

# **Luxembourg Income Study Working Paper Series**

**Working Paper No. 488**

**On the Evolution of Household Income**

**Giambattista Salinari and Gustavo De Santis**

**July 2008**



---

**Luxembourg Income Study (LIS), asbl**

---

# On the evolution of household income<sup>\*</sup>

*Giambattista Salinari and Gustavo De Santis*

Dept of Statistics "G. Parenti"; Un. of Florence, Italy, July 2008

"E non mi è incognito come molti hanno avuto et hanno opinione che le cose del mondo sieno in modo governate dalla fortuna e da Dio, che li uomini con la prudenzia loro non possino correggerle, anzi non vi abbino remedio alcuno ... Non di manco, perché il nostro libero arbitrio non sia spento, iudico potere essere vero che la fortuna sia arbitra della metà delle azioni nostre, ma che etiam lei ne lasci governare l'altra metà, o presso, a noi."<sup>§</sup>

(N. Machiavelli, *Il Principe*, 1513)

## Abstract

We present a markovian homogeneous model that mimics the evolution of household income. With three parameters only, the model generates a set of theoretical curves that closely fit actual income distributions, as observed in 19 advanced economies in the period 1967-2004. The fit is better, and theoretically more consistent, than that obtained with other models customarily used in the literature, for instance log-normal or power-law models.

## 1. Introduction

The goal of this article is to present and to test a markovian homogeneous model which we deem useful for the description of the evolution of household incomes at the micro level, and also income distributions at the macro level.

The basic idea is that the total income of a household at time  $t+1$  can be thought as the sum of two components: total income at time  $t$  and the variation that takes place in the interval  $(t, t+1)$ . This is not new in itself (see, e.g., Champernowne 1953, Labergott 1959, Goodman 1961, Lydall 1968, Hart 1976, Atkinson et al. 1992, Neal and Rosen 2000, Dutta *et al.* 2001): the novelty lies in the idea that this variation be a random variable, because it is influenced by so many different factors

---

<sup>\*</sup> Financial support from the UE – Sixth Framework Programme: "Major Ageing and Gender Issues in Europe – MAGGIE" (Contract no.: CIT5 – 028571) is gratefully acknowledged. We thank Daniele Vignoli for his comments on an earlier version of this paper.

<sup>§</sup> "It is not unknown to me how many men have had—and still have—the opinion that the world's affairs are governed by chance and by God, in such a way that the wisdom of man cannot channel them or even do anything about them ... Nevertheless, not wishing to dismiss our freedom of will, I believe that chance arbitrates one-half of our actions, but that she still leaves us to manage the other half, or perhaps a little less." (N. Machiavelli, *The Prince*, 1513)

that their combined effect defies understanding and escapes modeling - a dynamic that, in Machiavelli's view, characterizes most of the world's affairs. The characteristics of this random variable, the central point of the theoretical part of this article, are examined in section 4.

This conjecture on the evolution of household income permits us to generate a great variety of model income distributions, which differ by at least one of the three basic parameters of our model, and which can be compared to actual income distributions. Since the fit proves good, actual income distributions can be described with three parameters only, each of which, incidentally, is interpretable in socio-economic terms.

## **2. Sources of data: LIS and ECHP**

In what follows, we will compare model to actual income distributions. The latter come from two different sources: the *LIS* database (Luxemburg Income Study) and the *ECHP* (European Community Household Panel). Both are well known international databases that contain, *inter alia*, data on net individual and household income, detailed by type (e.g. labour, assets, rents, pension benefits, etc). In this paper we will only consider total net household income.

*LIS* data are freely accessible: their characteristics and the related documentation can be found at the web page <http://www.lisproject.org/>. In short, this is a collection of cross sectional surveys on households in various developed countries, for various years, at irregular intervals, in the period 1967-2004. We considered all the countries for which we could find at least two different datasets, spaced by at least 15 years: this left us with 19 countries and 114 different income distributions, summarized in Table 3.

The *ECHP* is a European survey, with a panel structure (that we will ignore here), that investigates several household dimensions. Of all these variables, in this paper we will only consider total net household income. The countries that we take into consideration, only for the years 1994 and 2001 (beginning and end of the survey), are listed in Table 2. The official web site of the *ECHP* is <http://epp.eurostat.ec.europa.eu> (see "Access to microdata", bottom right), but information on the pros and cons of the survey must be looked for elsewhere: for instance Locatelli, Moscato and Pasqua (2001)<sup>1</sup> highlight that imputation was relatively frequent for *ECHP* income data, and that not all countries and all incomes could be entered net of taxes: in some cases, conversion from gross to net took place at a later stage.

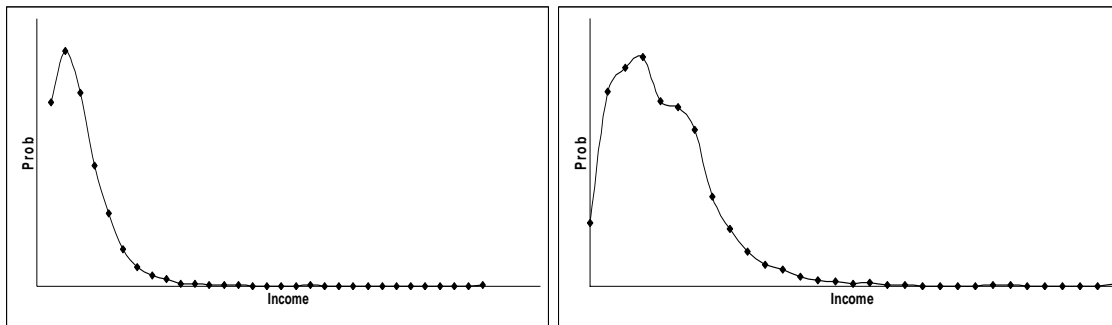
---

<sup>1</sup>See also [http://circa.europa.eu/Public/irc/dsis/echpanel/library?l=/doc\\_pan&vm=detailed&sb=Title](http://circa.europa.eu/Public/irc/dsis/echpanel/library?l=/doc_pan&vm=detailed&sb=Title), or <http://epunet.essex.ac.uk/echp.php>.

### 3. A conjecture to explain similarities over space and time

Income distributions from different countries in different epochs show some common and recurrent features (Neal and Rosen, 2000; Cowell, 2000), the most evident of which is their right-skewedness. Figure 1 provides two examples: Greece in 1994 and Finland in 2001.

**Figure 1.** Income distributions in Greece (1994) and Finland (2001).



a) GR 1994

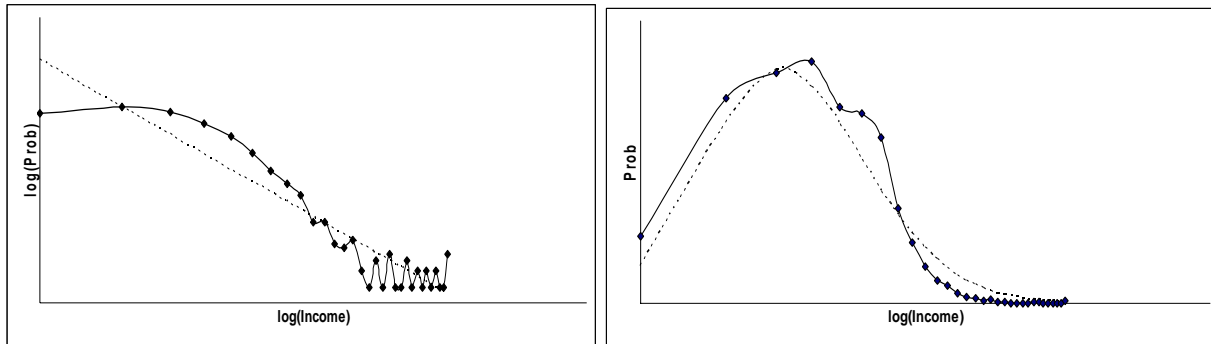
b) FI 2001

Source: own calculations on ECHP data

It is possible to fit empirical income distributions with model functions, for instance power-law or log-normal distributions (Majumder and Chakravarty, 1990; McDonald J.B., Mantrala A., 1995). However, fitting empirical income distributions with these theoretical distributions results in a few systematic fitting errors (Fig. 2). Power-law model distributions lead to an overestimation of the frequencies in the first (poorest) classes; conversely, log-normal fitting yields an overestimation of the terminal values.

These systematic fitting errors may be taken as an indication that supports the following conjecture: empirical income distributions are distinct empirical realizations of unique, but unknown, model of household income evolution. In this paragraph we will attempt to present some other indications supporting our conjecture that this unique, general model for the evolution of household income exists.

**Figure 2.** Interpolation with a power-law function (dotted line) of a) the Greek income distribution in 1994 (log-log scale); b) the Finnish income distribution in 2001 (log scale).



a) GR 1994

b) FI 2001

Source: Own calculations on *ECHP* data.

Let us first compare the income distributions of two rather different countries in two different years, Finland in 2001 and Greece in 1994. Note that we are taking income distributions as they were, without adjusting them, e.g. with equivalence scales<sup>2</sup>.

