Dinamiche e persistenze nel mercato del lavoro italiano ed effetti di politiche (basi di dati, misura, analisi)

Progetto di ricerca cofinanziato dal MIUR
(Ministero dell'Istruzione, dell'Università e della Ricerca) – Assegnazione: 2001
Coordinatore: Ugo Trivellato

Statistical imputation in conjunction with micro-simulation of income data

G. Betti*, V. Verma*, M. Natilli**, F.Ballini**

* *Dipartimento di Metodi Quantitativi, Univ. di Siena*
** *International Social Research Ltd e Univ. di Siena*

Working Paper n. 55   settembre 2003

Unità locali del progetto:

Dip. di Economia "S. Cognetti De Martiis", Univ. di Torino	(coord. Bruno Contini)
Dip. di Scienze Economiche, Univ. "Ca' Foscari" di Venezia	(coord. Giuseppe Tattara)
Dip. di Metodi Quantitativi, Univ. di Siena	(coord. Achille Lemmi)
Dip. di Scienze Statistiche, Univ. di Padova	(coord. Ugo Trivellato)
Dip. di Politiche Pubbliche e Scelte Collettive, Univ. del Piemonte Orientale	(coord. Alberto Martini)

Dip. di Scienze Statistiche
via C. Battisti 241-243, 35121 Padova

1 Introduction

In this paper we address issues of special concern in the imputation of *complex, composite variables*. Specifically, the paper identifies and discusses pertinent questions relating to the imputation of micro-level *longitudinal data* on income of households and persons by detailed income component *in conjunction with micro-simulation modelling* involved in the construction of household income variables in standardised analytical forms. *Imputation* refers to the process of using the information existing in a dataset, as well as external information where appropriate, to produce improved estimates for missing, implausible or inconsistent elements in the dataset. *Micro-simulation* in the context of household and personal income data, as interpreted here, means generating and relating detailed information on income, by source (component) and type in its different forms, according to 'destination', i.e. according to how gross income accrued by households and individuals is partitioned into taxes, social insurance contributions, and the remaining net amounts available for private consumption. Micro-simulation may encompass other aspects such as simulation, on the basis of characteristics and circumstances of individuals, of the benefits and other social transfers they *should* receive, but these types of issues are not addressed here.

These two statistical processes of imputation and modelling have to be implemented in conjunction with each other - in so far as imputation must be based on 'donor' data in a homogeneous form (which is the function of micro-simulation in the above sense to create), while micro-simulation generally requires data with no missing values (which is the function of imputation to ensure). This paper aims to discuss these issues in concrete terms. We believe this discussion to be rather unique in this respect, in so far as the issues of imputation and micro-simulation for income data have not been much discussed *together* in the literature.

Any good micro-level imputation procedure must meet some basic standards. The imputed values generated should preserve the correlation structure of the data, should be determined stochastically rather than deterministically, and should be consistent or at least plausible. There are added requirements when we are dealing with complex, composite variables such as survey information on household and personal incomes, especially in the longitudinal context. To impute where possible and reasonable is more critical for this type of data: total household income is made up of a large number of components, and rejecting a unit with incomplete information is unacceptable as it would result in the loss of much valuable information. Income components as variables do not form an independent set: they are mere components of the same 'organic' aggregate (total income of the household), and hence it is not meaningful to impute individual components separately. Even how that aggregate is partitioned is not pre-determined, and hence nor is the resulting correlation structure of the data. In longitudinal data, the relationship between lagged and current values of the same variables, as well as this relationship across variables, needs careful treatment, and we address this issue in this paper. Our special concern, however, is with issues arising from the fact that, on the one hand, the available information on income is in heterogeneous forms, and on the other, some of this information is missing and needs to be imputed. This requires the use of imputation and modelling techniques in conjunction.

The concrete context for this discussion is provided by longitudinal information on household and personal income, generated in a multi-national comparative form through major programmes developed and supported by the European Commission. These programmes are the European Community Household Panel (ECHP)[1], and its successor, EU Statistics on Income and Living Conditions (EU-SILC)[2]. Taking the kind of statistics on income generated from these sources as the point of reference, **Section 2** briefly reminds of the type of complex, composite variables we are dealing with. **Section 3** describes an imputation strategy appropriate for this type of data. It largely derives from the actual practice of ECHP, but incorporates suggestion for improvements which we have formulated on the basis of an evaluation of this experience. **Section 4** describes essentials of a micro-simulation model in this context. This is based on our work in relation to the Siena Micro-Simulation Model (SM2) which has been developed for *multi-country* (at least Europe-wide) *application*, specifically to meet requirements of the EU-SILC programme. Finally, **Section 5** aims to present a concise but precise description of the methodology of implementing imputation and modelling procedures in tandem.

