

ISAE ISTITUTO DI STUDI E ANALISI ECONOMICA

**ESTIMATION OF HOUSEHOLDS INCOME
FROM BRACKETED INCOME SURVEY DATA**

by

Enrico D'Elia and Bianca Maria Martelli

ISAE

Rome

April, 2003

The Series “Documenti di Lavoro” of the *Istituto di Studi e Analisi Economica* - Institute for Studies and Economic Analyses (ISAE) hosts the preliminary results of the research projects carried out within ISAE. The diffusion of the papers is subject to the favourable opinion of an anonymous referee, whom we would like to thank. The opinions expressed are merely the Authors' own and in no way involve the ISAE responsibility.

The Series is meant for experts and policy-makers with the aim of submitting proposals and raising suggestions and criticism.

La serie “Documenti di Lavoro” dell’Istituto di Studi e Analisi Economica ospita i risultati preliminari di ricerche predisposte all’interno dell’ISAE. La diffusione delle ricerche è autorizzata previo il parere favorevole di un anonimo esperto della materia che qui si ringrazia. Le opinioni espresse nei “Documenti di Lavoro” riflettono esclusivamente il pensiero degli autori e non impegnano la responsabilità dell’Ente.

La serie è destinata agli esperti ed agli operatori di politica economica, al fine di formulare proposte e suscitare suggerimenti o critiche.

ABSTRAT

As far as data on personal income are highly confidential and sensible, it is a common practice to collect such information by asking people to classify their own earnings along a discrete scale of income “brackets”. This procedure provides an unbiased estimation of average income, under fairly general conditions, but it is well known that standard error of estimates increases with brackets size. On the other hand, people tend to underreport income, and this bias is likely to increase as brackets width gets smaller. Thus, an optimal bracket size can be generally identified, that insures a reduction of underreporting without increasing estimate variance too much. The paper presents an evaluation of brackets size effect on various procedures for estimating Italian households’ income. The first result is that the most reliable and robust procedures are those based on the extrapolation of income distribution in the upper open class by means of very simple functions. Secondly, reducing of the number of income brackets from the actual 22 to 5-7 seems to improve the accuracy of indicators for every procedure.

JEL Classification: C82, D31

Key Words: Accuracy, Bracketing, Coarse data, Households’ income, Quantification.

NON TECHNICAL SUMMARY

This paper presents an evaluation of bracket size effect on various procedures for estimating Italian households' income collected from the ISAE Consumer Survey.

Data on personal income are highly confidential and sensible, thus it is a common practice to collect such information by asking people to classify their own earnings along a discrete scale of income "brackets". Since 1982, the Institute of Studies and Economic Analysis (ISAE) carries over a survey on a stratified random sample of 2,000 consumers about their economic behaviour and expectations, in the framework of the Harmonised Project of the European Commission. Among the questions included in the ISAE questionnaire, one concerns the monthly households' income, net of personal taxes and compulsory social contributions, but inclusive of wages, salaries, self-employed earnings, profits, capital incomes, pensions, social benefits, family's transfers, etc. The information is collected over 22 brackets with particular expedients to increase the response rate.

Such data can be used both in the analysis of individual behaviour, and for macroeconomic analysis. The first approach has been developed as a generalisation of usual censored variable models, such as the classical one proposed by Hsiao (1983) and the non-parametric approach suggested by Manski and Tamer (2002). Quite the reverse, only few papers, such as Cowell and Metha (1982) and D'Elia and Martelli (2000), deal with the problem of inferring the level, or at least the dynamics, of households' income from the bracket-type survey results.

Recently Winter (2002) summarised a number of reasons why the results may be affected by brackets definition and by the way brackets are presented to the respondents, as well. Notably, households could avoid extreme brackets, simply since they signal extreme poverty or wealth. In addition, if the unfolding bracket technique is adopted, interviewed persons can be influenced by the so called "anchoring" effect, associated to the value of the initial income threshold, that determines all the following steps of the interview.

Juster and Smith (1997) acknowledged that bracketing can improve the quality of economic data, since it reduces the non-responses dramatically. In addition, it seems that wider brackets encourage people to answer to the questionnaire, since they reduce the disclosure risk of data and make it easier to answer even in case of uncertainty about the exact amount of income. Of course, the accuracy of income estimates falls with the brackets width as well.

This paper deals with the problem of the optimal choice of brackets associated with various hypotheses about the treatment of open (lower and upper) classes. An optimal bracket size is identified, that insures a reduction of underreporting without increasing estimate variance too much.

Income data are firstly grouped in different brackets sets and a series of procedures aimed to estimate average income from bracketed data are analysed. Some of them are based on a special treatment of the data falling in the upper open bracket provided in the questionnaire. Others are based on the extrapolation of income distribution within the upper class. A third group of procedures relies on the use of robust statistics computed on the observed distribution of answers. Only the methods belonging to the second group, possibly based on very simple functions, proved to be robust and reliable, even in extreme situations

To evaluate the accuracy of the income indicators obtained from the monthly survey carried over in Italy by ISAE, they are compared with the corresponding estimates of households' disposable income elaborated by the National Accounts Department of the Italian National Institute of Statistics every year, and published approximately one year after the reference period. The main result is that the correlation between the two indicators does not fall monotonically as the number of income intervals reduces (and their average size increases). Quite the reverse, such correlation raises to a maximum reducing the number of income brackets from the actual 22 to 5-9, that is for a brackets width close to 600 €, and then drops as the income classes enlarge further. In addition, this result is quite robust for any change in the treatment of the upper open income class.

It is worth noticing, however, that the results on accuracy of indicators refer solely to the estimation of average income. Reducing the number of brackets might dramatically worsen the estimation of income quantiles, as the poor results obtained by using the median estimator seem to suggest. It implies that the actual survey, based on 22 brackets, could be continued to provide reliable indicators on income distribution.

LA STIMA DEL REDDITO DELLE FAMIGLIE CON I DATI PER INTERVALLO DELL'INCHIESTA PRESSO I CONSUMATORI

SINTESI

Questo lavoro analizza l'effetto dell'ampiezza delle classi in cui sono raccolte le informazioni sul reddito delle famiglie italiane, rilevate dall'inchiesta dell'ISAE presso i consumatori, al fine della costruzione di un indicatore sintetico mensile del reddito. Dato che le informazioni sul reddito personale sono particolarmente sensibili, vengono normalmente rilevate per classi. Si evidenzia come la dimensione degli intervalli influenzi l'affidabilità delle informazioni raccolte: classi di maggiore ampiezza offrono maggiore facilità di reperimento dei dati e permettono stime corrette del valore medio, ma con elevati errori standard. Classi di minore ampiezza inducono a maggiori errori di rilevazione, dovuti a minore partecipazione e affidabilità degli intervistati. Il lavoro evidenzia come le procedure più soddisfacenti per la stima dell'indicatore sintetico siano quelle basate sul valore medio delle classi di reddito intermedie e sull'estrapolazione della distribuzione del reddito nella classe superiore per mezzo di semplici funzioni. Correlata a questa conclusione è la individuazione di un numero di classi pari a 5-7 (rispetto alle 22 originali) come migliore compromesso fra variabilità e affidabilità delle informazioni.

Classificazione JEL: C82, D31

Parole Chiave: Precisione, classi (di reddito), dati perturbati, reddito delle famiglie, quantificazione.

INDEX

1 INTRODUCTION	Pag. 7
2 ACCURACY VS. MISREPORTING REDUCTION	“ 10
3 QUANTIFYING METHODS	“ 11
3.1 Assumptions about the dynamics of $y_{N,t}$	“ 12
3.2 Solution provided by interpolating functions	“ 14
3.3 Robust statistics on income distribution	“ 17
4 THE ISAE SURVEY ON HOUSEHOLDS' INCOME	“ 19
5 THE QUALITY OF INCOME ESTIMATES	“ 21
5.1 Concordance with the National Accounts estimates	“ 22
5.2 Volatility and robustness	“ 24
CONCLUSIVE REMARKS	“ 26
APPENDIX - FIGURES AND TABLES	“ 27
REFERENCES	“ 35

1 INTRODUCTION¹

Data on personal income are highly confidential and sensible, thus it is a common practice to collect such information by asking people to classify their own earnings along a discrete scale of income “brackets”. Since 1982, the Institute of Studies and Economic Analysis (ISAE) carries over a survey on a stratified random sample of 2,000 consumers about their economic behaviour and expectations, in the framework of the Harmonised Project of the European Commission. The survey comprises fifteen questions requested by the Commission and some structural information, notably a question on income of the household. More specifically, the question included in the ISAE questionnaire concerns the monthly households’ income, net of personal taxes and compulsory social contributions, but inclusive of wages, salaries, self-employed earnings, profits, capital incomes, pensions, social benefits, family’s transfers, etc. The information is collected over 22 brackets with particular expedients to increase the response rate.

