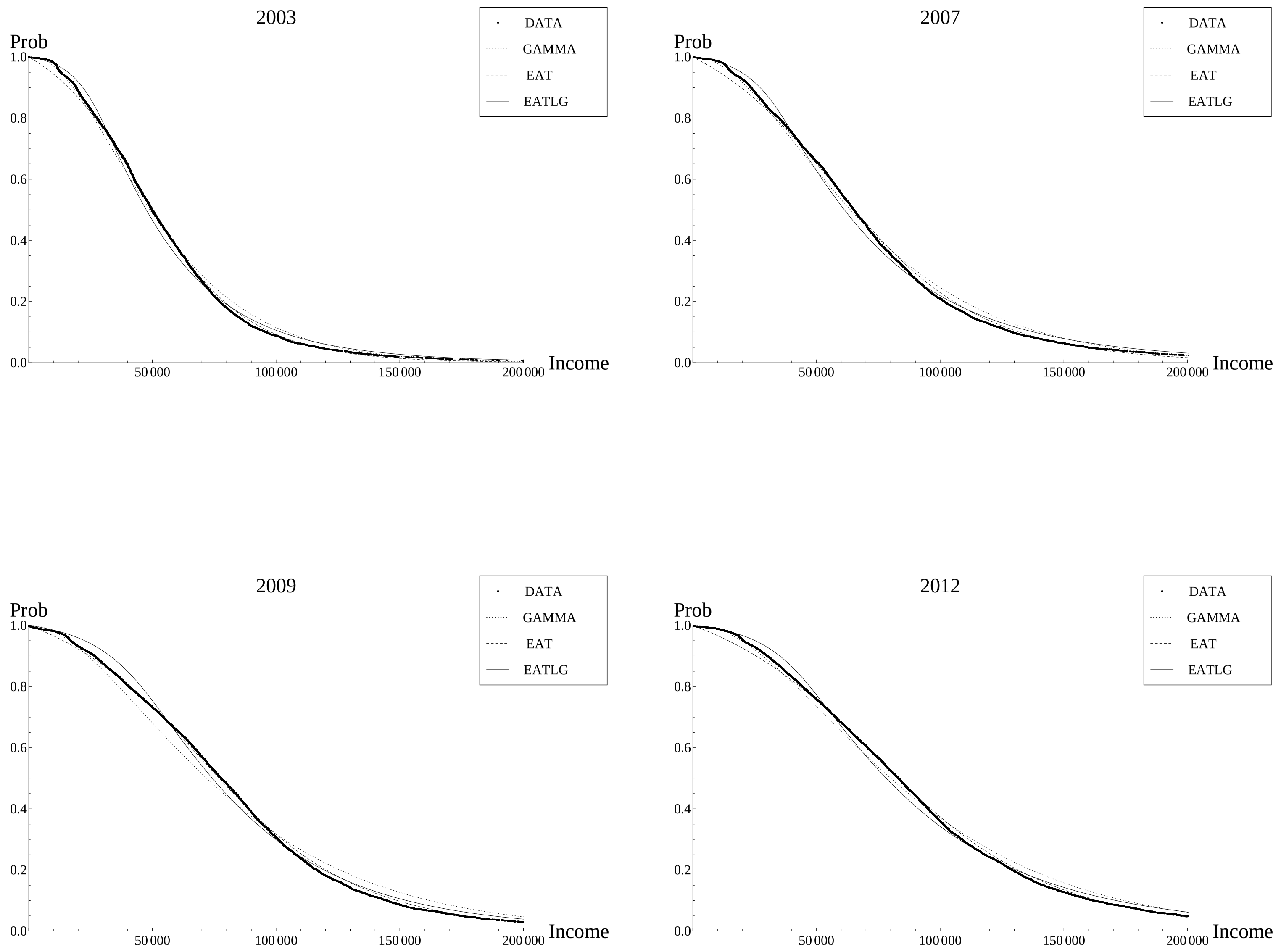

Figure 2: Graphical validation: plots for Australian household disposable income data.