2 Income variables

Obtaining comparable and detailed information on the household and a personal income means that in most circumstances the information is complex and diverse. This background factor has to be kept in view in the following discussion of statistical procedures for imputation and modelling of income variables.

<u>Collection variables</u>

Many factors contribute to this complexity and variation, especially if the information has to be collected through interview surveys. For instance:

- Diverse sources and circumstances. Because of the fact that households and persons may receive income from many diverse sources in diverse forms, and also as a matter of practical necessity of data gathering, information on income normally needs to be collected or compiled in great detail.

- Different reporting units. Major part of the income is enumerated at the individual level; individual incomes are then aggregated over members to obtain the household total. However, certain components, which are received by the household as a whole, may be recorded at the household level.

- Long reference period. Most commonly, information on income has to be collected with a long reference period, typically a whole year. Furthermore, the information has to be collected retrospectively, after the completion of the reference period and when the information has generally become available.

- Further decomposition of items. Difficulties in recall associated with the long period means that each item of information may have to be further broken down. It is common, for instance, to ask for income from each source in the form of the normal amount received per month (or in terms of another reference period if more appropriate) and the number of months received during the fixed reference year. But annual totals may have to be directly recorded if the payment is a lump sum, irregular or the number of months received cannot be specified for other reasons. For some items simpler but approximate procedures may be unavoidable.

[1] The European Community Household Panel (ECHP): Survey methodology and implementation. Volume 1.

[2] SILC065: Description of target variables – Version 2004

° Different forms. The information may be net or gross amounts or in some other form. This may vary from one respondent to another, and also from one item to another for the same respondent.

Analysis or 'target' variables

Usually, data analysis requires much more aggregated and uniform set of variables, which are meaningful for analysis.

Imputation and modelling variables

The imputation variables and modelling variables would normally represent intermediate levels of breakdowns of total income between the collection and analysis sets[3]. However, these two sets themselves are not necessarily identical. In fact, in determining how 'detailed' a set of variables is needs to take into account at least two dimensions. The first is the degree of breakdown in terms of substantive content, i.e. how finely the variables divide the total income. The second dimension, determining how 'detailed' a set of variables is, concerns the population unit to which they refer – for example whether a variables refers to the whole household, an individual person, or to some other intermediate 'tax unit'. One set of variables may be more detailed than another set in terms of one criterion (such as the partitioning of income into components), but less detailed in terms of the other criterion (such as income at person versus household level).

In the combined implementation of the imputation and modelling systems considered in this paper, these variables of different types and the relationship between them, need to be clearly defined. In practice, imputation and modelling procedures will generally involve different sets of variables, the latter normally representing less detailed breakdowns of total income. A modelling variable (income component) may correspond to a single imputation variable, or an aggregation of a group of imputation variables. Furthermore, while imputation variables normally involve individual as the unit, much of modelling involves their aggregation over individuals in the same tax unit.

This relationship of variables is a complicating issue in the development, implementation and exposition of the complex statistical procedures being considered here. For simplicity and to explain the interaction between imputation and modelling more clearly, *we have assumed throughout in the present discussion that the two operations involve an identical set of variables.*

3 The imputation procedure

Some important requirements

An important characteristic of the income variables is that these variables form a set in which there is an interdependence between all the components. Hence an appropriate approach is through a multivariate model involving a multiple regression sequence[4]. The sequential multivariate model used makes for more complete imputation of the variables, while at the same time safeguarding their variance and their inter-correlation.[5] In brief outline the approach may be described as follows.

[3] DOC.PAN 164/2001-12, imputation of income in ECHP.

[4] IVE*ware*: Imputation and Variance Estimation Software. User guide, *Survey Methodology Program. Survey Research Center, Institute for Social Research, University of Michigan.*

[5] Our approach uses (and is almost entirely based on) the EM procedure based on the work of a team at the Institute of Social Research (ISR) of the University of Michigan, who have developed the "IVE-Ware" software for the purpose.