Respondents are asked to report if their income falls within a sequence of brackets, so that, at time t , the survey provides an income interval $[y_{t,i}, y_{t,i+1}]$ for each household, instead of a point estimate. Such data can be used both in the analysis of individual behaviour, and for standard macroeconomic analysis. The first approach has been developed as a generalisation of usual censored variable models, such as the classical one proposed by Hsiao (1983) and the non-parametric approach suggested by Manski and Tamer (2002). Quite the reverse, only few papers, such as Cowell and Metha (1982) and D’Elia and Martelli (2000), deal with the problem of inferring the level, or at least the dynamics, of households’ income from the bracket-type survey results.

If the income of each i^{th} household, say Y_i , lays between the lower bound $b_{L,i}$ and the upper bound $b_{U,i}$, $b_{L,i}$ underestimates the actual income of the i^{th} household, while $b_{U,i}$ overestimates it. Thus, a measure of the average income Y , suitable to be included in macroeconomic models, reads

$$\sum_{i=1}^H w_i b_{L,i} \leq Y < \sum_{i=1}^H w_i b_{U,i} \quad [1]$$

¹ The paper took great advantage from the suggestions of an anonymous referee and the participants to the 26th CIRET Conference. Of course, the views expressed in this paper are those of the authors, and do not involve any responsibility of ISAE and referees. Although the paper reports the main results of a joint research of both authors, Enrico D’Elia wrote sections 2, 3 and 5; Bianca Maria Martelli is responsible for section 4.

where w_i is the weight of the i^{th} household; that is

$$\sum_{j=1}^N f_j b_{L,j} = \sum_{i=1}^H w_i b_{L,i} \leq Y < \sum_{i=1}^H w_i b_{U,i} = \sum_{j=1}^N f_j b_{U,j} \quad [2]$$

where f_j is the observed frequency of households responding that their income falls in the j^{th} bracket; the lowest and highest bounds $b_{L,1}$ and $b_{U,N}$ are usually unknown parameters. [2] can be simplified by assuming that Y lays just in the middle point, say d_i , of the two bounds of the inequality.

Hsiao (1983) remarked that using d_i as an explanatory variable in modelling individual behaviour may bias the estimates of the parameters, but this fact does not necessarily imply that

$$Y^* = \sum_{j=1}^N f_j d_j = \frac{1}{2} \sum_{j=1}^N f_j (b_{L,j} + b_{U,j}) \quad [3]$$

is a biased estimation of aggregated income as well. The only required assumption is that the deviations between individual incomes y_i and the mid point d_j of the corresponding income class compensate within the sample, that is

$\sum_{j=1}^N \sum_{i=1}^{R_j} (y_i - d_j) = 0$, where R_j is the number of households reporting an income included in the j^{th} bracket. Notably, this requirement is much more weaker than the condition that the distribution of y_i within each bracket is such that its average turns out to be exactly d_i .

It is worth noticing that [2] and [3] do not depend on the formulation of behaviour relation under examination, but rely only on the hypothesis that it is linear with respect to income. Thus, in principle, the approximation [3] provides the researchers and market analysts with a measure of the aggregate households' income that is available timely and even monthly.

However, to make this approach operational, at least two problems should be solved. The first one is the treatment of open brackets, if any, since [3] requires the knowledge of both bounds of every bracket. The second point relates to the optimal definition of income brackets. In fact, estimator [3] is not independent from the choice of bounds, neither the results of the survey are. Recently Winter (2002) summarised a number of reasons why the results may be affected by brackets definition and by the way brackets are presented to the respondents, as well. Notably, households could avoid extreme brackets, simply since they signal extreme poverty or wealth. In addition, if the unfolding bracket technique is adopted, interviewed persons can be influenced by the so called "anchoring" effect, associated to the value of the initial income threshold, that determines all the following steps of the interview.

In any case, Juster and Smith (1997) acknowledged that bracketing can improve the quality of economic data, since it reduces the non-responses dramatically. In addition, it seems that wider brackets encourage people to answer to the questionnaire, since they reduce the disclosure risk of data and make it easier to answer even in case of uncertainty about the exact amount of income. Of course, the accuracy of income estimates falls with the brackets width as well. Thus there is a case for choosing the brackets not so narrow to discourage respondents, but not so large to raise the variance of estimates too much.

This paper deals with the problem of the optimal choice of brackets associated with various hypotheses about the treatment of open classes in [3]. In order to evaluate the accuracy of the income indicators obtained from the monthly survey carried over in Italy by ISAE, they are compared to the corresponding estimates of households' disposable income elaborated by the National Accounts Department of the Italian National Institute of Statistics every year, and published approximately one year after the reference period. The main result is that the correlation between the two indicators does not fall monotonically as the number of income intervals reduces (and their average size increases). Quite the reverse, such correlation raises to a maximum for a brackets width close to 600 €, and then drops as the income classes enlarge further. In addition, this result is quite robust for any change in the treatment of the upper open income class.

The remainder of the paper is organised as follows. The next paragraph provides a framework for taking into account the contrasting effects of brackets reshaping on the accuracy of income estimates. Basically, it derives the conditions for the variance induced by widening income classes being partially compensated by the related reduction of underreporting or, more generally, misreporting. Section 3 discusses several procedures for mapping survey results into a quantitative index of household's income. Methods are based either on some treatment of the open upper bracket, or on robust statistics derived from the observed frequency distribution of households' responses. The fourth section describes the actual framework of ISAE survey. The fifth paragraph summarises the results of an application of various quantification procedures to Italian data. First of all, the concordance between survey indicators and National Accounts estimates is analysed for different bracketing schemes. In addition the volatility of monthly indicators is considered as a further possible selection criteria. Some conclusive remarks and suggestions for further improvements of the ISAE survey close the paper.

2 ACCURACY VS. MISREPORTING REDUCTION

Data collected by using bracketing do not necessary provide biased estimates of underlying quantities. As matter of facts, bracketing is a special case of rounding or, more generally, of coarsening. Heitjan and Rubin (1991) demonstrated that it is generally possible to treat coarsened data as if they were simple grouped data, ignoring the special nature of data both in standard likelihood and in Bayesian inference, unless bracketing itself induces a bias in respondents' behaviour. Notably, the latter case can occur when households avoid choosing extreme brackets, since they signal extreme positions, and when "anchoring" effect is huge.² Otherwise, it can be assumed that bracketing provides an unbiased estimation of average income.

Even if this is the case, the standard error of estimates increases with brackets size. In the special case of equally spaced brackets of width 2δ , it is easy to show that estimated average has the standard error

$$\sigma = \frac{\delta}{\sqrt{3}} \quad [4]$$

(see Dempster and Rubin, 1983).

On the other hand, people tend to underreport income, and this bias is likely to increase as brackets width gets smaller. As a limiting case, underreporting is approximately null for a two bracket question, asking whether the interviewed person earns more than a given amount of money. Notably, assuming, for sake of simplicity, that each household wish to underreport its own income by H €, and the average bracket width is $W > H$, then only a fraction of people $H/W < 1$ succeeds in giving a wrong answer to the questionnaire.

More generally, bracketing helps in reducing non-ignorable non-responses and random misreporting of income, specially associated to casual earnings and expenditures (e.g.: back pay, some taxes, etc.). Of course, larger brackets reduce the risk of such errors. Regardless to the causes and systematicity of respondents' misreporting (u), it can be assumed that it is a non increasing function of δ , and thus overall mean square error of estimated income (mse) is

$$mse = \frac{\delta}{\sqrt{3}} + u(\delta) \quad [5]$$

² Tourangeau, R., L. J. Rips, and K. Rasinski (2000) review the psychological literature on the bracketing effect and anchoring.

Of course, mse achieves its minimum for

$$\frac{du}{d\delta} = -\frac{1}{\sqrt{3}} \quad [6]$$

As a consequence, an optimal bracket size can be generally identified, that insures a reduction of underreporting without increasing estimate variance too much. Figure 1 illustrates how this optimal size can be graphically detected for a suitable shape of $u(\delta)$. The latter is supposed to be concave, in the relevant region at least, in order to have $\lim_{\delta \rightarrow \infty} u(\delta) > 0$, however $u(\delta)$ might be convex for small brackets sizes as well. In the latter case, respondents do not change their propensity to lie too much, since they are likely to feel that narrow brackets width does not protect their privacy enough. In both cases, equation [6] implies that optimal size δ_{min} turns out to be non null if

$$\frac{du}{d\delta} < -\frac{1}{\sqrt{3}} \quad [7]$$

for some δ . Otherwise the optimal δ is simply zero, and a conventional direct survey overwhelms a bracketed one. However, the condition [7] is not very binding: for instance, it requires that widening brackets size by 1% of average income, misreporting falls more than 0.6% of average income.

Equation [6] has some interesting consequences for data collection. First of all, where misreporting (or general voluntary bias in answers) is a severe problem, bracketing is a feasible strategy in formulating questions. In addition, sometimes it is possible to choose brackets bounds in such a way that respondents are induced to be fair, because their own true position fall far from the bounds. For instance, if it is known that incomes concentrates around some discrete levels (e.g.: 500, 1,000 and 1,500 euro per month) brackets bounds should be centred around those values, so that people willing to shadow their actual income may choose the right bracket as well.

Secondly, data collected by using narrow brackets could be less reliable than those collected and elaborated by using larger ones. This consequence has been put under test in Section 5.