Table 1 displays the first nine deciles of the Finnish 2001 and the Greek 1994 income distributions. Figure 3, instead, shows the Greek income deciles plotted against the Finnish income deciles.

**Table 1.** Deciles of income distributions in Finland (2001) and in Greece at (1994). Data are expressed in local currencies of the time. In 2001, it took about 5.95 Finnish Markkas to buy one Euro; in 1994, it took about 45.2 Greek Drachmas to buy one Finnish Markka<sup>3</sup>.

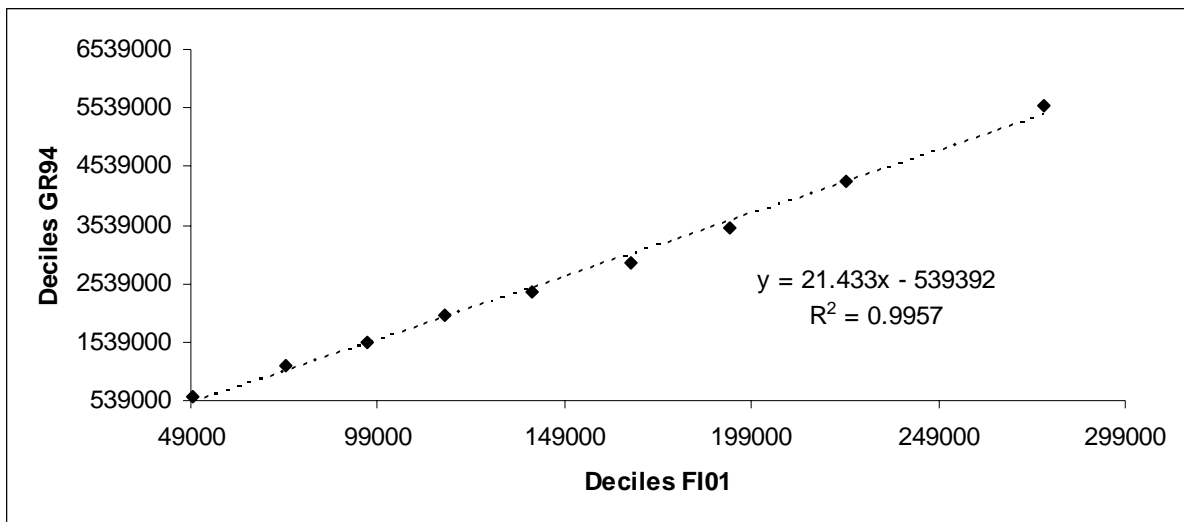
Decile	FI01	GR94
1	49497	605883
2	74505	1120000
3	96253	1540000
4	116957	1988000
5	140256	2389000
6	166929	2877192
7	192996	3480000
8	224240	4269051
9	277008	5567947

Source: own calculation on *ECHP* data.

<sup>2</sup> Our first checks on Italy (*LIS* data) suggest that the model presented below works *better* with equivalence scales. However, since we could not find any convincing hypothesis about the underlying causal mechanism, and since equivalence scales are in large part arbitrary, we decided to stick to unadjusted income data for this paper.

<sup>3</sup> Conversion rates obtained through <http://www.oanda.com/convert/fxhistory>.

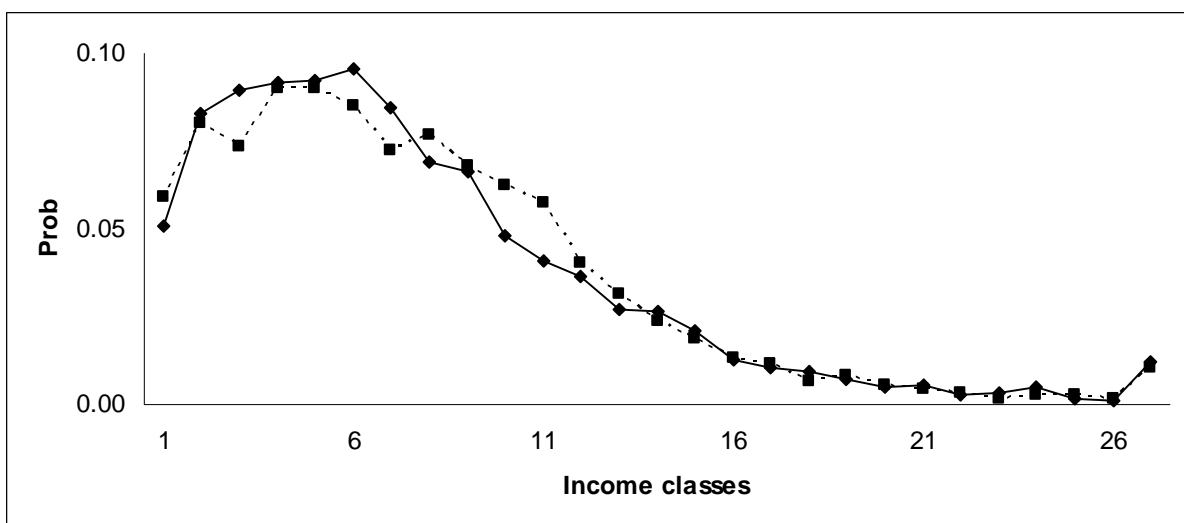
**Figure 3.** Q-Q Plot of Greek income deciles against Finnish income deciles.



Source: Table 1.

The diagram of Figure 3 can be regarded as a simple non-parametric test on the similarity of the shape of the two distributions. The more the points on the Q-Q plot approximate a straight line, the more precisely the two distributions overlap. Fig 3 uses the squared correlation coefficient ( $R^2$ ) to evaluate how well a straight line fits the nine points of the Q-Q plot. The result is good ( $R^2=99.6\%$ ) and this implies that we can convert the Finnish income distribution into the Greek one with a simple linear transformation - more precisely with the transformation given by the linear regression shown in Figure 3. Figure 4 represents the original Greek income distribution against the modified Finnish distribution.

**Figure 4.** Greek original (continuous) and Finnish transformed income distribution (dotted line). The transformation is  $Y = 21.4X + 539,392$  (where  $Y$ =New and  $X$ =Original Finnish income).



Source: own calculation on ECHP data.

Let us now repeat the experiment, "crossing" the income distributions of 12 countries in 1994 and in 2001<sup>4</sup>: the corresponding R squared are shown in Table 2.

**Table 2.** Comparisons between 12 income distributions for year 2001, and 12 for year 1994..

	D94	DK94	NL94	BE94	LU94	FR94	UK94	IL94	IT94	GR94	SP94	PT94
<b>DK01</b>	0.992	0.999	0.996	0.994	0.984	0.987	0.982	0.993	0.993	0.987	0.976	0.964
<b>NL01</b>	0.997	0.995	1.000	0.995	0.987	0.991	0.984	0.996	0.995	0.992	0.982	0.969
<b>BE01</b>	0.996	0.997	0.993	0.999	0.996	0.997	0.996	0.998	0.998	0.996	0.993	0.986
<b>FR01</b>	0.998	0.992	0.992	0.999	0.999	1.000	0.998	0.999	0.998	1.000	0.998	0.992
<b>IL01</b>	0.999	0.997	0.997	0.999	0.995	0.997	0.993	0.999	0.997	0.997	0.991	0.982
<b>IT01</b>	0.998	0.995	0.993	1.000	0.999	0.999	0.998	0.999	0.999	0.999	0.996	0.990
<b>GR01</b>	0.996	0.991	0.988	0.998	0.999	0.999	1.000	0.997	0.996	0.999	0.998	0.994
<b>SP01</b>	0.996	0.989	0.988	0.997	1.000	0.999	0.999	0.997	0.996	0.999	0.999	0.995
<b>PT01</b>	0.994	0.985	0.983	0.995	0.999	0.998	0.999	0.994	0.993	0.998	1.000	0.998
<b>AT01</b>	1.000	0.995	0.997	0.999	0.997	0.999	0.994	1.000	0.999	0.999	0.993	0.985
<b>FI01</b>	0.999	0.997	0.999	0.998	0.993	0.996	0.990	0.999	0.998	0.996	0.988	0.978
<b>SV01</b>	0.996	0.993	0.999	0.993	0.986	0.990	0.982	0.995	0.992	0.990	0.980	0.967

The values shown in the table are those of  $R^2$  (squared correlation coefficient) calculated on the quantile-quantile points, as in Fig. 3. D=Germany (Deutschland); DK=Denmark; LU= Luxembourg; NL= The Netherland; BE= Belgium; FR= France; IL= Ireland; IT= Italy; GR= Greece; SP= Spain; PT= Portugal; AT= Austria; FI= Finland; SV= Sweden; UK= United Kingdom.

Source: own calculation on ECHP data.

Three main observations emerge from Table 2:

- 1) The  $R^2$  values are always extremely high: they range between 0.964 (DK01-PT94) and 1. Their average is .994.
- 2) The  $R^2$  values are higher when the same country is compared in two different epochs: the lowest value for this combination is .998, in Portugal.
- 3) The  $R^2$  values are comparatively low when the combination is between a northern European country in 2001 (FI01, SV01, DK01, NL01, BE01) and a Mediterranean country in 1994 (FR94, IT94, GR94, ES94, PT94). The average value of the  $R^2$  coefficients for this type of comparisons is about 0.985.