- The variables are divided into two types: auxiliary variables used to impute the others, and target variables which are the subject of the imputation. For instance, in the study of household income, the auxiliary variables can be those relating to the demographic characteristics (sex, age) and to the characteristics of the labour force. The target variables are components of income defined at an appropriate level of aggregation.

- Some preliminary work may be required in defining the target variables and coveting them in a suitable form for application of the imputation procedure. This requires in particular dealing with categorical variables such as whether a particular component of income is received ("yes/no"), values specified as ranges for certain variables, number of months an income has been received when the average monthly amount has been specified, etc. These details are very data-specific and need not concern us here.

- The target variables are arranged in a sequence, starting with those with the smallest proportion of (or with no) missing values. Going down in sequence, each target variable is imputed using all the variables above it, for which all information is available (or has been previously imputed), as auxiliary variables in the multivariate regression.

- Once a variable with missing values has been imputed, it is moved from the second set to the first, i.e. used as an auxiliary variable in imputation of the next variable in the list.

- After all variables in the list have been dealt with as above, the process is started again with the first variable in the target set, but this time using all the other variables as predictors, using for each the given or the most recently imputed value. The process is performed for each variable in turn, and is repeated iteratively.

- Units for which the income information available is so incomplete that no reasonable basis exists for imputation should be excluded, however.

- The variables in the auxiliary set U (and any lagged variables if used in the regressor set) are required not to have any missing values. If not, they themselves have to be imputed in some way. Ad hoc procedures may suffice to meet this requirement. However, it is desirable to standardise these and incorporate these into the standard imputation procedure to the extent possible. This desirable feature is incorporated into the procedure described below.

- Normally, one item of information is still available for cases where the actual amount of income is missing: namely that a non-zero amount was received. Hence *the "donors" for the imputation should only be those with non-zero amounts* reported for the variable concerned.

- Where supplementary information is available on the permitted range or plausible values, the imputations must be confined within those limits. Here as well, the donors

In the context of the topic of this paper, a most fundamental consideration is the need to integrate imputation and modelling procedures. Such integration becomes all the more important with greater diversity in the forms on which the income information has been recorded. As will be seen, a fundamental requirement of such integration is that the *imputation and modelling routines need to be interwoven and applied variable-by-variable in combination*. It happens that the above mentioned imputation program (IVE-Ware) cannot be applied in an automatic manner to a whole set of variables in one go. To overcome this limitation, we have developed appropriate SAS routines which permit repeated "calling" of IVE-Ware at each step of the imputation-modelling procedure, with the flexibility to alter parameters of the program application as required.

with zero amount have to be excluded if the known range for the imputation does not include zero.

The model is as follows. With U as the matrix containing variables with no missing data (including as a result previous imputation), and Y1, Y2...Yk are variables with increasing rates of missing data, the sequence of imputations is determined by the following factorisation:

[Y1¦U] [Y2¦U, Y1] ...[Yk¦U, Y1, ..., Yk-1]

where [Y¦X] is the conditional joint distribution of Y where X is known.

The form of regression depends on the nature of Y, such as a generalised linear regression for continuous variables (as in the case of income amounts), a logistical regression for binary variables, etc. In the application of the above procedure, ideally, set U should also include, for the particular variable being imputed, its lagged variable (value from the preceding wave) as a regressor. Clearly, in order not to introduce multi-collinearity, the current and the lagged versions of any of the other variables cannot both be included as regressors. This means that the set of regressors has to vary somewhat from one variable to another.

The following subsections describe main aspects of the imputation procedure. It is assumed in this section that all income values have been specified in a standard form, so that issues relating to modelling do not arise. This is helpful in describing the imputation procedures more simply and clearly. Their application in the real situation, i.e. in combination with modelling, will be elaborated in subsequently.

Imputation of regressor and lagged variables

Application of the procedure requires that variables used as regressors in the model contain no missing values. These regressors include relevant background characteristics of the unit and also, where available, values of the target variables at the preceding wave ("lagged variables"). Missing values are of course encountered in regressor variables. The following systematic approach to deal with this problem.

Let us denote with Y the set of *k* income Imputation Variables $\mathrm{Y} \equiv (\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_k)$, with U the <u>basic set of *r* regressors</u> to be used for imputation, and with $\mathrm{Y}^{(t-1)}$ the set of *k* lagged variables corresponding to the set Y as the <u>supplementary set of *k* regressors</u>. The set of regressors can be more flexible $\mathrm{U_j}$, with some variation by variable if required.