3 QUANTIFYING METHODS

There are two main ways to evaluate income dynamics from bracketed data: the first one tends to estimate the average of income distribution, the second

exploits position indexes of the same distribution, such as median and mid-mean. Following the first approach, let be:

$$\bar{y}_t = \int_{y_{0,t}}^{y_{N,t}} f(y_t) y_t dy_t \quad [8]$$

the average income at time t computed from the *true* frequency distribution of incomes, where $y_{0,t}$ and $y_{N,t}$ are the lower and higher income levels, and $f(y_t)$ is the inherent frequency distribution. Where the discrete N class distribution of incomes is observed in place of the continuous one, only the approximation

$$\hat{y}_t = 1/2 \sum_{i=2}^{N-1} (y_{i,t} + y_{i-1,t}) f_{i,t} + 1/2 (y_{1,t} + y_{0,t}) f_{1,t} + 1/2 (y_{N,t} + y_{N-1,t}) f_{N,t} \quad [9]$$

can be computed; where $f_{i,t}$ are the observed frequencies; $y_{i,t}$ for $i=2, N-1$, are the known limits of income classes, while the extrema $y_{0,t}$ and $y_{N,t}$ are generally unknown. The approximated formula [9] matches [3], and relies on the middle-point Cauchy theorem. Thus the estimate \hat{y}_t can be biased if the $f(y)$ function is not linear between each specified couple of class bounds. Anyway, this bias should be negligible if the classes are narrow enough and non-linearities tend to compensate over the income brackets as a whole. The latter assumption is quite reasonable for a frequency distribution that alternates peaks and throats, so that the function is not allowed to be concave or convex anywhere, apart from particular cases (i.e.: an income distribution highly concentrated toward the lowest or highest incomes).

The estimate \hat{y}_t turns out to depend on two unknown parameters (i.e.: $y_{0,t}$ and $y_{N,t}$). Thus some hypothesis is needed to proceed. In what follows we concentrate on the assumptions regarding the upper bound only, since empirical evidence shows that the value assigned to the lowest bound $y_{0,t}$ is not so relevant.

3.1 Assumptions about the dynamics of $y_{N,t}$

The simplest assumption about $y_{N,t}$ is that it is fixed over time or, more realistically, that it varies with price level p_t (Brandolini and Parigi, 1993). According to this latter assumption, [9] reads

$$\hat{y}_t^P = B_t + 1/2 (y_{N,0} p_t + y_{N-1,t}) f_{N,t} \quad [10]$$

where B_t is the observable part of [9], and corresponds to the first two member of the right hand side of [9], under the hypothesis that $y_{0,t} = 0$; $y_{N,0}$ is a suitable fixed value, set on the basis of some information about the true distribution of households' income (e.g.: derived from a standard households' budget survey). Whereas some independent evaluation of y_t is available (for instance from annual or quarterly National Accounts), say Y_t , it is possible to get an estimation of $y_{N,0}$ regressing Y_t on the terms: $(B_t + 1/2 y_{N-1,t} f_{N,t})$ and $1/2 p_t f_{N,t}$. The ratio between the last and the first estimated coefficients is a guess of the unknown parameter.

The solution [10] has the advantage to be very simple, but it implies a strong assumption about the neutrality of inflation on the dynamics of highest incomes. Quite the reverse, economic theory states that highest incomes are more volatile compared with the others, because richest people revenues are likely to depend more heavily on capital incomes and profits. Thus, every cyclical variation in productivity and mark-up affects the level of highest incomes stronger than prices. As a result, an index based on the previous hypothesis tends to smooth the true dynamics of income (i.e.: it overestimates lower incomes and underestimates richest people earnings).

An other assumption about the unknown parameters of [9] states that the width of the upper open bracket ($y_{N,t} - y_{N-1,t}$) is proportional to the average level of households' income \hat{y}_t , that is

$$\hat{y}_t^Y = B_t + 1/2 (\alpha_N \hat{y}_t^Y + y_{N-1,t}) f_{N,t} \quad [11]$$

The underling hypothesis is that the structure of relative income differentials is constant over time, and households move from one point of the distribution to the other, changing the observed frequencies $f_{i,t}$, but not the underlying relative income gap between poor and wealthy people. This conjecture is subject to a criticism similar to the assumption on price-varying bounds. Anyway the link established between average income and unknown upper bound could take into account the productivity and inflation dynamics effect at least (i.e.: as far as productivity and/or prices increases the average income bounds vary with productivity and/or prices, but not with mark-up and profit share).

According with the average-income-varying bounds assumption, an estimate of average income, derived by rearranging the terms of [11], is:

$$\hat{y}_t^Y = \frac{B_t + 1/2 y_{N-1,t} f_{N,t}}{1 - 1/2 \alpha_N f_{N,t}} \quad [12]$$

The term $(1 - 1/2 \alpha_N f_{N,t})$ in [12] plays the role of a time varying scale factor for the observable part $(B_t + 1/2 y_{N-1,t} f_{N,t})$ of the estimator. To the aim of having a positive estimate of average income, the additional constraint $\alpha_N < \frac{2}{f_{N,t}}$ must hold. As far as income grows, $f_{N,t}$ increases over time, but the latter constraint becomes binding only if bound $y_{N-1,t}$ is too low, so that the proportion of households falling in the upper open bracket is very high. In any case, given α_N , the scaling factor tends to increase over time, inflating the dynamics of the index \hat{y}_t^Y , all other things being equal.

3.2 Solution provided by interpolating functions

Another way to get an estimate of the crucial term $(y_{N,t} + y_{N-1,t})f_{N,t}$ in [9] relies on the assumption that the frequency of incomes in the right tail of the frequency distribution follows some known, non-increasing, possibly time varying, function $g_t(y)$ of income in the interval $[y_{N-1,t}, y_{N,t}]$. It is not necessary that $g_t(y)$ is chosen among the theoretical distribution functions, however it shares some of the properties with such functions, since it is required that

$$f_{N,t} = \int_{y_{N-1,t}}^{y_{N,t}} g_t(y) dy \quad [13]$$

$$g_t(y) \geq 0, \text{ for } y_{N-1,t} \leq y \leq y_{N,t} \quad [14]$$

$$g_t(y) = 0, \text{ for } y \geq y_{N,t} \text{ if } y_{N,t} \text{ is finite} \quad [15]$$

$$0 < f_{N,t} m_{N,t} < \infty, \quad \text{where} \quad m_{N,t} = \frac{1}{f_{N,t}} \int_{y_{N-1,t}}^{y_{N,t}} g_t(y) y dy \quad [16]$$

In addition, it is possible to restrict further the set of admissible functions by assuming that

$$g_t'(y) \leq 0, \text{ for } y_{N-1,t} \leq y \leq y_{N,t} \quad [17]$$

and

$$g_t\left(\frac{y_{N-1,t} + y_{N-2,t}}{2}\right) = \frac{f_{N-1,t}}{y_{N-1,t} - y_{N-2,t}} \quad [18]$$

Condition [17] follows from the traditional economic theory about personal income distribution, establishing that the number of households earning the net income y does not increase with y , at least among the richest households. Condition [18] derives from two simplifying (even if reasonable) assumptions. The first one requires that in the middle point of the $N-1$ -th bracket the interpolating function assumes exactly the value of the mean value of the density function of income distribution between $y_{N-2,t}$ and $y_{N-1,t}$, that is $\frac{f_{N-1,t}}{y_{N-1,t} - y_{N-2,t}}$. This

is true for every piecewise linear frequency distribution functions, and for every linear approximation of general functions as well, provided that they are monotonic between $y_{N-2,t}$ and $y_{N-1,t}$. The second assumption requires that the frequency distribution of income is continuous passing from the next to last bracket to the last one. As far as the aforementioned approximation of $g(y)$ holds in the middle point of the interval $y_{N-2,t}$ and $y_{N-1,t}$, condition [18] follows.

Of course conditions [13] - [18] do not allow to identify univocally either the functional form and parameters of $g_t(y)$, or the upper bound $y_{N,t}$. However, once the functional form of $g_t(y)$ has been chosen (complying to [14]-[17]), conditions [13], [15] and [18] provides a set of equations that usually determine either three parameters of $g_t(y)$, assuming that $y_{N,t} = \infty$; or two parameters plus the finite value for the upper bound $y_{N,t}$.

In principle, the problem of evaluating the term $(y_{N,t} + y_{N-1,t})f_{N,t}$ in [9] could be solved also by interpolating the known part of the income distribution by means of some theoretical function. Notably, this is the solution proposed by Kakwani (1976) and Cowell and Metha (1982), in order to deal with the estimation of inequality measures from data published as grouped data. Of course, closed brackets provides a set of (highly non-linear) restrictions suitable to determine up to $N-1$ parameters of $g_t(y, \theta)$, whose general form reads

$$\int_{y_{j-1,t}}^{y_{j,t}} g_t(y, \theta) dy = f_{j,t} \quad [19]$$

where θ is a vector of unknown parameters.