These observations are consistent with the following set of conjectures:

- a) all the income distributions considered derive from the same generating mechanism (our main hypothesis - point 1);

<sup>4</sup> We also crossed these distributions in the same years (1994 and 2001, respectively): results are at least as good as those presented in Table 2.

b) "local" peculiarities produce some relatively minor dissimilarities (points 2 and 3). These "peculiarities" may be due to different national redistributive policies, e.g. the type of welfare state at work (Esping-Andersen 1990, 1999)<sup>5</sup>, or to differences in the quality of the data.

#### 4. A markovian homogeneous model for income evolution

Total household income at time<sup>6</sup>  $t+1$  can be defined as  $Y_{t+1}=Y_t+\Delta_t$ , that is the sum of two terms: household income in the preceding year and the variation that has taken place in between. What do we know about this variation? The range of possible events that can modify household income in any given time interval is ample and variable over time and space. To the best of our knowledge, the literature has thus far concentrated on the systematic part of this change, both at the individual level (that is linking it to age, education, ethnicity, household dimension, ...), at the intermediate level (e.g., area of residence) and at the macro level (e.g., economic growth, type of welfare state, ...). But these exogenous<sup>7</sup> variables can explain only a limited part of the variation: most of the change  $\Delta_t$  evades explanation and modelling. This is even more true when one tries to "explain" a series of such variations:  $\Delta_t, \Delta_{t+1}, \Delta_{t+2}, \dots, \Delta_{t+n}$ .

It does not seem unreasonable, therefore, to assume that  $\Delta_t$  is (largely) a random variable. Let us try to work a little bit on this idea.

##### 4.1 General assumptions of the model

To start with, let us subdivide household incomes  $Y$  into an enumerable infinity of income classes 0, 1, ... of the same width  $w$ . For example, class 0 can be defined as ranging from 0\$ to less than 5,000\$; class 1 as 5,000\$—10,000\$, and so on. Given  $w$ , let  $C_t$  ( $C$  for class) be the income class of a household at time  $t$ , and let

$$1) \quad p(C_t = 0)$$

be the probability of finding a given household in the first income class at time  $t$ . Let us also define the probability of moving from class  $h$  to class  $k$  during a give temporal unit ( $t, t+1$ ) as

---

<sup>5</sup> In this case, another mechanism may be at work: the *ECHP* is a panel and the households interviewed in 2001 are basically the same that were interviewed in 1994. However, the same phenomenon (better fit when the same country is compared in two different years) can be observed also on *LIS* data, which are repeated cross sections.

<sup>6</sup> Income is normally defined over a time interval (e.g. 1 Jan. to 31 Dec.), but in this paper, for the sake of simplicity, we will assume that it is concentrated at mid-year. Similarly, variations are normally instantaneous (at midnight of 31 Dec.), but we will instead assume that they take place in the interval between two successive mid-years.

<sup>7</sup> Some, actually, are at least partly endogenous to income: e.g. education, household dimension, place of residence, ...

$$2) \quad p(C_{t+1} = k \mid C_t = h)$$

In order to identify income dynamics, we need to assign a value to all the conditional probabilities indicated by expression 2), and globally described by the (infinite) transition matrix  $\mathbf{M}$ :

$$2') \quad \mathbf{M} = \begin{pmatrix} p(C_{t+1} = 0 \mid C_t = 0) & p(C_{t+1} = 1 \mid C_t = 0) & p(C_{t+1} = 2 \mid C_t = 0) & \dots \\ p(C_{t+1} = 0 \mid C_t = 1) & p(C_{t+1} = 1 \mid C_t = 1) & p(C_{t+1} = 2 \mid C_t = 1) & \dots \\ p(C_{t+1} = 0 \mid C_t = 2) & p(C_{t+1} = 1 \mid C_t = 2) & p(C_{t+1} = 2 \mid C_t = 2) & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

The notation becomes simpler if we denote with  $m_h$  the mean income of class  $h$  at time  $t$ , and with  $Y_h$  the income distribution at time  $t+1$  of the households whose income class was  $h$  at time  $t$ . With this new notation we can now re-write expression (2) as:

$$3) \quad p_{h,k} = p(kw < Y_h \leq kw+w)$$

For instance, with  $w=5$  (in thousand dollars),  $h=6$  and  $k=4$ , the notation in (3) gives the probability that a household, whose income was originally (at time  $t$ ) of class 6 (i.e. 30-35 thousand dollars), ends up in class 4 (20-25 thousand dollars).

Let us now consider the main assumptions of the model.

- 1) household income derives from several potential sources, for each household member: wages, pensions, rents of various nature (financial, land, real estate, etc.), subsidies, insurances, bequests, etc. Each of these sources of income is influenced by several possible determinants: ability, luck, health, external constraints, etc. We can therefore imagine that the random variable  $Y_h$  (income at time  $t+1$  for households belonging to class  $h$  at time  $t$ ) can be approximated by a normal variable ( $Y_h \sim N$ ), if the time lag ( $t, t+1$ ) is not too long;
- 2) higher incomes are generally characterized by more income sources than lower ones (wages, rents, interests, etc), and their variability should therefore be correspondingly greater. Let us simply assume that the variance of  $Y_h$  is a monotonically increasing function  $g$  of the starting class  $h$ :  $\text{Var}(Y_h) = g(m_h)$ ;
- 3) there exists a function  $f$  linking the starting class  $C_h$ , and its average  $m_h$ , to the expected valued of the distribution of end points  $E(Y_h) = f(m_h)$ , with  $f(\cdot) > 0$ , because we assume incomes to be non negative.
- 4) The probability for a household in income class  $h$  at time  $t$ , to end up in class  $k$  at time  $t+1$  (where  $k$  does not necessarily differ from  $h$ ) depends, among other things, on the value of  $w$ , that is on how large our income classes are;

5) the probability  $p_{h,k}$  is constant over time.

In short: we have defined a set of normal distributions  $Y_h$ , depending on the income mean  $m_h$  of class  $h$  at time  $t$ , as follows:

$$4) \quad E(Y_h) = f(m_h) [f(m_h) > 0]$$

$$5) \quad \text{Var}(Y_h) = g(m_h) [g(m_h) > 0; g'(m_h) > 0]$$

#### 4.2 The parameters of our model

Let us first see how to determine  $\text{Var}(Y_h)$ . One of the simplest assumptions that we can make is that the variance of  $Y_h$  increases in proportion to the average income  $m_h$ . If we approximate  $m_h$  (unknown) with the center of its class  $h$  ( $hw + \frac{1}{2}w$ ), and let  $b$  be the proportionality factor, we get:

$$6) \quad \text{Var}(Y_h) = b\left(h + \frac{1}{2}\right)w$$

Where  $b$  is the first parameter of the model<sup>8</sup> and  $w$  the width of class  $h$ .

In order to determine the mean of the random variable  $Y_h$ , three simple solutions may be considered. The first is to imagine that all incomes, independently of the class  $C_h$  from which they come, undergo the same expansion process:

$$7) \quad E(Y_h) = \left(hw + \frac{1}{2}w\right) + e$$

Where  $e$  is the expansion parameter, the same for all income classes.

Alternatively, we can imagine that incomes undergo an expansion process that is proportional to the class to which they belong:

$$8) \quad E(Y_h) = \left(hw + \frac{1}{2}w\right)e$$

---

<sup>8</sup> Eq. (6) forces the variance to increase with  $m_h$  (and  $h$ ), because  $b$  cannot be negative, by construction. However, we have also verified that, with a slightly more complex model  $\text{Var}(Y_h) = a + b(h + 1/2)w$ ,  $\hat{b}$  is still positive and significant. In order to minimize the number of parameters, we decided to keep in eq. (6) only the parameter that interests us most, that is "b".

In this case, expansion is greater for the rich. A third possibility is that incomes expand in a way that is inversely proportional to the class to which the household belongs:

$$9) \quad E(Y_h) = (hw + \frac{1}{2}w) + \frac{e}{hw + \frac{1}{2}w}$$

which means that the poor improve more than average. These three hypotheses can be collapsed into a unique expression with the introduction of a new parameter  $c$  (for *clinamen*<sup>9</sup>):

$$10) \quad E(Y_h) = (hw + \frac{1}{2}w) + e(hw + \frac{1}{2}w)^c$$

Equation 10 encompasses equations (7-9), and more. If  $c=0$ ,  $c=1$  or  $c=-1$  we find, respectively, eqs. (7), (8) or (9). But since  $c$  can be any real number (for instance,  $c=-0.2$ ) we can also find other, more general cases.

Equations (10) and (6) describe the assumed relations between  $m_h$  on the one hand, and the mean and the variance of the variable  $Y_h$  on the other. Transition probabilities, from any starting class to any destination class, depend on a number of assumptions (1 to 4 above) plus three parameters only:  $b$ ,  $c$ , and  $e$ .