We have missing values in Y, but we may also have missing values in the regressor sets $\mathrm{Y}^{(t-1)}$ and even in U. Now the first step consists in imputing missing values in any variable in order to have full information in the set of regressors.[6] Let $\mathrm{Z} \equiv \left(\mathrm{Y}, \mathrm{U}, \mathrm{Y}^{(t-1)}\right)$ be the entire set of (2k+r) variables considered, and order variables in Z according to the ascending incidence of (proportion of applicable) values missing. Denote with z_i any ordered variable and with Z_i the set of ordered variables from 1 to i. The first set of imputations is defined as follows (see Table 1):

1. Let be j the number of variables in Z with full information.
2. Consider variable $\mathrm{z_{j+1}}$ and the set of all preceding variables $\mathrm{Z_j}$.

[6] Variables for which the necessary imputations cannot be performed may be dropped from the sets $\mathrm{Y}^{(t-1)}$ and U.

- For the imputation of a variables belonging to the subsets Y or U: remove from Z_j all lagged variables for which the corresponding current variable is also present in the set (i.e., do not include a lagged variable $y_i^{(t-1)}$ if the corresponding current variable y_i^t is already present in the set).

- For the imputation of a variables belonging to the subset $Y^{(t-1)}$: remove from Z_j all current variables for which the corresponding lagged variable is also present in the set. (i.e., do not include a current variable y_i^t if the corresponding lagged variable $y_i^{(t-1)}$ is already present in the set).

This gives the reduced regressor set Z_j'.

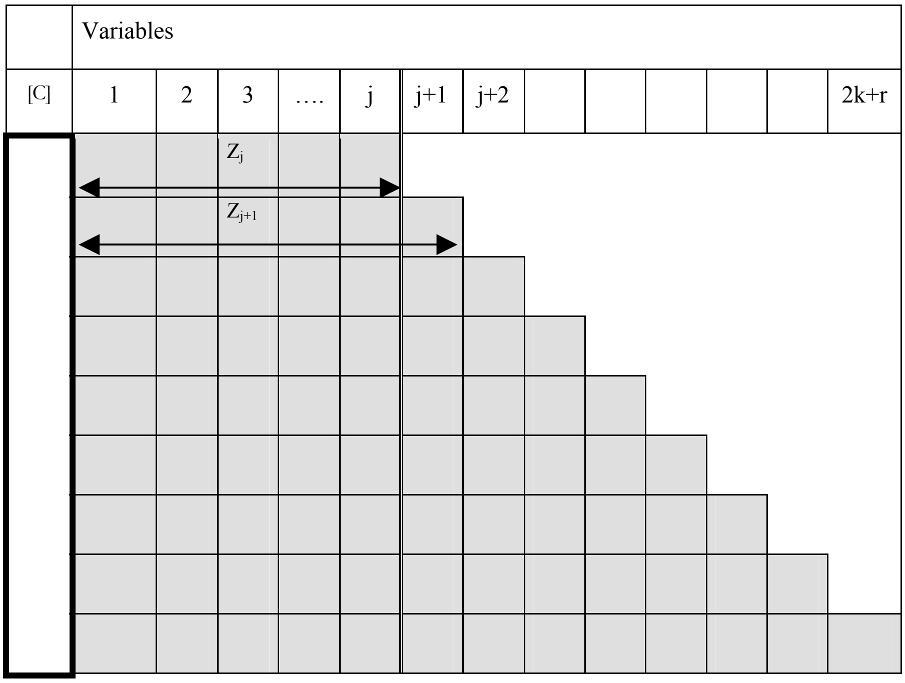
3. Impute missing values of z_{j+1} using the set Z_j' as regressors.

4. It is assumed that whether or not a quantitative (income) variable in set Z has a non-zero value is known (or that such information has been already imputed in a preliminary step). The imputation is done for cases with z_{j+1} missing but known to be non-zero. The donor cases determining the regression are those with z_{j+1} *known and non-zero*.

The inclusion of zero values in the donor set to impute missing values when they are known to be non-zero can seriously distort the results. This has been amply demonstrated by our detailed empirical work. For example, it has been found that including zeros among the donor population, when the recipients are known to have a non-zero (but unknown, of course) values, often gives imputed values concentrated at the extremities of the permitted ranges, or values which are implausibly different in distribution from the donor set. Detailed results from such empirical investigations are not presented here for reasons of space.