The parameters estimated empirically from the available data can be used to compute the unknown term of [9]. Visco (1984) provides an exemplum devised for the estimation of price expectations from a tendency survey (that are supposed normally distributed). However, the empirical income distribution is usually far more complicated than the one proposed in the economic literature. Overall incomes frequency seem to be the sum of several frequency distributions, each related to some special socio-economic group of households. As a result, the final empirical distribution of income is usually everything but unimodal, and thus usual theoretical functional forms hardly fit the actual data, especially in the tails. Therefore the estimation of $(y_{N,t} + y_{N-1,t})f_{N,t}$ may be very poor.

Many hypotheses about $g_t(y)$ satisfy conditions [13] - [18]. Table 1 provides a list of candidates, along with the related estimators of the average income within

the upper bracket, that is $m_{N,t} = \frac{1}{f_{N,t}} \int_{y_{N-1,t}}^{y_{N,t}} g_t(y) y dy$. Of course, the parameters of the

extrapolating functions are those consistent with [13], [15] and [18].³ The first two functions are the uniform and the triangular density functions. They are definite only within a finite interval of income values, that is they imply a finite upper bound for every income. Formally, this bound can be expressed as a function of h_t , $f_{N-1,t}$, and $f_{N,t}$, but it should not be interpreted necessarily as the top income within the sample of households at time t . However, the two functions can be also regarded as simple local interpolators of the true income distribution in the neighbourhood of $m_{N,t}$. In other words, the shape of the two first functions is optimised to fit the data from the beginning to the middle of the last bracket, but not necessarily in the right upper tail.

³ For instance, in the case of triangular distribution, for each period, $g(y) = a-by$, thus the conditions $g(y) = 0$, for $y = y_N$ and $h^* = g\left(\frac{y_{N-1,t} + y_{N-2,t}}{2}\right) = \frac{f_{N-1,t}}{y_{N-1,t} - y_{N-2,t}}$ imply that

$$a = h^* \frac{y_N}{(y_N - y_{N-1}) + \frac{1}{2}(y_{N-1} - y_{N-2})} \text{ and } b = \frac{h^*}{(y_N - y_{N-1}) + \frac{1}{2}(y_{N-1} - y_{N-2})}. \text{ Thus, } y_N \text{ derives from the condition}$$

$$f_N = \int_{y_{N-1}}^{y_N} g(y) dy = \frac{1}{2} b (y_N - y_{N-1})^2. \text{ It is easy to demonstrate that } y_N = y_{N-1} + \frac{f_N + \sqrt{f_N(f_N + f_{N-1})}}{h^*}.$$

The details of computation for the remaining cases are available from the authors on request.

Other two interpolating functions are derived from the exponential and the Pareto distribution respectively. The latter has been widely used in the analysis of income distribution (see Atkinson and Bourguignon, 1998). It assumes that, in each given point of time, the number of households decreases with the inverse of their income raised to the b^{th} power. That is, passing from the income level E to E plus $l\%$, one expects to find in the sample approximately $b\%$ households less than those earning E . Conversely, the exponential distribution says that the logarithm of the number of households in the sample falls linearly with their income. Even if the two hypotheses seem to differ very much, there are only minor differences in the upper tail of the distribution.

More generally, it is easy to demonstrate that all the four distributions listed in Table 1 provide an estimate of $m_{N,t}$ that share the general form

$$m_{N,t} = y_{N-1} + \alpha \frac{f_{N,t}}{h_t} \quad [20]$$

where $h_t = g(y_{N-1})$, that is the value of the extrapolating function $g(y)$ in the lower bound of the last bracket. When $\alpha = 1/2$ [20] gives the solution associated with the rectangular distribution; if $\alpha = 2/3$ it provides the solution consistent with the hypothesis that the income distribution is triangular in the upper tail; when $\alpha = 1$ [20] corresponds to the assumption that incomes follows the exponential distribution and, under the further (mild) constraint that $f_{N,t}$ is small compared to y_{N-1} , also the Pareto distribution.

According to equation [20], the unknown average income within the highest bracket rises with the percentage of households classifying themselves in the upper income class, but decreases as the number of households falling in the $N-1$ th bracket increases, since h_t is proportional to $f_{N-1,t}$. As a consequence, income indexes provided by interpolating function listed in Table 1 are very sensible to the dynamics of the two upper income brackets. More specifically, a shift in the share of upper class incomes has a more than proportionate effect on average income, while a variation in $f_{N-1,t}$ has an adverse effect.

3.3 Robust statistics on income distribution

In the previous section, the income indicator has been assumed to track the average of households' income in each point of time. An alternative approach to the estimation of income dynamics is based on the use of some "robust"

statistics on income distribution, such as median or mid-mean, instead of the mean.⁴

Notably, a quick estimate of median can be achieved, by assuming that the distribution of incomes within the M^{th} class for which $\sum_{j=1}^M f_{j,t} < \frac{1}{2} < \sum_{j=1}^{M+1} f_{j,t}$ is flat. In that case it reads

$$y_t^M = y_{M,t} + \frac{\frac{1}{2} - \sum_{j=1}^M f_{j,t}}{f_{M+1,t}} (y_{M+1,t} - y_{M,t}) \quad [21]$$

A geometrical demonstration of equation [21] is provided in fig. 1. Given that the cumulative frequency of income lower or equal to $y_{M,t}$ is $F_t = \sum_{i=1}^M f_{i,t} < \frac{1}{2}$, by hypothesis, than the frequency of income included in the interval $[y_{M,t}, y_t^M]$ is $(\frac{1}{2} - F_t)$. If the distribution of income in the $M+1^{\text{th}}$ class is rectangular, it follows that the density function of income inside that class is $h_t = \frac{f_{M+1,t}}{(y_{M+1,t} - y_{M,t})}$. Thus, the

Value d_t such that $d_t h_t = (\frac{1}{2} - F_t)$ is given by $d_t = \frac{\frac{1}{2} - F_t}{h_t}$. Equation [21] follows.

Adopting a similar hypothesis it is easy to get an estimation of the mid-mean y_t^Q , that is the simple average of 25th and 75th centiles of income distribution. Similar procedures share the advantage that they do not depend on the estimation of income bounds and, in principle, could turn out to be more robust than the average based methods whereas the income distribution is multimodal, as it is usually the case. The conjecture about the distribution of income within the median and quartile classes is usually not very binding, and any other assumption requires strong (and hard to be verified) hypothesis about the distribution of incomes.

A drawback of procedures based on median and mid-mean estimated from bracketed data is that the approximation [21] implies a discontinuity in the neighbourhood of each class bound. Therefore, if the estimated median (or a quartile) falls near one of these bounds, a slight variation of observed frequencies may produce a large variation in the index. In addition, the computation of quantiles is allowed to be estimated only from closed classes, for

⁴ See Hoaglin, Mosteller and Tukey (1983) for a discussion of robust statistics.

which the parameter h_t is easily evaluated. Thus, if a frequency higher than 0.5 (or 0.25 for mid-mean) is attached to one of the extreme classes, some artefact is needed (for instance h_t could be estimated from the nearest class). This constraint is quite unimportant for median estimation, but could be binding for mid-mean.

4 THE ISAE SURVEY ON HOUSEHOLDS' INCOME

Since 1982,⁵ within the European Harmonised Survey Programme, ISAE carries out a monthly Consumer Survey on a representative sample of 2,000 respondents.

The sample is random, built proportionally to the full-aged people universe and stratified by six main geographical areas and seven demographic classes of municipalities. It is a two stage sample, the first one being the telephonic number selected, the second being the consumer selected within the contacted household. It is also a quota sample, since the proportion between men and women has been set to the universe one (respectively 48% and 52%).⁶ The survey is carried out using telephonic interviews and with the aid of a Computer Aided Telephonic Interviewing (CATI) software.

The ISAE Consumer Survey comprises the harmonised fifteen qualitative questions characterised by three-to-five reply options based on three main topics: notably, opinions on the overall situation, opinions on the households' situations, plans to purchase durable goods, cars or homes. Beside a set of structural questions ISAE asks information on income.

As this kind of information is highly sensitive and often implies a high non-response rate, ISAE adopts two main specific criteria to face this difficulty.

First, as usual in these kinds of problems, data are collected asking the interviewed persons not to give a punctual information, but to choose between a

⁵ Properly, ISAE Survey started in 1973, on a four-monthly basis.

⁶ The sample design and survey features changed over the years. Since 1982, the main innovations have been the following. In 1995 the survey changed from a household to a consumer survey, that is the sample unit passed from the "breadwinner" to the "consumer" intended as a full-aged person belonging to the household. Also in 1995 the interview technique became telephonic, while previously the survey was carried out by face-to-face interviews. In 1998 the sample design has been updated diminishing the stages (from 3 to 2) and therefore increasing the precision of the sample. For details see Martelli (1998).

predefined range of 22 brackets. The consumer, however, is not aware of this, being asked according to an unfolding brackets sequence. He or she is firstly asked to choose between two main ranges (e.g. more or less a predetermined average earning), than to choose between two further sub-ranges of the previous chosen bracket, and so on until the final bracket is identified apparently avoiding to ask too specific information.