In order to calculate the transition probability  $p_{h,k}$  (from class  $h$  to class  $k$ ), we must calculate the integral of the density function of  $Y_h$  in the interval  $(kw, kw+w)$ . This proves easier after standardization of the lower  $L$  and the upper  $U$  limit of the class

$$11) \quad L_{h,k} = \frac{kw - E(Y_h)}{\sqrt{\text{Var}(Y_h)}}$$

$$12) \quad U_{h,k} = \frac{kw + w - E(Y_h)}{\sqrt{\text{Var}(Y_h)}}$$

The transition probability will be then:

$$13) \quad p_{h,k} = \Phi(U_{h,k}) - \Phi(L_{h,k})$$

where  $\Phi$  is the repartition function of the standardized normal distribution.

---

<sup>9</sup> In the Physics of the Greek philosopher Epicurus (341 BC-270 BC), the *clinamen* was the inclination that atoms were thought to follow during their assumed perpetual falling.

Let us now consider the extreme destination classes. If  $k=0$  (that is, income remains or becomes low, and the household ends up in the first class), its lower standardized limit is  $L_0=-\infty$ , and the notation simplifies to

$$14) \quad p_{h,0} = \Phi(U_{h,0})$$

The last class does not exist, strictly speaking, because, as mentioned, we are considering an enumerable infinity of income classes. Empirically, however, we need to define a terminal class  $\omega$  with infinite width, the standardized upper limit of which is therefore  $U_\omega=\infty$ .

The probability transition to such a class from a generic class  $h$  ( $p_{h,\omega}$ ) is given by:

$$15) \quad p_{h,\omega} = 1 - \Phi(L_{h,\omega})$$

Unfortunately, the definition of the transition probability from class  $\omega$  to a generic class  $h$  ( $p_{\omega,h}$ ) creates a few additional obstacles: while all other starting classes have the same width  $w$ , class  $\omega$  has an infinite width. This difficulty can be partially circumvented by choosing a very high lower limit for class  $\omega$ , so as to reduce the probability of finding anybody so rich as to belong to this class.

The transition matrix  $\mathbf{M}$  for the entire system becomes:

$$16) \quad \mathbf{M} = \begin{pmatrix} \Phi(U_{0,0}) & \Phi(U_{0,1}) - \Phi(L_{0,1}) & \dots & 1 - \Phi(L_{0,\omega}) \\ \Phi(U_{1,0}) & \Phi(U_{1,1}) - \Phi(L_{1,1}) & \dots & 1 - \Phi(L_{1,\omega}) \\ \dots & \dots & \dots & \dots \\ \Phi(U_{\omega,0}) & \Phi(U_{\omega,1}) - \Phi(L_{\omega,1}) & \dots & 1 - \Phi(L_{\omega,\omega}) \end{pmatrix}$$

For instance, with  $w=5$  (thousand dollars),  $b=0.65$ ,  $c=-0.64$ ,  $e=2.16$ ,  $\mathbf{M}$  becomes:

$$17) \quad \mathbf{M} = \begin{pmatrix} 0.897 & 0.103 & 0.000 & \dots \\ 0.041 & 0.817 & 0.142 & \dots \\ 0.000 & 0.101 & 0.715 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

#### 4.3 Markov's condition

Let us now imagine that  $\mathbf{M}_t$  and  $\mathbf{M}_{t+1}$  are two stochastic matrices describing the transition probabilities between income classes in time intervals  $(t, t+1)$  and  $(t+1, t+2)$ , respectively. We will assume that they belong to the same family,  $\mathbf{M}$ , but are characterized each by its specific combination of the three basic parameters  $(b, c, e)$ . If we also assume that the transitions of the

second temporal interval  $(t+1, t+2)$  are independent<sup>10</sup> of what happened during the first interval  $(t, t+1)$ , we can calculate the transition probability from  $h$  to  $k$  in the interval  $(t, t+2)$  through matrix multiplication

$$18) \quad p(C_{t+2} = k | C_t = h) = (M_t \times M_{t+1})_{h,k}$$

which is another way of affirming the Markovian nature of the process.

Let now define  $\mathbf{r}_t$  as a row-vector such that each cell  $r_{t,h}$  of  $\mathbf{r}_t$  contains the probability that a household belongs to class  $h$  at time  $t$ :

$$19) \quad r_{t,h} = p(C_t = h)$$

In simpler words,  $\mathbf{r}_t$  describes the income distribution of the population at time  $t$ . Using equation (18) we can now calculate  $\mathbf{r}_t$  from any original distribution  $\mathbf{r}_0$ , with a series of matrix products:

$$20) \quad \mathbf{r}_t = \mathbf{r}_0 \times \mathbf{M}_0 \times \mathbf{M}_1 \times \dots \times \mathbf{M}_{t-1}$$

All the matrices  $\mathbf{M}_0, \mathbf{M}_1, \dots, \mathbf{M}_{t-1}$  belong to the same type, but their parameters  $(b_t, c_t, e_t)$  are not necessarily the same, because they vary with time. It is possible, however, to simplify matters by using only one, "average" transition matrix  $\mathbf{M}$  for all the unitary intervals in the time span  $(0, t)$ . This matrix  $\mathbf{M}$  will be characterized by the mean value of the three basic parameters,  $b, c,$  and  $e$  respectively. With this simplification in mind, we can rewrite the process, described by (20), as

$$21) \quad \mathbf{r}_t = \mathbf{r}_0 \times \mathbf{M}^{t-1}$$

As for the original distribution  $\mathbf{r}_0$  we will use a row-vector such that:

$$22) \quad r_{0,h} = \begin{cases} 1 & \text{for } h = 0 \\ 0 & \text{otherwise} \end{cases}$$

In short, we are assuming that there was an original epoch, no matter how remote, when all households were relatively poor, and belonged to the first class ( $C_0$ ). From this original state of generalized poverty, evolution begins ...

---

<sup>10</sup> This assumption is implicit in our original assumption concerning the independence of the forces affecting income during the temporal unit  $(t, t+1)$ .

## 5. Income structure evolution in 19 countries

Let us now turn to the question: how well does our model (eq. 21) fit empirical data? In order to see this, we selected 19 countries from the *LIS* database, with at least 15 years of presence in the database, not necessarily in consecutive years. We thus collected information on 114 different income distributions, over the years 1967 to 2004. Income here means: total household net disposable income (DPI variable in the *LIS* database), converted into international dollars through the Pen World Table (<http://pwt.econ.upenn.edu/>).

Table 3 summarizes the main characteristics of our data.

**Table 3.** Summary of the main characteristics of data.

Country	Start	End	Span	N Surveys	Size	Gini	H
AT	1987	2000	13	5	7,706	0.26	8.3
AU	1981	2003	22	6	10,908	0.3	12
BE	1985	2000	15	6	3,904	0.25	6.5
CA	1971	2000	29	9	26,192	0.29	10.9
CH	1982	2002	20	4	5,125	0.29	8
DK	1987	2004	17	5	54,106	0.23	7.7
FI	1987	2004	17	5	10,899	0.23	5.4
FR	1979	2000	21	6	9,571	0.29	8.4
IE	1987	2000	13	5	2,868	0.32	12.4
IL	1979	2001	22	5	4,696	0.33	12.9
IT	1986	2000	14	8	7,981	0.32	12.1
MX	1984	2002	18	8	11,467	0.48	21
NL	1983	1999	16	5	4,551	0.25	5.6
NO	1979	2000	21	5	9,274	0.24	6.4
SE	1967	2000	33	7	11,202	0.23	7.1
SP	1980	2000	20	4	13,936	0.33	12.5
TW	1981	2000	19	6	15,059	0.28	7
UK	1969	1999	30	7	12,252	0.31	10.5
US	1974	2004	30	8	42,713	0.35	17

Start= first distribution of the series. End= last distribution. Span= End-Start. N. Surveys= number of known income distributions during period considered. Size= mean sample size along Period. Gini= average Gini index for the period. H= average headcount poverty index for the period. AT=Austria; AU=Australia; BE=Belgium; CA=Canada; CH=Switzerland; DK=Denmark, FI=Finland; FR=France; IE=Israel; IL=Ireland; IT=Italy; MX=Mexico; NL=The Netherlands; No=Norway; SE=Sweden; SP=Spain; TW=Taiwan; UK=United Kingdom; US=United States.

Source: own calculations on *LIS* data.

The mean length of these historical series (span) is about 20 years. We work with 30 classes of income, with a width of 5 (thousand dollars), plus a final open class for incomes greater than 150,000\$. Therefore, for each country, we obtain a set of vectors  $\{\mathbf{r}_{t1}, \mathbf{r}_{t2}, \dots, \mathbf{r}_T\}$  with 31 cells, describing the nominal income distribution for the years  $t1, t2, \dots, T$ .