5. Add imputed variable z_{j+1} to the 'full information' set, and continue the above process till the last (imputable) variable has been imputed.

Table 1. Arrangement of variables according to the proportion of missing values



Shaded cells indicate variables with available data, and blank cell with missing data.

[C]: Columns (Y-axis) shows the proportion of applicable cases for which data are available.

The objective of this initial imputation is only to complete the set of regressor variables in a reasonable way, and not to provide the 'final' imputations for the target variables Y. Hence only the imputed values in U and $Y^{(t-1)}$ are retained at the end of this cycle, and using these a more precise imputations for the target set Y are performed as described in the next section.

Imputation of target variables Y

As before, let Y denote the set of *k* income Imputation Variables $Y \equiv (y_1, y_2, ..., y_k)$, with U the basic set of *r* regressors to be used for imputation, and with $Y^{(t-1)}$ the set of *k* lagged variables corresponding to the set Y. We have missing values in Y, but all values in $Y^{(t-1)}$ and U are available or have been imputed at the preceding step. We begin by ordering variables in Y according to the ascending number of (proportion of applicable) values missing; denote with y_i any ordered variable and with Y_i the set of ordered variables from 1 to i.

The first set of imputations is defined as follows:

Let be j the number of variables in Y with full information. Consider variable y_{j+1} and the set of all preceding variables Y_j.

For imputing missing values of y_{j+1}, the regressor set Z_{j+1} consists of

$$Z_{j+1} = \left(U_{j+1} + Y_j + y_{j+1}^{(t-1)}\right) \quad \Big| \quad \textbf{donors } y_{j+1} > 0, \quad \textbf{recipients } y_{j+1} \text{ missing but } >0.$$

Here Y_j is the *set* of variables which had none or a lower proportion of missing values (among the applicable cases for each of the variables) than current variable y_{j+1}, and for which any of those missing values have already been imputed. $y_{j+1}^{(t-1)}$ is the lagged variable corresponding to y_{j+1}, and U_{j+1} is the regressor set for $y_{j+1.}$ As in the previous subsection, it is assumed that whether or not a variable in set Y has a non-zero value is known or has been already imputed. The imputation is done for cases with y_{j+1} missing but known to be non-zero, and the donor cases are those with y_{j+1} *known and non-zero*.

The requirement to include only the lagged variable corresponding to the current variable being imputed and to confine the donor population to those with *known and non-zero* values on the variable of interest, means that the set of regressors has to be varied from one variable to another. This necessitates calling the imputation routine separately for each variable in the sequence.

The process is continued till the last (imputable) variable in Y has been imputed.

Table 2. The first ("triangular") cycle of imputation

		Regressors:		
Imputation Variable	Donor population	Basic set (background variables)	Fully available or previously imputed variables (in the sequence ordered according to increasing proportion of missing values)	Lagged Variable
Y_j=(y_1,y_2, .. ,y_{j-1},y_j)	(Variables with full information)			
y_{j+1}	$y_{j+1} > 0$	U_{j+1}	Y_j= (y_1,y_2, .. ,y_{j-1},y_j)	$y^{(t-1)}_{j+1}$
y_{j+2}	$y_{j+2} > 0$	U_{j+2}	Y_{j+1}= (y_1,y_2, .. ,y_{j-1},y_j,y_{j+1})	$y^{(t-1)}_{j+2}$
y_{j+3}	$y_{j+3} > 0$	U_{j+3}	Y_{j+2}= (y_1,y_2, ... ,y_{j-1},y_j,y_{j+1},y_{j+2})	$y^{(t-1)}_{j+3}$
…..	…..			
y_k	$y_k > 0$	U_k	Y_{k-1}= (y_1,y_2, ….. ,y_{j-1},y_j,y_{j+1}, ….. ,y_{k-1})	$y^{(t-1)}_{k}$

Imputation cycle for the full set:

Once all variables in Y have been imputed once, the following cycle is applied iteratively. The variables are ordered as before, in increasing proportion of the originally missing values. In each application, the full set of regressor variables is used, using all values, whether originally available or imputed in previous cycles. As before, the imputation routine is called separately for each variable in the sequence. The sequence is repeated a number of times. Based on our experience with ECHP data, a few (3-5, say) cycles should be sufficient in most cases.