A second device to induce a favourable psychological attitude in the consumer is asking two specific questions on income in a strictly defined order. The first one is asked about in the middle of the interview, when the interviewer has already established a favourable contact with the consumer. The wording of the first questions is: “How much money a household like yours would need to live without luxuries, but having all the necessary available?”. This question represents an interesting tool to investigate the issues of relative poverty and consumers’ satisfaction. In addition, it is important for better collecting the information of the second question about actual income. It is intended that consumers are more willing to answer to an indirect question like this. Having answered to the question about necessary income is likely to lower the concern of respondents and, in addition, it creates a more relaxed climate when the most sensitive question on actual income is asked, at the end of the interview. In this way it is possible to cut the non-response rate.⁷ This core question on actual net disposable income of the household represents the source of the data analysed in this paper.

The household’s net disposable income amount is requested in an extended and general formulation comprising wages, salaries, self-employed earnings, profits, capital incomes, pensions, social benefits, family transfers, etc. However, it is very likely that the consumers tend to underreport this amount. First of all, consumers are obviously reluctant to declare their irregular activities, if any. Secondly, they usually try to hide extreme positions, notably very low or high incomes. In addition, often the consumer has not a clear opinion about the monthly quotas of income due to profits, capital interests and so on that do not have a monthly periodicity. Thus, the answers could mainly approximate the monthly perceived wages and salaries, rather than overall income. Other surveys ask more detailed information about family’s revenues to face this problem,⁸ but the ISAE survey aims to obtain quick information at a high frequency (monthly)

⁷ The average non-response rate ranges about to 10-12% in the last ten years.

⁸ For example, in Italy, the Central Bank and the National Statistical Institute carry out surveys which collect more detailed information on household incomes, but with a lower frequency.

and is therefore bounded in formulating plain questions. The general formulation could, on the other side, lower the non-response rate.

The number of income brackets adopted by ISAE is wider than the one recommended by the Commission, that is 22 classes instead of four (see Table 2), so that the information tend to be almost punctual. The classes range from the first one (up to about 500€) to the upper one comprising incomes over 3100€. ⁹ Further, as brackets have not been revised since 1990, ¹⁰ there is a finer detail in the lower ones. The frequency distribution, however, is still satisfactory, giving only about 4% of answers falling in the higher class.

A further problem arising from the use of a broad classification in 22 brackets is due to a “heaping effect”:¹¹ that is the presence of frequency peaks corresponding to few main brackets. The respondents tend to approximate the answers and to round the declared amount of income to the nearest 500,000 Italian Lira (250€) class. Notably, Figure 3 shows that many answer concentrate around 2,000,000 (1000€) and 2,500,000 Lira (1300€).

5 THE QUALITY OF INCOME ESTIMATES

By using the monthly data described in section 4, an experiment has been carried out with a twofold aim. The first one is to evaluate the reliability of various procedures in providing a timely estimate of households income, consistent with the final National Accounts estimates, usually published one year after the reference period. The second aim is to test empirically the result presented in section 2, where it is suggested that the reliability of income estimates may unexpectedly improve by widening the brackets used in the questionnaire.

In order to simulate the effect of adopting different (wider) brackets, the data collected by using the original brackets have been aggregated, and the different procedures have been applied to such derived data. In doing so, it has been implicitly assumed that “anchoring”, “heaping”, and other bracket effects on the respondents can be neglected. That is, it is assumed that the brackets’ bounds did not influence the way the households would have answered. In other words,

⁹ As the classes are defined in Italian liras the corresponding amount in € is approximated.

¹⁰ Until 1989 income was grouped in 7 classes.

¹¹ See, for a discussion this feature on the unemployment duration data in Italy, Torelli et al. (1993).

it could be argued that carrying out the survey by using the different bracketing schema analysed here would have given different results. Indeed this is a very strong assumption (see section 1), however it should be argued that this further restriction is unfavourable to the assumption under examination: that misreporting reduces as the brackets width widen. As a consequence, tests based on the simple aggregation of data collected by means of the original bracketing are likely to be unduly severe.

Of course, the original income groups used in the ISAE questionnaire can be aggregated in many ways. The aggregation schema presented in the following sections is only one of about 20 tested during the research programme. It is consistent with the main aims of this paper: that are to compare quantification procedures, and to test the effects of changing the average brackets width. First of all, the upper bound of the next last brackets has been let unchanged, in order to improve the comparability among results, since fixing y_{N-1} is crucial for most of the quantification procedures. Secondly, the number of classes has been progressively reduced to 15, 11, 9, 7, 5, 4, and 3, as reported in Table 2, trying to include in each bracket approximately the same number of households, as can be verified by looking at the average percentage of respondents falling in each of the original brackets over the past decade, reported in the first column of the table. Unfortunately, the actual distribution of incomes, and the size of original brackets, allowed achieving this target only in part. However, the derived closed classes have been defined so that their average width increases approximately by one third from one group to the following, passing from 141€ in the original bracketing to 1,033€ in the 3-class simulation. Thus, the range of the derived brackets seems wide enough to test if the result showed in section 2 holds. Anyway, the results do not change too much by using different aggregation schemas.¹²

5.1 Concordance with the National Accounts estimates

Although the definition of income provided in the questionnaire of ISAE is necessarily more fuzzy than the one formally stated in the SEC 95, the estimates published by the Italian National Accounts Department represent a reasonable benchmark for the results of the survey carried out by ISAE. In addition, a reference series is needed in order to estimate the parameter α_N in [11]. Notably, National Accounts measure the sum of Italian households' income, while ISAE refers to the monthly average income of each household. As a consequence,

¹² Of course, detailed results for several aggregations of original brackets are available from the authors on request.

each quantification of the ISAE survey has been multiplied by the average number of families reported yearly by the Population Registers, to make the indexes comparable to the National Accounts estimates.¹³

Anyway, regarding the National Accounts as the most accurate measure of the households disposable income, the quantification procedures applied to ISAE survey can be considered as reliable as the higher is the correlation between the time series of quantified indexes and the corresponding National Account data. Table 3 summarises the correlation coefficients computed for the eight combinations of brackets detailed in Table 2 and the eight procedures for the quantification of survey results. The correlation has been reported both between the levels of indexes vs. the corresponding National Accounts aggregate, and between the annual percentage changes of the same time series.¹⁴ Correlations have been estimated from simple linear regressions (including a constant term) run on a sample ranging from 1985 to 2001.

First of all, it should be noted that there are only minor differences among the correlations between each couple of time series. Unexpectedly, also the procedure based on [12] does not show any strong advantage compared to the others, even if it implies the estimation of the parameter (α_N), that potentially maximises the correlation between the index and the National Accounts data. Updating y_N by using also information on the dynamics consumer prices, according to [10], does not provide better results than other procedures, as well.

The R^2 statistics ranges between a minimum of 0.93, attained for the median computed on 22 brackets, to 0.97, reached by applying various procedures to 5, 7 and 9 class bracketing. On average, the relation between brackets width and R^2 is the one reported in Figure 4. As predicted in section 2 and in Figure 1, the best approximation of National Accounts estimates of disposable income is not the one obtained from the original 22 class data. As a matter of facts, by widening the brackets to 450-600€ almost every quantification procedure provides more accurate estimates. Thus, the optimal size of brackets seems to be

¹³ The aim of the experiment is simply to verify the correlation between the dynamics of income indices and disposable income estimated in national accounts. Thus, the possible integration of ISAE data for the so-called 13th month wage of employees is not relevant as far as it implies a simple multiplication of original indices by a constant (i.e.: 13/12).

¹⁴ The correlations between income indices and the national accounts estimates have been estimated by using different concept of disposable income. Specifically, the capital income has been excluded, since it is hardly reported correctly by respondents; taxes on capital and wealth have been included, considering that households are unable to distinguish the basis on which taxes are levied.

3-4 times larger than the actual one. Since, according [4], the coarsening component of total variance of estimates amounts to 250-350€ for the optimal brackets, it implies that misreporting of incomes is huge, and thus [7] holds for a reasonable range of bracketing schemas.

In order to select among the procedures and among the aggregations, it may be more informative to look at the second part of table 2, reporting the correlation coefficients computed for the yearly changes of the same time series. Many differences became apparent. First of all, some combination of bracketing and procedures are quite inefficient (e.g.: indexes based on median and mid-mean, specially associated with a fine bracketing). Secondly, Figure 4 shows that R^2 is quite stable, if not decreasing with the brackets width, until the width is less than 500€, then it rises up to a maximum corresponding to 4-5 income classes (600-800€ wide). Compared with the correlation among the levels of time series, these findings seem to strengthen the conclusion that large brackets may improve dramatically the accuracy of indexes derived from ISAE survey. In addition, several procedures fit satisfactory the incomes level but not its dynamics.

Looking at the two lines in Figure 4 together, the optimal bracketing for the ISAE data seems to be close to 600 €, since this width grants a good correlation with the level of income, and a reasonable concordance with its dynamics as well. As far as the overall evaluation of procedures listed in Table 1 is concerned, those based on rectangular, triangular and exponential interpolation functions provide both a first class approximation of income levels, and a good estimate of income dynamics. In turn, the procedures based on an update of the upper bound of the income distribution give a better approximation for the dynamics of households' net earnings. Finally, using the Pareto distribution, median and mid-mean provide unexpectedly bad results, although they share appreciable theoretical characteristics. Noticeably, indexes based on robust statistics are very sensitive to different bracketing and sometimes give very volatile results, marked by large outliers. Thus, they cannot be considered as reliable as the other procedures.