In each country, the process is assumed to start from an original distribution  $\mathbf{r}_0$  as in (22) above. Our goal is to fit an income transition process that, starting from  $\mathbf{r}_0$ , goes through  $\mathbf{r}_{t1}$ , then  $\mathbf{r}_{t2}$ , and finally reaches  $\mathbf{r}_T$  (latest distribution available). We fit the incomes evolution process using natural number  $n_1, n_2, \dots$  and an average transition matrix  $\mathbf{M}$ , and look for the best approximation that we can get:

$$23) \quad \begin{cases} \mathbf{r}_{t1} \approx \mathbf{r}_0 \times M^{n_1} \\ \mathbf{r}_{t2} \approx \mathbf{r}_0 \times M^{n_2} \\ \dots \\ \mathbf{r}_T \approx \mathbf{r}_0 \times M^{n_T} \end{cases}$$

In practice we let the three model parameters ( $b$ ,  $c$ , and  $e$ ) vary, so as to generate the set of theoretical distributions ( $\mathbf{r}_0 \times M^{n_1}, \mathbf{r}_0 \times M^{n_2}, \dots, \mathbf{r}_0 \times M^{n_T}$ ) that get closest to the empirical income distributions ( $\mathbf{r}_i, \mathbf{r}_j, \dots, \mathbf{r}_T$ ).<sup>11</sup>

$$24) \quad b, c, e, n_1, n_2, \dots, n_T : \max[\rho(r_h, r_0 \times M^{n_1}) + \rho(r_k, r_0 \times M^{n_2}) + \dots + \rho(r_T, r_0 \times M^{n_T})]$$

Our results<sup>12</sup> are shown in Table 4. Figures 5-7, below, show a detailed description of the incomes evolution in three different countries - United States, France and Sweden – taken as representative of different welfare regimes (liberal, conservative, social democratic), in Esping-Andersen's (1990, 1999) classification.

The empirical application is, all in all, satisfactory. The country for which the model returns the best fit Mexico, where the mean  $\rho^2$  (on 8 income distributions, 1984 to 2002) is 0.998. The country where the model produces the worst result is Denmark, for which  $\rho^2=0.949$  (on 5 income distributions, 1987-2004). The Danish case, however, is rather an exception: elsewhere the mean  $\rho^2$  is always greater than 0.979.

Table 4 also compares the goodness of fit that we obtain with our model (col. 4) and with a more standard approach: log-normal interpolation (col. 5). The global mean  $\rho^2$  for our model application is 0.987, against 0.977 for the log-normal interpolation - which, given the small range of variation in this type of models, means that improvement is substantial. Notice, moreover, that our model uses much fewer parameters: for example, we describe the eight income distributions of the United States in 1974-2004 with only 3 parameters, while the 8 log-normal distributions that do the same need 16 parameters.

---

<sup>11</sup> "Closest" here means "yielding the highest sum of correlation coefficients  $\rho$  between the empirical and the model distributions". Other optimizing procedures yield basically the same estimates (not shown here).

<sup>12</sup> We used both a program of our own creation in *R* and the Excel solver, obtaining very similar results in both cases.

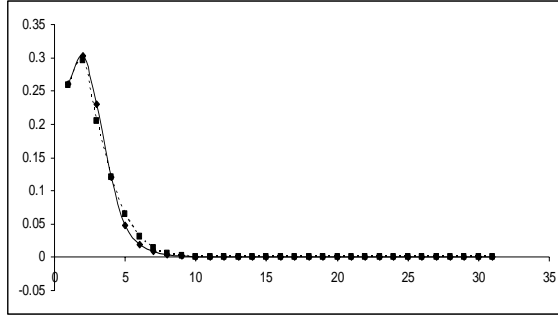
**Table 4.** Results for the model application to the description of income evolution in 19 countries.

<b>Country</b>	$\hat{b}$ (1)	$\hat{c}$ (2)	$\hat{e}$ (3)	av( $\rho^2$ )model (4)	av( $\rho^2$ ) lognorm (5)
<b>MX</b>	0.485	-0.774	2.477	0.998	1.000
<b>FR</b>	0.486	-0.593	2.615	0.997	0.992
<b>CA</b>	0.624	-0.607	2.345	0.996	0.982
<b>US</b>	0.650	-0.644	2.164	0.995	0.963
<b>IL</b>	0.846	-0.599	2.910	0.994	0.983
<b>SP</b>	0.480	-0.854	2.970	0.993	0.992
<b>AT</b>	0.376	-0.524	2.495	0.991	0.980
<b>IE</b>	0.525	-0.915	3.013	0.991	0.981
<b>CH</b>	0.530	-0.324	2.034	0.991	0.971
<b>IT</b>	0.567	-0.695	2.804	0.989	0.990
<b>AU</b>	0.515	-0.772	2.559	0.988	0.976
<b>NL</b>	0.609	-0.718	3.544	0.986	0.976
<b>UK</b>	0.403	-0.810	2.896	0.985	0.986
<b>FI</b>	0.504	-0.527	2.425	0.984	0.962
<b>BE</b>	0.574	-0.814	3.583	0.983	0.982
<b>TW</b>	0.495	-0.652	3.227	0.981	0.977
<b>NO</b>	0.321	-0.377	2.440	0.980	0.947
<b>SE</b>	0.537	-0.705	2.933	0.979	0.972
<b>DK</b>	0.385	-0.778	3.025	0.949	0.955
<b>Mean</b>	<b>0.522</b>	<b>-0.667</b>	<b>2.761</b>	<b>0.987</b>	<b>0.977</b>
<b>St. dev.</b>	<b>0.116</b>	<b>0.155</b>	<b>0.425</b>	<b>0.011</b>	<b>0.013</b>

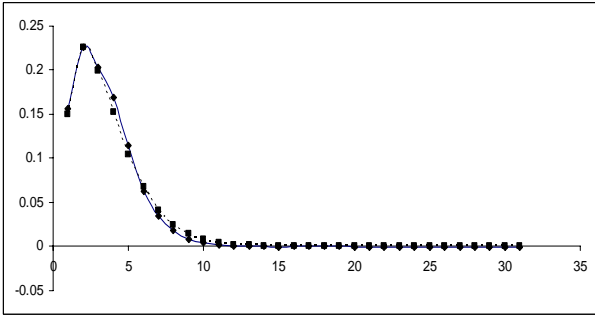
The width of income classes is always  $w=5$  (thousand dollars).  $\hat{b}$ ,  $\hat{c}$ ,  $\hat{e}$  = estimates of model parameters. Av( $\rho^2$ )-model = mean squared correlation coefficient measured between empirical income distributions and the theoretical curves produced by the model for the given values of the  $b$ ,  $c$ ,  $e$  parameter. Av( $\rho^2$ )-lognorm = mean squared correlation coefficient measured between empirical income distributions and log-normal distributions. data sorted by Av( $\rho^2$ )-model.

Source: own calculations on LIS data.

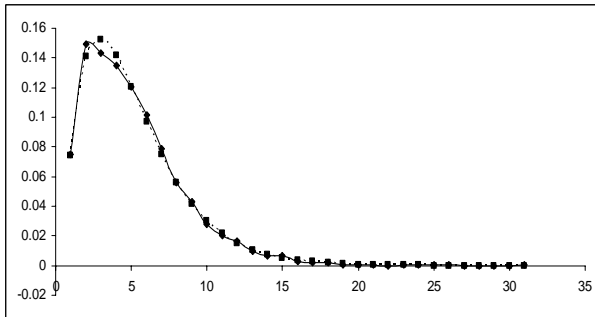
**Figure 5.** Nominal incomes evolution in the United States 1974-2004.



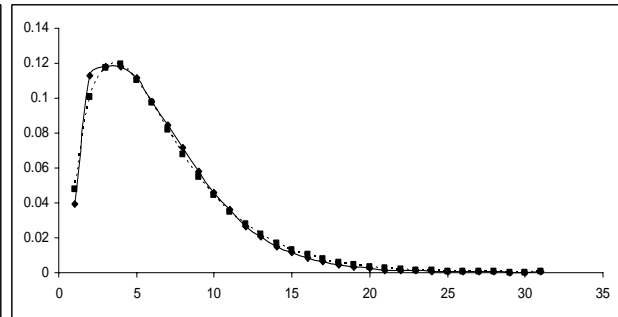
**a)** US 1974 - it. 18,  $\rho^2 = 0.995$



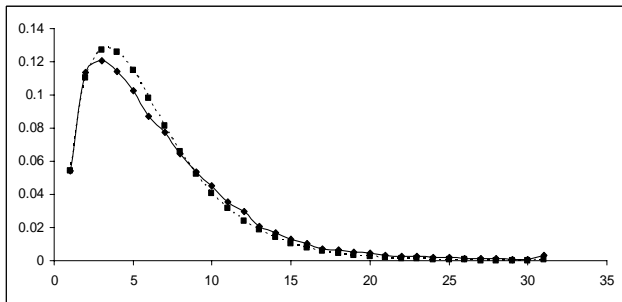
**b)** US 1979 - it. 29,  $\rho^2 = 0.996$