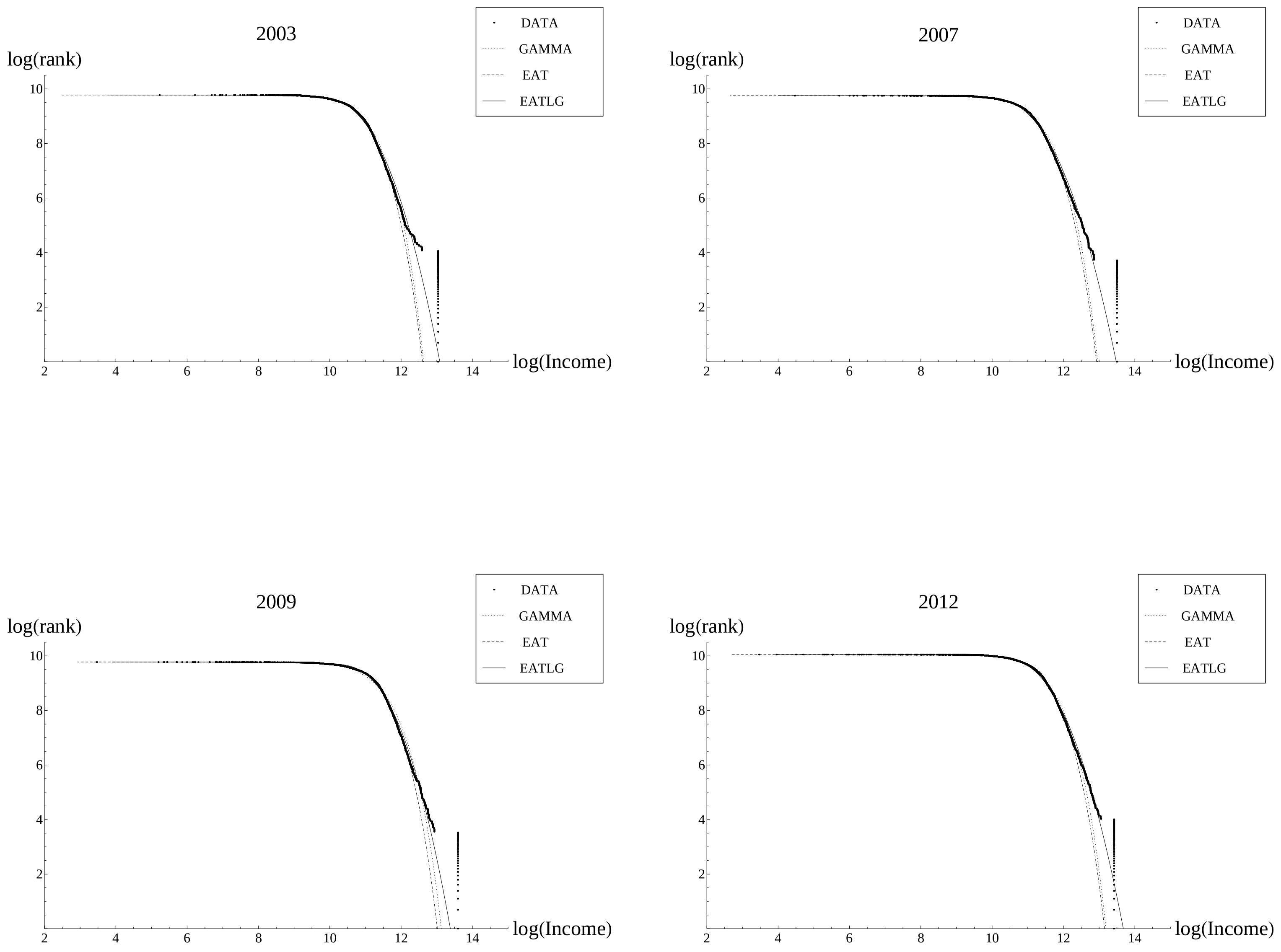

Figure 3: Graphical validation: log-log plots for Australian household disposable income data.