5.2 Volatility and robustness

Concordance between quantified indexes and the official estimates is only one of the selection criteria for optimal bracketing and quantification procedures. Another important issue concerns the stability over time of the monthly time series of indicators. In fact, it would be almost valueless an index even highly correlated to yearly National Accounts data, but very volatile from one month to the other as well. In addition, robustness to changes in bracketing would be

highly appreciated in case of periodic update of the bracketing schema, requested by the evolution of households income as time goes on.

Table 4 reports some statistics about monthly volatility and average (between 1990 and 2002) of indicators computed under different combinations of bracketing and quantification procedures. It is apparent that only 5 out of 8 procedures provide time series whose average is almost insensitive to the bracketing schema. Notably, passing from the original 22 class bracketing to the 3 class grouping, procedures based on robust statistics give estimates of average income that differ less than 4% each other. In turn, assuming rectangular, triangular or exponential distributions for income in the upper class, the averages of resulting indexes differ by 8% at most. Quite the reverse, by widening the brackets, the level of indexes computed assuming the Pareto distribution sharply reduces, while it raises by applying [10] and [11].¹⁵ This fact may prevent researchers to adopt the last three procedures when optimal bracketing is uncertain, and when revisions of income grouping occurred. However, even for the most stable procedures, larger brackets seem to give slightly lower indexes, as it is shown in Figure 5.

As far as volatility of indexes is concerned, the lower section of Table 4 shows that the standard deviation of monthly changes is quite large only for the procedures based on [10], [11] and median. In the latter case, by using 9 brackets few outliers in the time series raise the volatility by 220. Unexpectedly, the Pareto distribution does not provide badly volatile time series, even if such indexes are very sensitive to grouping. Figure 5 shows that, in general, the standard deviation of monthly changes of indexes raises as brackets widen, even if grouping households' answers in 3 groups only reduces volatility slightly, compared with 4, 5 or 7 group bracketing. This fact implies that the optimal number of brackets cannot be reduced too much, if quite smooth monthly time series are requested. Notably, brackets size should not exceed 450-600€ to ensure reasonably smooth time series.

¹⁵ In order to estimate a series of indexes based on [Pesaran0], the requested parameter α_N has been fixed to 7%, according the results of the non-linear regressions against the yearly National Account data.

CONCLUSIVE REMARKS

In this paper monthly survey data on households income, collected by using “brackets”, instead of a direct question, have been proved to be successful in providing timely and reliable estimates of households disposable income. Several procedures, discussed in Section 3, provide unbiased estimations of average income, under fairly general conditions, but the standard error of estimates increases with brackets size as well. However, people tend to underreport income, and this bias is likely to decrease with brackets width. Thus an optimal bracket size can be generally identified, that insures a reduction of underreporting without increasing estimate variance too much.

The paper described a series of procedures aimed to estimate average income from bracketed data. Some of them are based on a special treatment of the data falling in the upper open bracket provided in the questionnaire. Others are based on the extrapolation of income distribution within the upper class. A third group of procedures relies on the use of robust statistics computed on the observed distribution of answers. Only the methods belonging to the second group, possibly based on very simple functions, proved to be robust and reliable, even in extreme situations.

According to the empirical evidence showed in Section 5, concerning the survey carried out monthly by ISAE on Italian consumers, reducing the number of income brackets from the actual 22 to 5-9 seems to improve slightly the accuracy of indicators for every procedure. This finding may be quite unexpected, but is consistent with the arguments presented in Section 2.

Results on accuracy of indicators refer solely to the estimation of average income. Reducing the number of brackets might dramatically worsen the estimation of income quantiles, as the poor results obtained by using the median estimator seem to suggest. It implies that the actual survey, based on 22 brackets, could be continued in order to provide reliable indicators on income distribution.

Of course a more sound evaluation of optimal bracketing requires a real experiment involving parallel interviewing, devised to evaluate “anchoring”, “heaping”, and other bracketing effects on respondents. However, it should be stressed that the finding presented in this paper are likely to be even biased against the hypothesis that reducing brackets size may improve the accuracy of income indicators, as discussed in Section 5. Therefore, the results about optimal bracketing and procedures selection seem to be almost robust, even lacking for a proper and complete ad hoc experimental survey.