Table 3. Iterative cycle

Imputation Variable	Donor population	Regressors:		
		Basic set	All Y except y_j: available or previously imputed values	Lagged Variable
y_j	$y_j > 0$	U_j	$Y_{(j)}$= (y_1,y_2, ….. ,y_{j-1}) + (y_{j+1}, ….. ,y_k)	$y^{(t-1)}_{j}$

4 Micro-simulation[7]

Information on income of households and persons collected from surveys or similar sources may be reported in different forms by different units and even for different

[7] Verma, Betti, Ballini, Natilli, Galgani, *Personal income in the gross and net forms: Applications of the Siena Micro-Simulation Model (SM2).*

components by the same unit: forms such as net income or income gross of taxes and social insurance contributions. *Assuming that information is available on all income components in some form*, in this section we summarise essentials of the micro-simulation procedure for estimating, on the basis of the prevailing fiscal system, the full set of information at the micro-level in a homogeneous form: gross income by income component, with breakdown into taxes, social insurance contributions and net amounts.[8]

Terminology: forms of income and their relationship

Table 4 summarises the terms and relationships between income forms.[9] Gross income (G) refers the total income from all sources received during a reference period by an individual, household or other 'tax unit', before payment of tax or social insurance contributions. This income may be divided into components (i) depending on the nature of the income and its source, such as income from employment, self-employment, pensions, capital, social transfers, etc. Some components of gross income are subject to social insurance contributions (S). Normally these contributions apply only to income from work, but in any case tend to be *component-specific*, i.e. for any particular component determined independently of other components of income. Gross taxable income (H) is gross income less social insurance contributions: H = G – S. Deductions (D) refers to part of gross taxable income which is exempt from tax. Net taxable income (Y) is obtained by subtracting from gross taxable income this part which is tax exempt: Y = H - D.

Initial tax due (W) is determined by the prevailing income tax schedule, normally as some function of net taxable income, W = W(Y). This functional relation tends to be very complex, however. *For the main part, it is applied not to income components one at a time, but to net taxable income pooled over all or most of the components.* Furthermore, the unit for the purpose of tax assessment may not be the income of a single individual but pooled income all individuals forming a tax unit. This is a crucial point. As will be discussed in more detail later, it is because of this requirement of *pooling of income over components and individuals* that the modelling and imputation processes cannot be applied to the data independently of each other.

The tax due is normally reduced by tax credits (C). Deduction of these tax credits from the tax due gives the final tax paid: X = W – C. Total net income (N) is total gross taxable income less tax paid: N = H - X.

In certain systems, income as initially received is subject to retention at source of tax and/or social insurance contributions; for instance, for income subject to both these retentions we have: XTS = G-S(G)-T(H) = H-T(H). Unlike the amount of 'final tax due' W or X which is determined through complex rules involving pooled income over individuals and components, the relationship T(H) determining tax retention at source is often component-specific and much simpler.[10] The same applies to any income components which are taxed separately from pooled income as defined above (such as at a flat or some other component-specific rate).

[8] The micro-simulation procedure as described here has been implemented in the Siena Micro-Simulation Model (SM2).

[9] Actual fiscal systems tend, of course, to be somewhat more complex than implied in this overview.

[10] However, sometimes the amount retained may be determined by individual arrangements rather than on the basis of fixed rules of the fiscal system, in which case the relationship T(H) has to be determined at the micro-level.

Table 4. Basic relationship among forms of income

	form	relationship	comment
1	Gross income	G	
2	Social insurance contributions	$S = S(G)$	
3	Gross taxable income	$\mathbf{H} = G - S$ ↖	
4	⇓	tax and SI contributions at source	$XS = H$ $XST = H – T(H)$ $XT = H + S(G) – T(H)$
5	Deductions	$D = D(H)$	↓
6	Net taxable income	$Y = H - D$	
7	Tax due	$W = W(Y)$	
8	Tax credits	$C = C(Y)$	
9	Tax paid	$\mathbf{X} = W - C$	
10	Net income	$\mathbf{N} = H - X$	

Retention at source: XS=social insurance (SI) only; XST=tax and SI; XT=tax only

Returning to the pooled income, it should be noted that while some deductions (D) may be component-specific, others apply to the pooled income as defined above, and the same is true of tax credits (C). Consequently, normally a major part of quantities Y and W, and hence also X and N, relates to the pooled income. Henceforth we will use these symbols to denote only that 'pooled' part of these quantities, since that is what is relevant in the discussion of imputation in conjunction with modelling.