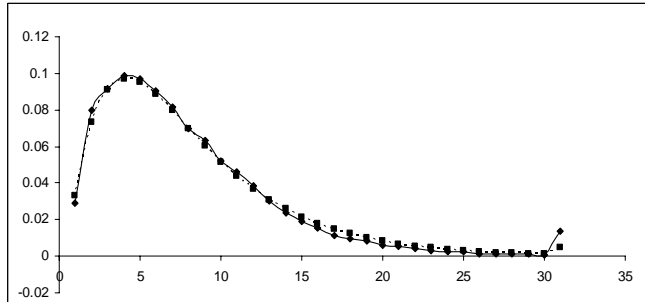
**c)** US 1986 - it. 51,  $\rho^2 = 0.997$



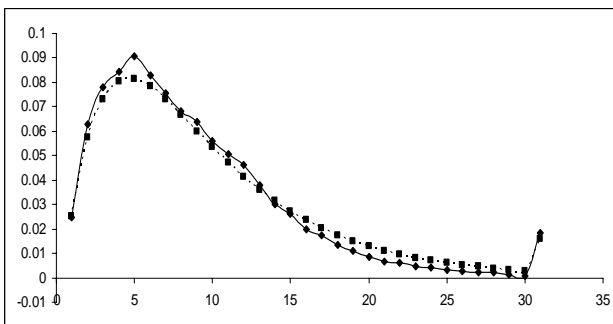
**d)** US 1991 - it. 73,  $\rho^2 = 0.995$



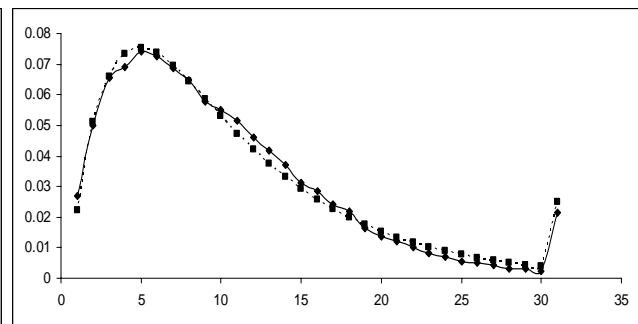
**e)** US 1994 - it. 66,  $\rho^2 = 0.994$



**f)** US 1997 - it. 101,  $\rho^2 = 0.995$



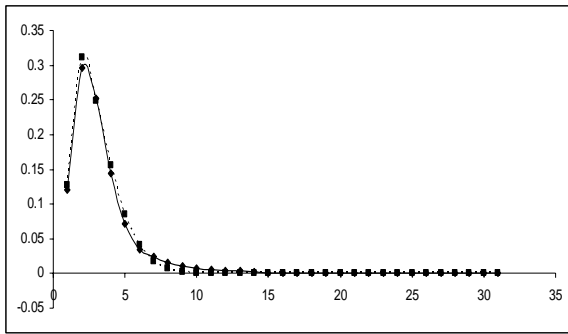
**g)** US 2000 - it. 132,  $\rho^2 = 0.996$



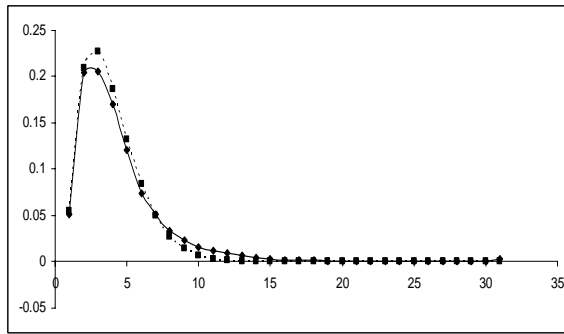
**h)** US 2004 - it. 150,  $\rho^2 = 0.990$

X axis = income classes (1-31). Y axis= Probability (theoretical, dotted line) and relative frequency (empirical, continuous line). With  $w=5$  (thousand dollars), the model parameters are  $\hat{b}=0.65$ ,  $\hat{c}=-0.64$ ,  $\hat{e}=2.16$ . Under each diagram, iteration (it.) says when the model produces the theoretical distribution that best fits the empirical one, and the  $\rho^2$  coefficient measures the fit between these two distributions. The mean  $\rho^2$  value for the all the process is 0.995. *Source:* own calculations on LIS data.

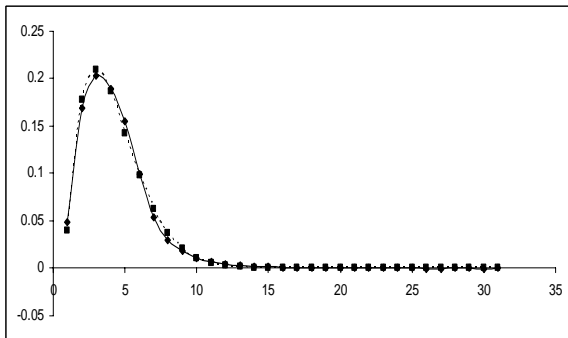
**Figure 6.** Nominal incomes evolution in France 1979-2000.



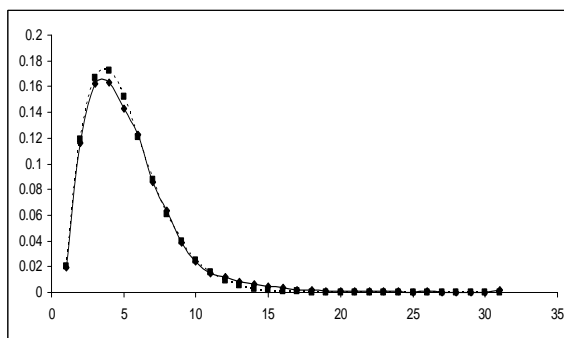
a) FR 1979 – it. 22,  $\rho^2 = 0.997$



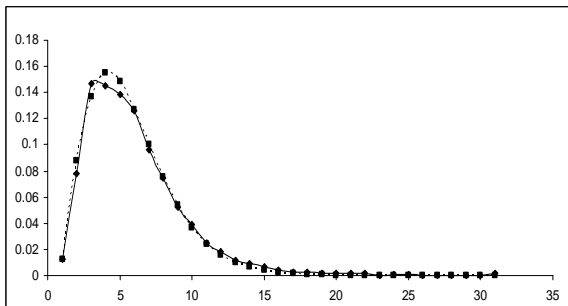
b) FR 1984 – it. 32,  $\rho^2 = 0.996$



c) FR 1989 – it. 36,  $\rho^2 = 0.996$



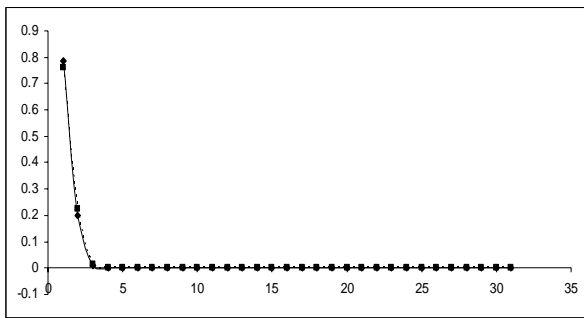
d) FR 1994 – it. 46,  $\rho^2 = 0.999$



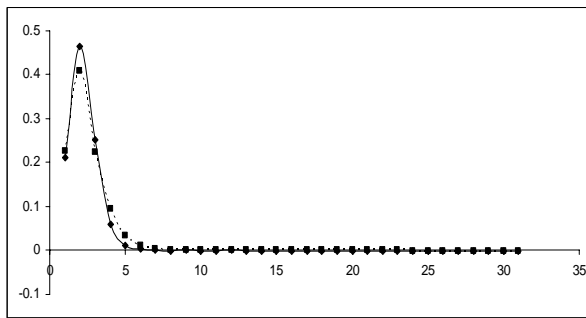
e) FR 2000 – it. 54,  $\rho^2 = 0.995$

X axis= income classes (1-31). Y axis= Probability (theoretical, dotted line) and relative frequency (empirical, continuous line). With  $w=5$  (thousand dollars), the model parameters are  $\hat{b}=0.49$ ,  $\hat{c}=-0.59$ ,  $\hat{e}=2.61$ . Under each diagram, iteration (it.) says when the model produces the theoretical distribution that best fits the empirical one, and the  $\rho^2$  coefficient measures the fit between these two distributions. The mean  $\rho^2$  value for the all process is 0.997.  
 Source: own calculations on LIS data.