APPENDIX - FIGURES AND TABLES

Figure 1 – The optimal brackets size

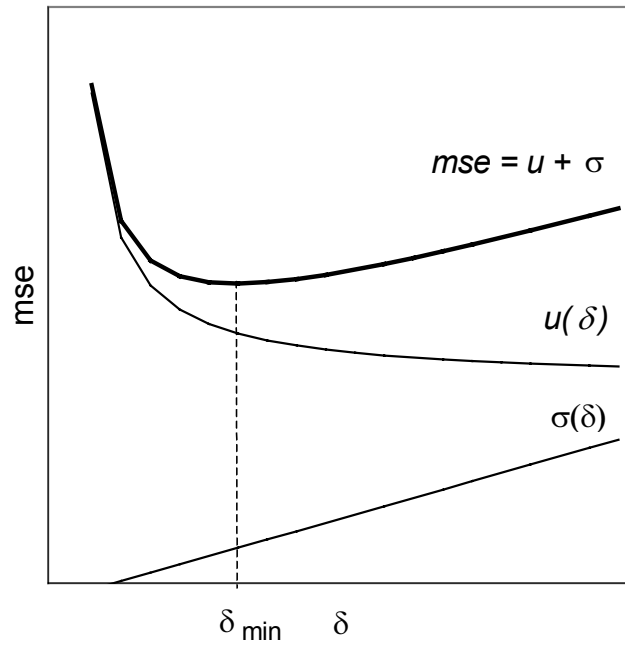


Figure 2: Estimation of the median from grouped data

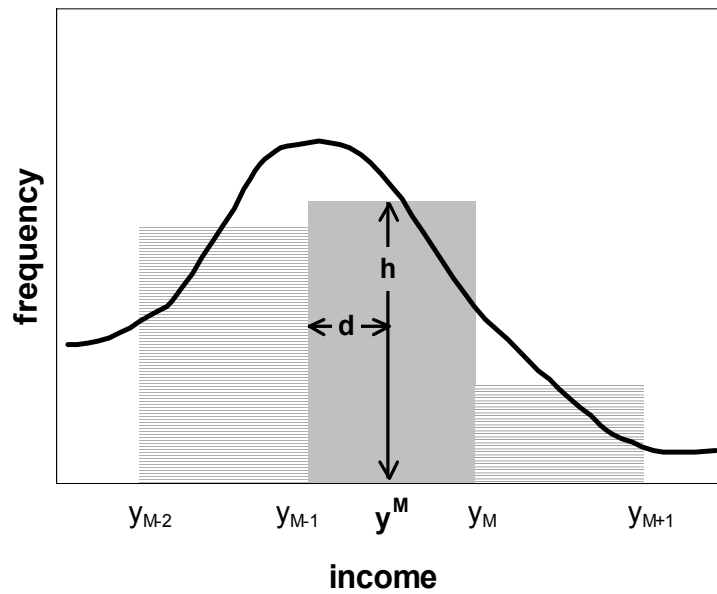


Figure 3: Households' income distribution according to the ISAE survey

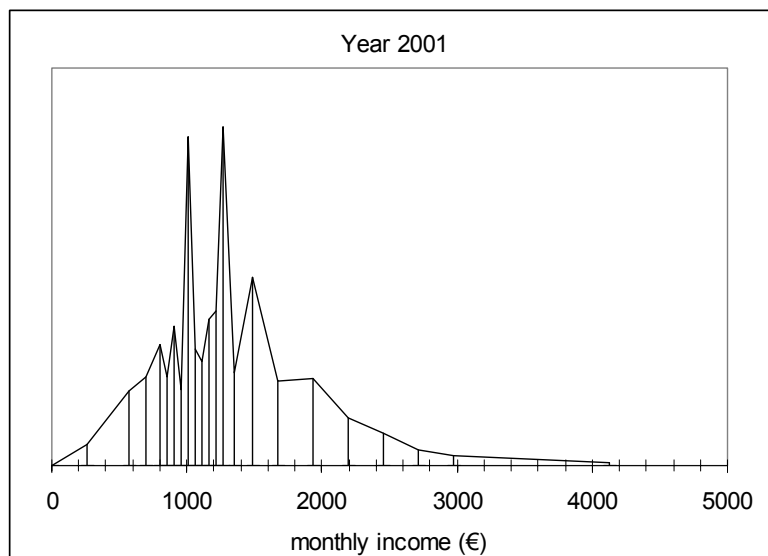
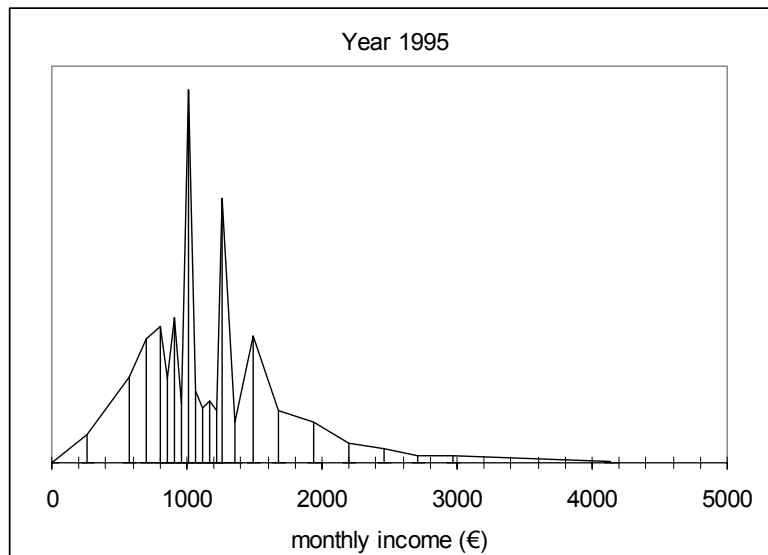
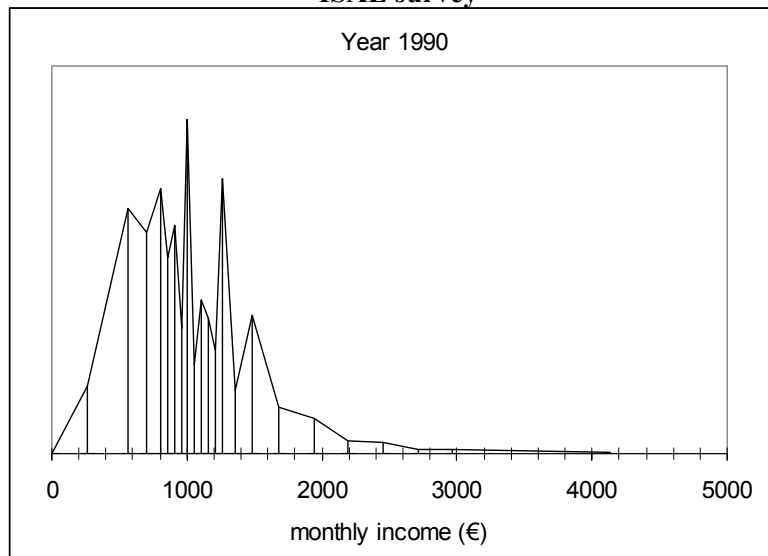


Figure 4: Brackets width and accuracy of income estimation
 (correlation with National Accounts estimates: average of various procedures)

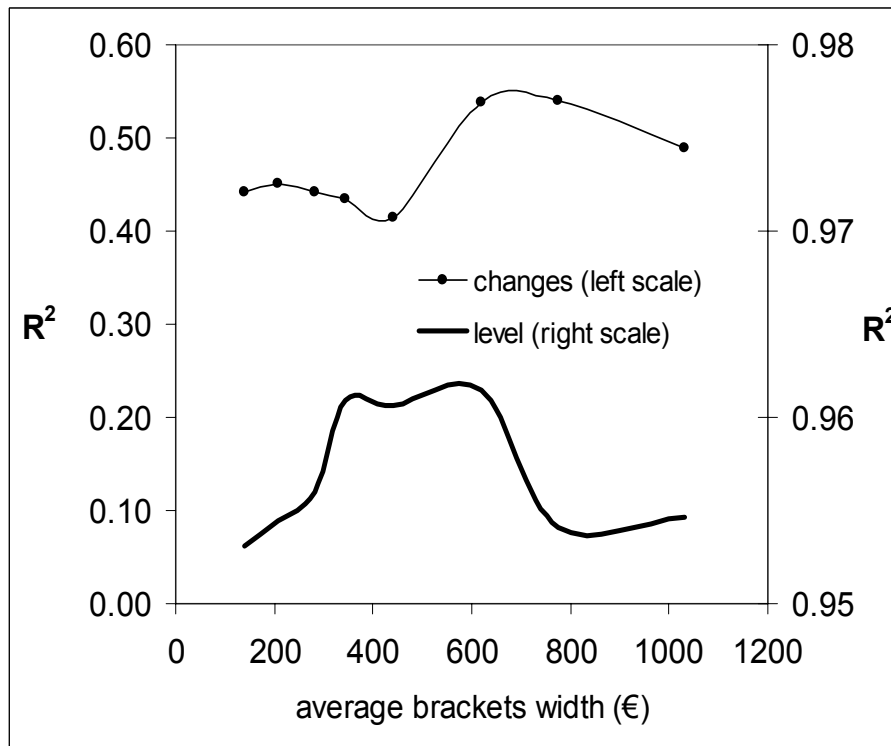


Figure 5: Mean and volatility of income indicators
 (average of various procedures)

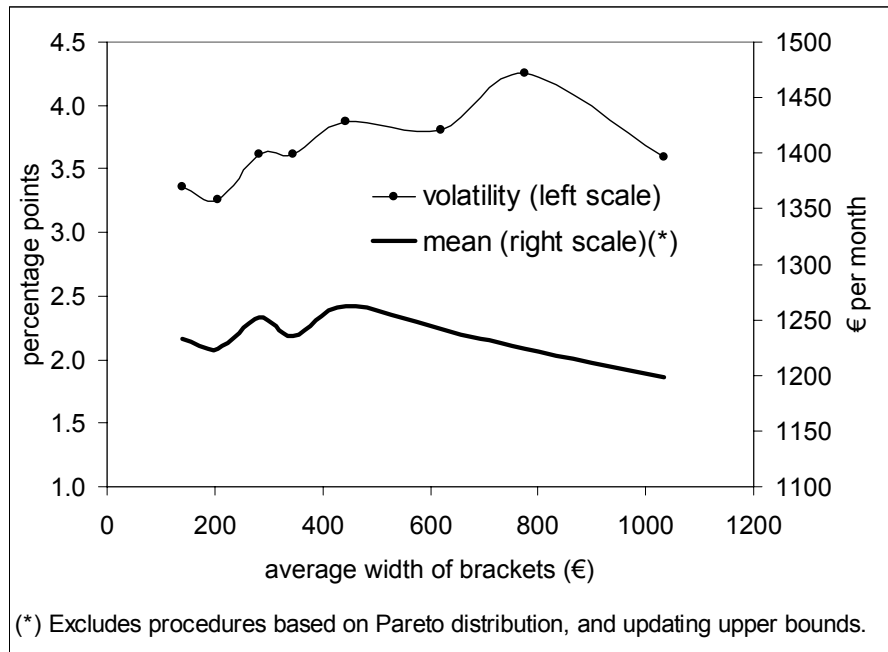
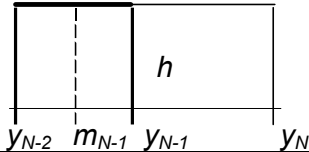
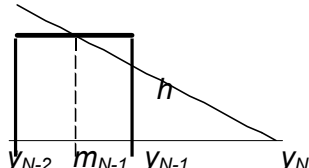
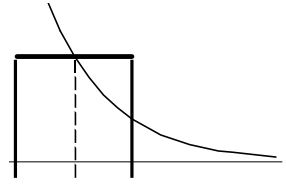
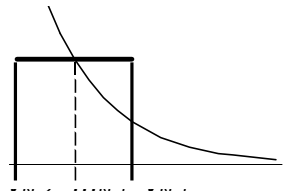


Table 1: The treatment of the upper open class

Hypotheses	Extrapolating functions		m_N
Escalating upper bound	<i>not specified (m_N changes proportionally to consumer prices P)</i>		$y_{N-1} + k P$
Proportional upper bound	<i>not specified (m_N changes proportionally to income Y)</i>		$y_{N-1} + k Y$
Rectangular (uniform) distribution	$g(y) = h^* = \frac{f_{N-1}}{y_{N-1} - y_{N-2}}$		$y_{N-1} + \frac{1}{2} \frac{f_N}{h}$ $h = h^*$
Triangular distribution	$a = h^* \frac{y_N}{(y_N - y_{N-1}) + \frac{1}{2}(y_{N-1} - y_{N-2})}; b = \frac{h^*}{(y_N - y_{N-1}) + \frac{1}{2}(y_{N-1} - y_{N-2})}$ $y_N = y_{N-1} + \frac{f_N + \sqrt{f_N(f_N + f_{N-1})}}{h^*}$		$y_{N-1} + \frac{2}{3} \frac{f_N}{h}$ $h = h^* \frac{f_N + \sqrt{f_N(f_N + f_{N-1})}}{f_N + \frac{1}{2}f_{N-1} + \sqrt{f_N(f_N + f_{N-1})}}$
Exponential distribution	$g(y) = ae^{-by}$ $a = \ln(bf_N) + by_{N-1}; b = \frac{\ln\left(1 + \frac{f_{N-1}}{f_N}\right)}{y_{N-1} - y_{N-2}}$		$y_{N-1} + \frac{f_N}{h}$ $h = f_N \frac{\ln\left(1 + \frac{f_{N-1}}{f_N}\right)}{y_{N-1} - y_{N-2}}$
Pareto distribution	$g(y) = ay^{-b}$ $a = \frac{(b-1)f_N}{y_{N-1}^{1-b}}; b = \frac{1 + \frac{\ln\left(1 + \frac{f_{N-1}}{f_N}\right)}{\ln\left(\frac{y_{N-1}}{y_{N-2}}\right)}}{1}$		$y_{N-1} + \frac{f_N y_{N-1}}{h y_{N-1} - f_N} \cong y_{N-1} + \frac{f_N}{h} \text{ if } \frac{f_N}{y_{N-1}} \cong 0$ $h = f_N \frac{\ln\left(1 + \frac{f_{N-1}}{f_N}\right)}{\ln\left(\frac{y_{N-1}}{y_{N-2}}\right)}$

Note: The time subscript t is omitted for the sake of notation simplicity.

Table 2: Plan of the simulated brackets

Original brackets		Upper bounds of aggregated brackets (€)						
Average frequency (%)	Upper bounds (€)	(A)	(B)	(C)	(D)	(E)	(F)	(G)
4.8	500	600	600	850	850	1000	1000	1200
3.4	600	750	850	1050	1100	1300	1400	3100
5.9	750	900	1000	1250	1300	2050	3100	n.a.
2.8	850	1000	1100	1400	1550	3100	n.a.	
2.1	900	1050	1250	1550	2050	n.a.		
3.1	950	1150	1400	1800	3100			
1.8	1000	1250	1550	2050	n.a.			
7.5	1050	1300	1800	3100				
2.7	1100	1400	2050	n.a.				
2.4	1150	1550	3100					
3.4	1200	1800	n.a.					
3.5	1250	2050						
7.6	1300	2300						
5.2	1400	3100						
10.5	1550	n.a.						
9.5	1800							
9.9	2050							
5.4	2300							
3.6	2600							
1.7	2850							
1.1	3100							
2.0	n.a.							
Number of classes	22	15	11	9	7	5	4	3
Average width of closed brackets	141	207	282	344	443	620	775	1033

Table 3: National Account estimates and income quantifications

Number of classes	22	15	11	9	7	5	4	3
Average width of brackets (€)	141	207	282	344	443	620	775	1033
<i>Correlation with income level</i>								
Rectangular distribution	0.963	0.963	0.964	0.967	0.967	0.968	0.959	0.963
Triangular distribution	0.962	0.962	0.964	0.967	0.967	0.968	0.959	0.963
Exponential distribution	0.960	0.961	0.963	0.967	0.967	0.967	0.958	0.963
Pareto distribution	0.956	0.957	0.959	0.961	0.955	0.967	0.960	0.960
Escalating upper bound	0.959	0.957	0.957	0.961	0.963	0.959	0.952	0.963
Proportional upper bound	0.945	0.947	0.948	0.958	0.958	0.955	0.946	0.951
Median	0.933	0.938	0.933	0.942	0.944	0.940	0.941	0.936
Mid-mean	0.949	0.951	0.960	0.964	0.964	0.968	0.958	0.938
<i>Correlation with yearly changes of income</i>								
Rectangular distribution	0.521	0.510	0.512	0.483	0.455	0.565	0.566	0.518
Triangular distribution	0.526	0.515	0.515	0.487	0.460	0.568	0.568	0.519
Exponential distribution	0.537	0.524	0.522	0.495	0.468	0.573	0.570	0.521
Pareto distribution	0.359	0.385	0.403	0.366	0.329	0.470	0.534	0.499
Escalating upper bound	0.529	0.534	0.543	0.527	0.514	0.564	0.599	0.580
Proportional upper bound	0.499	0.509	0.519	0.500	0.484	0.551	0.584	0.551
Median	0.264	0.308	0.253	0.257	0.242	0.552	0.585	0.150
Mid-mean	0.307	0.327	0.267	0.364	0.359	0.459	0.317	0.580

Table 4: Statistics on monthly time series of income indexes

Number of classes	22	15	11	9	7	5	4	3
Average width of brackets (€)	141	207	282	344	443	620	775	1033
Mean of indexes								
Rectangular distribution	1271.0	1256.7	1297.3	1264.3	1298.5	1273.7	1257.0	1200.4
Triangular distribution	1274.1	1260.0	1304.0	1272.1	1315.6	1290.7	1274.1	1226.6
Exponential distribution	1280.3	1266.5	1317.4	1287.6	1349.7	1324.8	1308.1	1279.0
Pareto distribution	1184.4	1154.3	1121.3	1052.2	968.0	943.1	926.5	785.0
Escalating upper bound	1352.8	1397.1	1535.0	1686.5	2072.1	2047.2	2030.6	2377.6
Proportional upper bound	1343.0	1375.9	1490.1	1593.1	1873.7	1834.5	1807.8	1937.1
Median	1135.5	1135.3	1136.1	1148.4	1147.0	1134.7	1119.0	1108.8
Midmean	1202.6	1201.5	1203.8	1203.3	1199.8	1182.9	1161.7	1177.2
Volatility of indexes (standard error of monthly percentage changes)								
Rectangular distribution	2.92	2.67	3.06	2.96	3.14	3.20	3.29	2.79
Triangular distribution	2.95	2.71	3.09	3.02	3.24	3.31	3.39	2.95
Exponential distribution	3.02	2.78	3.17	3.16	3.47	3.54	3.63	3.30
Pareto distribution	2.42	2.39	2.43	2.46	2.67	2.64	2.69	2.80
Escalating upper bound	3.99	4.00	4.80	5.33	5.44	5.54	5.61	5.03
Proportional upper bound	4.04	3.99	4.90	5.37	5.76	5.84	5.91	5.17
Median	4.46	4.48	4.31	n.a.	4.53	3.04	6.18	3.46
Midmean	3.11	3.10	3.17	3.00	2.72	3.39	3.35	3.26

REFERENCES

- Atkinson A. B. and F. Bourguignon (1998), *Handbook of Income Distribution*, (eds.), North Holland, Amsterdam.
- Brandolini A. and G. Parigi (1993), Il reddito disponibile delle famiglie e la sua distribuzione: una stima derivata dall'inchiesta mensile dell'ISCO, *mimeo*, Bank of Italy, Rome.
- Cowell F. A. and F. Metha (1982), The estimation and Interpolation of Inequality Measures, *Review of Economic Studies*, Vol. 49, 273-90.
- D'Elia E. (1991), La quantificazione dei risultati dei sondaggi congiunturali: un confronto fra procedure, *Rassegna di lavori dell'ISCO*, p. 1-71, ISCO, Rome.
- D'Elia, E. and B.M. Martelli (2000), Households' Incomes, Professional Status, and Inflation in Italy in the 1990s: Evidence from the ISAE Consumer Survey, in the *Proceedings of the XIV CIRET Conference*, Ashgate, London, 2000.
- Dempster A. P. and D. B. Rubin (1983), Rounding Error in Regression: The Appropriateness of Sheppard's Corrections, *Journal of the Royal Statistical Society, Series B*, Vol. 45, Issue 1, 51-59.
- Heitjan D. F. and D. B. Rubin (1991), Ignorability and coarse data, *Annals of Statistics*, 19(4), 2244–2253.
- Hoaglin D. C., Mosteller F. and J. W. Tukey (1983), *Understanding Robust and Exploratory Data Analysis*, (eds.) John Wiley & Sons, New York.
- Hsiao, C. (1983): Regression analysis with a categorized explanatory variable. In S. Karlin, T. Amemiya, and L. Goodman (Eds.), *Studies in Econometrics, Time Series, and Multivariate Statistics*. New York: Academic Press.
- Juster F. T. and J. P. Smith (1997), Improving the quality of economic data: Lessons from the HRS and AHEAD, *Journal of the American Statistical Association*, 92(440), 1268–1278.
- Kakwani N. (1976), On the estimation of income inequality measures from grouped observations, *Review of Economic Studies*, Vol. 43, 483-92.

- Manski C. F. and E. Tamer (2002), Inference on regressions with interval data on a regressor or outcome, *Econometrica*, 70(2), 519–546.
- Martelli B. M. (1998), "Le inchieste congiunturali dell'ISCO: aspetti metodologici", in "Le inchieste dell'ISCO come strumento di analisi della congiuntura economica", *Rassegna di Lavori dell'ISCO*, Anno XV, n.3, 13-67.
- Pesaran I. (1984), Expectations Formations and Macroeconomic Modelling, in *Contemporary Macroeconomic Modelling*, P.Malgrange & P.A. Muet (Eds.), Blackwell, Oxford.
- Torelli N. and U. Trivellato (1993), Modelling Inaccuracies in Job-Search Duration Data, *Journal of Econometrics*, Vol. 59.
- Tourangeau R., Rips L. J. and K. Rasinski (2000), *The Psychology of Survey Response*. New York, NY and Cambridge, UK, Cambridge University Press.
- Visco I. (1984), *Price Expectation in Rising Inflation*, North Holland, Amsterdam.
- Winter J. K. (2002), Bracketing effects in categorized survey questions and the measurement of economic quantities, Paper No. 02-34, University of Mannheim.