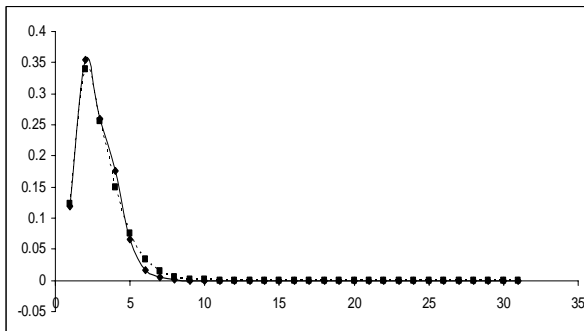
**Figure 7. Nominal incomes evolution in Sweden 1967-2000.**



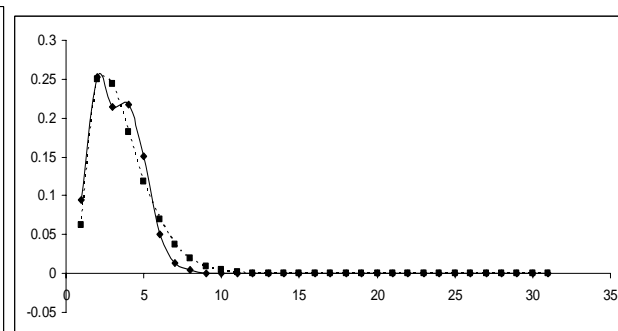
**a) SE 1967 – it. 3,  $\rho^2 = 0.999$**



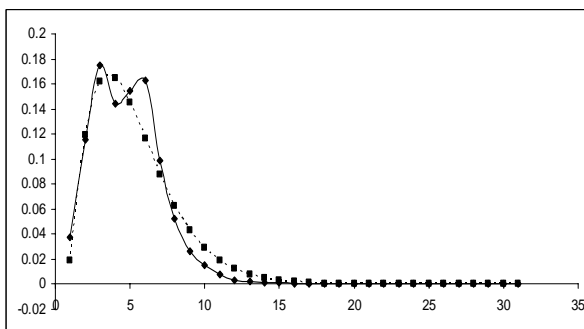
**b) SE 1975 – it. 13,  $\rho^2 = 0.986$**



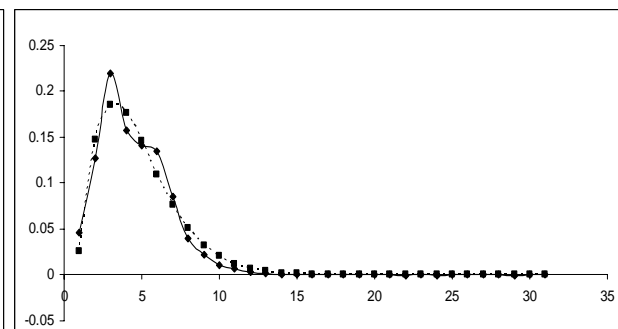
**c) SE 1981 – it. 19,  $\rho^2 = 0.994$**



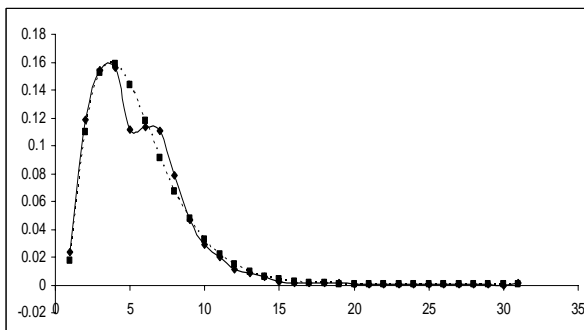
**d) SE 1987 – it. 27,  $\rho^2 = 0.967$**



**e) SE 1992 – it. 49,  $\rho^2 = 0.960$**



**f) SE 1995 – it. 42,  $\rho^2 = 0.968$**



**g) SE 2000 – it. 52,  $\rho^2 = 0.978$**

X axis= income classes (1-31). Y axis= Probability (theoretical, dotted line) and relative frequency (empirical, continuous line). With  $w=5$  (thousand dollars), the model parameters are  $\hat{b}=0.54$ ,  $\hat{c}=-0.7$ ,  $\hat{e}=2.93$ . Under each diagram, iteration (it.) says when the model produces the theoretical distribution that best fits the empirical one, and the  $\rho^2$  coefficient measures the fit between these two distributions. The mean  $\rho^2$  value for the all process is 0.979.  
 Source: own calculations on LIS data.

## 6. Parameter interpretation

The parameters of our model can be interpreted in socio-economic terms. Note, first (from table 4, columns 1 to 3, and from Fig. 8), that their empirical estimates are relatively close in all the countries. For instance, the  $c$  parameter, our *clinamen*, discussed in section 3.2, is always negative (mean = -0.67; standard deviation = 0.15). This means that income expansion tends to be greater, the lower the starting point, and better-off households tend to improve less than others<sup>13</sup>. This is not necessarily good news in redistributive terms, because a simple life cycle of earnings is consistent with our results: young households are relatively poor at the start, but they tend to improve over time. The top of one's career and earnings is typically reached shortly before retirement, after which earnings start to decrease. However, the finding is interesting and the underlying mechanism is worth closer scrutiny in future research.

The  $b$  parameter is positive (average = 0.522; standard deviation = 0.12): as discussed in footnote 8, this is not merely an artefact of our model, and this means that the variance of the income distribution in the next period is greater, the higher the starting point. The rich run higher risks of abrupt changes: both for the better and for the worse, actually, but the latter is more likely, given the negative value of  $c$  discussed before.

The  $e$  parameter has a positive mean (2.76; standard deviation = 0.42). This parameter can be interpreted, as a first approximation, as the rate of growth of nominal income. In this application, we did not separate the effect of inflation from that of real income growth. This is something which may well be worth doing in future applications, but at the price of introducing an additional parameter into the model.

One possible conclusion of our work could be as follows: countries differ in the way they let individual household incomes vary over time. These differences have long been discussed, but it is ultimately hard to state something fully convincing about them, because the process is multidimensional, and several of its aspects are poorly quantified. Our approach permits us to translate the performance of each country into three numbers (our three parameters  $b$ ,  $c$ , and  $e$ ): this makes it easier to see which country is similar to, or differs from, which other, and in what respect.

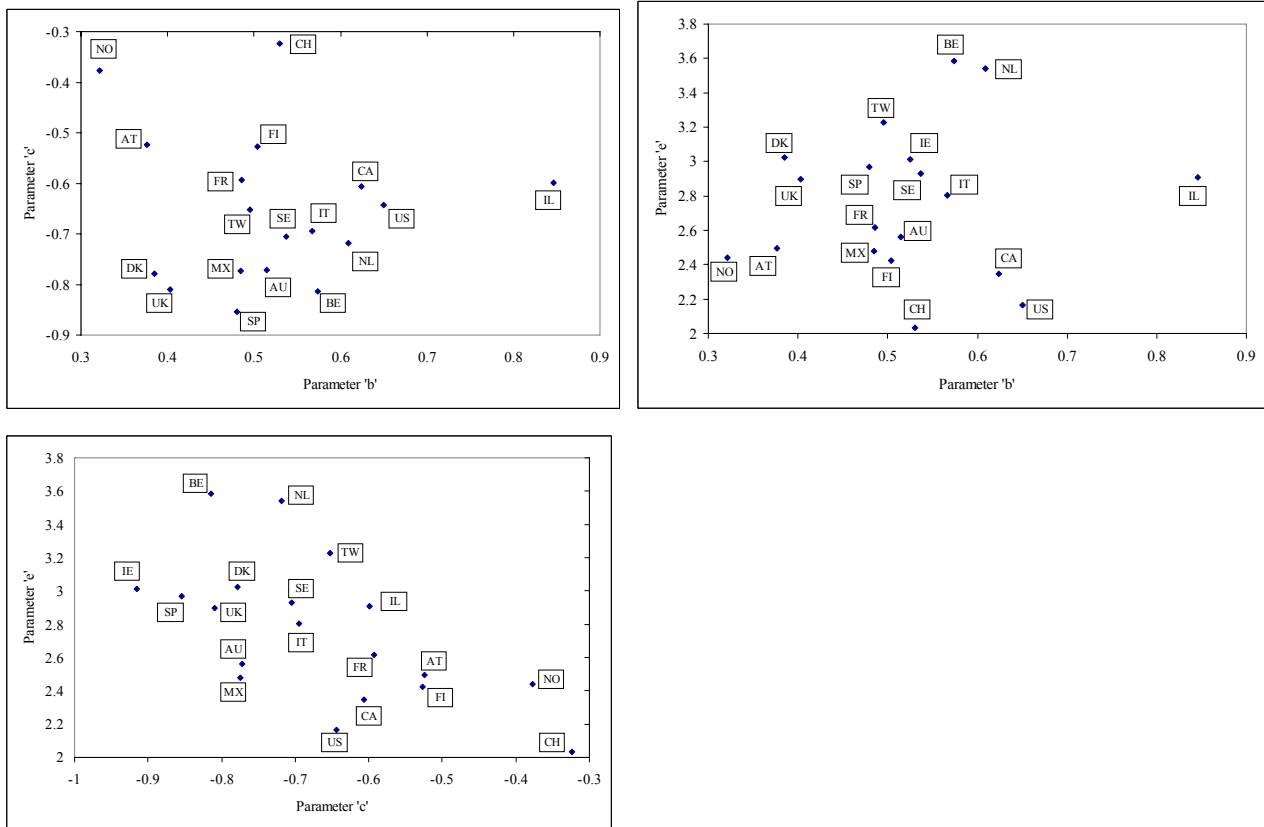
Consider for instance Fig. 8, where we plotted our 19 countries according to the value of their parameters in our model. There are expected similarities: Belgium, for instance, is always close to the Netherlands, and Canada and the United States behave similarly. But we also find that Mexico approaches Australia; that Denmark and the United Kingdom lie very close to each other, and that

---

<sup>13</sup> This phenomenon has been empirically verified for Italy and the Netherlands, using the ECHP panel data, in the years 1994-2001.

Norway and Sweden have very little in common. In short: the traditional classifications of welfare states do not emerge here.

**Figure 8.** Scatter plot of 19 LIS countries according to the estimated values of the three parameters of the model ( $b$ ,  $c$ ,  $e$ )



AT=Austria; AU=Australia; BE=Belgium; CA=Canada; CH=Switzerland; DK=Denmark; FI=Finland; FR=France; IE=Israel; IL=Ireland; IT=Italy; MX=Mexico; NL=The Netherlands; No=Norway; SE=Sweden; SP=Spain; TW=Taiwan; UK=United Kingdom; US=United States.

Source: own calculations on LIS data.

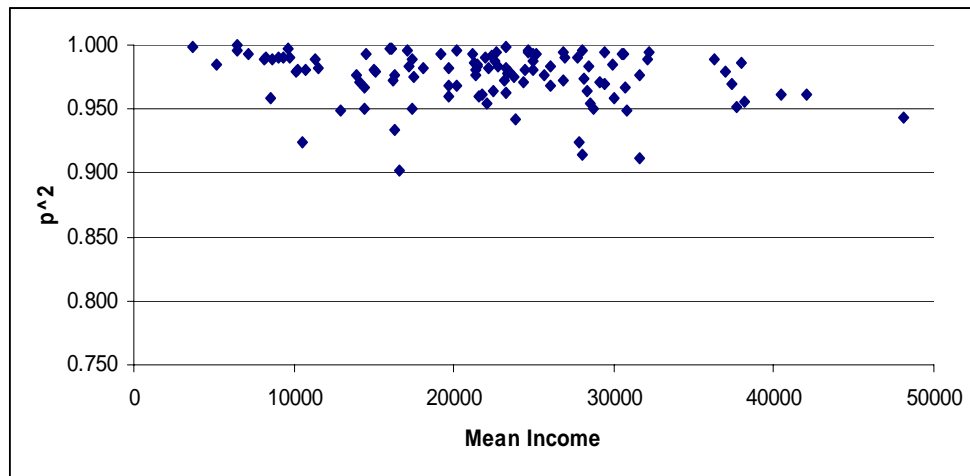
One possible explanation is that our model is unfit to describe reality. Another interpretation, though, is that, in the long run, economic systems do not work as we tend to believe: the underlying forces - at least in terms of variance (parameter  $b$ ), redistribution towards the poor (parameter  $c$ ), and nominal growth (parameter  $e$ ) - operate in a different and somewhat surprising way.

## 7. An alternative conclusion

However, a different, perhaps even more astonishing, conclusion is also possible. Shapiro's tests on the 19 values of the estimates of, separately,  $b$ ,  $c$ , and  $e$  (Table 4, columns 1 to 3), suggests that these values *could* derive from the same normal distributions. In other words, the idea cannot be rejected that "normal" values for  $b$ ,  $c$ , and  $e$  exist, and that the deviations from these values that we observe empirically (in the various countries, in the various years) are mainly due to chance.

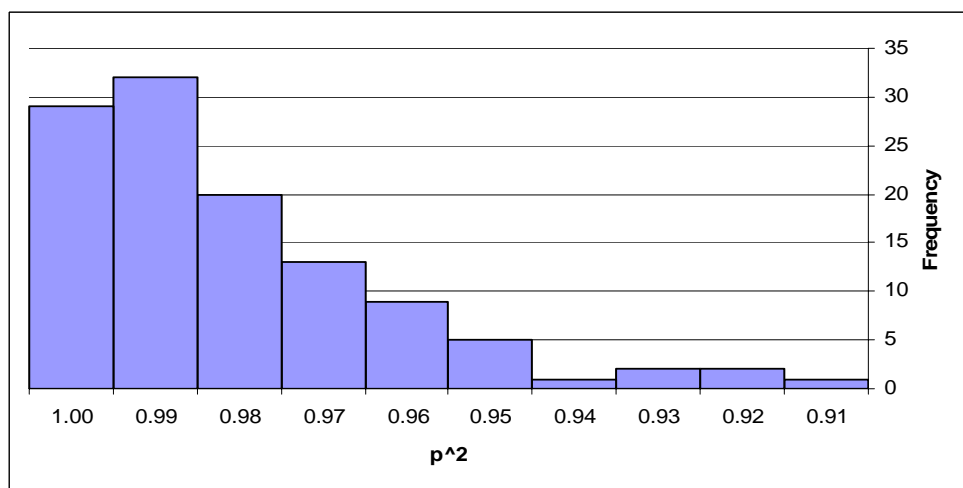
Could the process be unique? Could the "true" values of  $b$ ,  $c$ , and  $e$ , be the same everywhere, and appear different in different countries only (or at least mainly) because of chance? Let us run our model on the entire set of the 114 income distributions, as if they were the result of a single process, produced by the general structure of our model and three constant parameters  $b$ ,  $c$ , and  $e$ . Figures 9 and 10 show what happens.

**Figure 9.** Model application to all the 114 income distributions of the 19 countries



Parameters:  $w=5$  thousand dollars,  $\hat{b}=0.53$ ,  $\hat{c}=-0.65$ ,  $\hat{e}=2.58$ . Every income distributions is described by two values: 1) the mean income of the empirical distribution; 2) the  $\rho^2$  coefficient measured between the empirical and the theoretical distributions.  
 Source: own calculation on LIS data.

**Figure 10.** Frequency distribution of the  $\rho^2$  values calculated between the 114 empirical income distributions and the theoretical distribution generated by the model



Parameters:  $w=5$  thousand dollars,  $\hat{b}=0.53$ ,  $\hat{c}=-0.65$ ,  $\hat{e}=2.6$ .  
 Source: own calculations on LIS data.

The model parameters that yield the best global fit, for  $w=5$  thousand dollars, are  $\hat{b}=0.53$ ,  $\hat{c}=-0.65$ , and  $\hat{e}=2.60$ . The resulting mean  $\rho^2$  (our measure of the goodness of fit between the empirical and

the theoretical distributions) is 0.976. In 92 cases out of 114 (82%)  $\rho^2 > 0.97$ . In short: it does not seem unwarranted to state that the income evolutions of these 19 countries in about 40 years could be satisfactorily described with our model and just three constant, common parameters.

The conclusion implied by this interpretation is a very strong one. All the countries would be basically on the same path of income evolution. Household income would expand much in the same way everywhere. The only difference is that each country would be in a different phase of the process. Attempts to influence the "natural" distribution and evolution of individual incomes (e.g. to combat poverty, to foster economic growth, etc.) are either basically the same everywhere (despite some superficial differences in welfare regimes) or totally ineffective, and nature follows its course in each of the 19 countries considered.

## References

- Atkinson A.B., Bourguignon F., Morrisson C. (1992) *Empirical Studies of Earnings Mobility*, "Fundamentals of Pure and Applied Economics", Chur, Harwood Academic Publisher.
- Champernowne D.G. (1953) "A Model of Income Distribution", *The Economic Journal*, Vol. 63, No. 250, pp. 318-315.
- Cowell F.A. (2000) "Measurement of Inequality", in Atkinson A.B., Bourguignon F. (eds.) "Handbook of Income Distribution", North Holland, pp. 89-150.
- Dutta J., Sefton J.A., Weale M.R. (2001) "Income Distribution and Income Dynamics in the United Kingdom", *Journal of Applied Econometrics*, Vol. 16, No. 5, pp. 599-617.
- Esping-Andersen G. (1990) *Three worlds of welfare capitalism*, Cambridge, Polity Press.
- Esping-Andersen G. (1999) *Social foundations of post industrial economies*, Oxford/New York, Oxford University Press.
- Goodman L.A. (1961) "Statistical Methods for the Mover-Stayer Model", *Journal of the American Statistical Association*, Vol. 56, No. 296, pp. 841-868.
- Hart P.E. (1976) "The Dynamics of Earnings, 1963-1973", *The Economic Journal*, Vol. 86, No. 343, pp. 551-565.
- Labergott S. (1959) "The Shape of the Income Distribution", *The American Economic Review*, Vol. 49, No. 3, pp. 328-347.
- Locatelli M., Moscato V., Pasqua S. (2001) *The European Community Household Panel (ECHP): elements for users with special focus on labour and household economics*, "Child Working papers", 24 (<http://www.child-centre.it/>)
- Lydall H.F. (1968) "The Distribution of Employment Incomes", *Econometrica*, Vol. 27, No. 1, pp. 110-115.

- Majumder A., Chakravarty S.R. (1990) "Distribution of Personal Income: Development of a new Model and its Application to U. S. Data", *Journal of Applied Econometrics*, Vol. 5, No. 2, pp. 189-196.
- McDonald J.B., Mantrala A. (1995) "The Distribution of Personal Income: Revisited", *Journal of Applied Econometrics*, Vol. 10, No. 2, pp. 201-204.
- Neal D., Rosen S. (2000) "Theories of the Distributions of Earnings", in Atkinson A.B., Bourguignon F. (ed.) "Handbook of Income Distribution", North Holland, pp. 380-423.