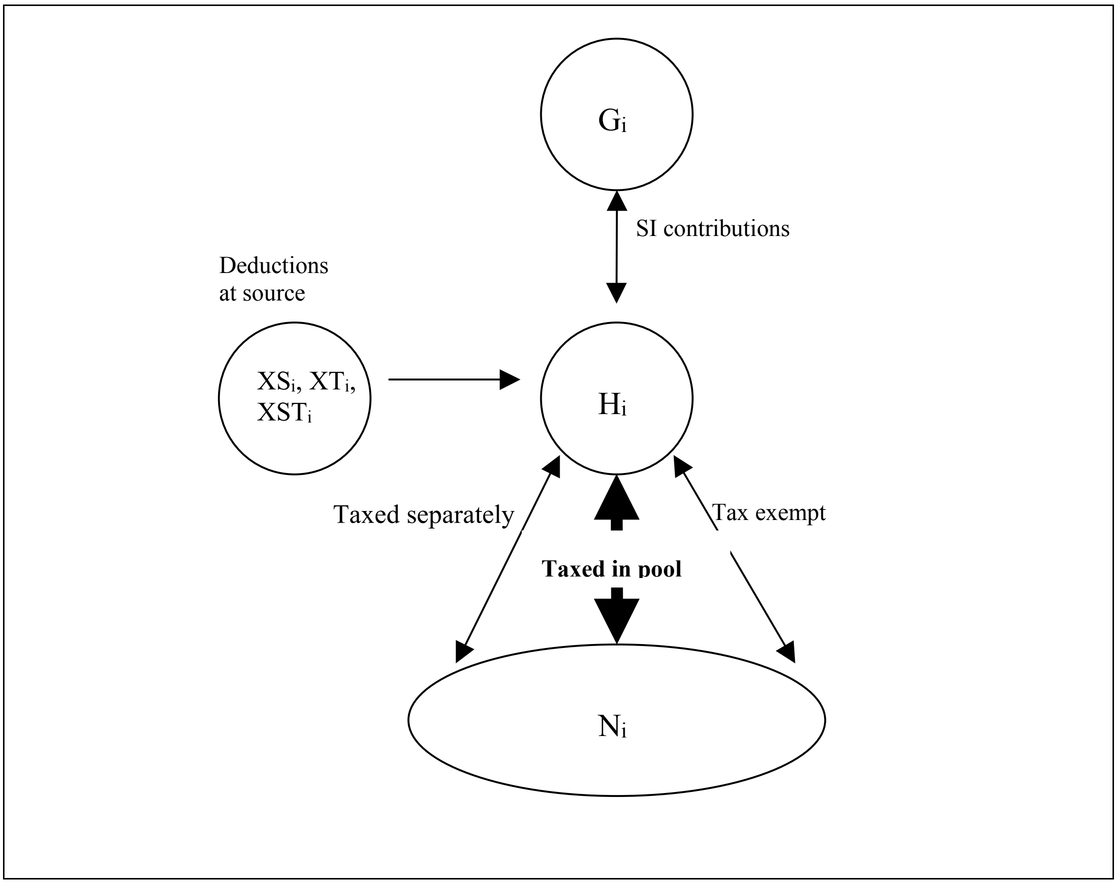
The concept of Tax Rate

We introduce the concept of tax rate (R) as an analytical measure pertaining to the pooled income of a tax unit. It is defined as the ratio of the total amount of tax due to total net taxable income: $R = W/Y$. To be precise, the tax in the numerator excludes *component-specific tax credits,* and the taxable income in the denominator excludes *component-specific deduction;* the common parts of deductions and tax credits applicable to the whole pooled income need not be excluded. By removing all known component-specific aspects, that is component-specific deductions and tax credits, R is the *common rate* which applies to all taxable income from, from whatever source, which has been pooled and subject to a common tax schedule. Parameter R has two functions.

Firstly, it provides a means for the disaggregation of tax and net income by component when required. This 'common tax rate' can be seen as a rate applying to each component individually, and not merely some average rate applicable only at the level of total income. Hence tax due (before component-specific tax credits) can be distributed among income components subject to the common tax schedule in proportion to their share in the total taxable income (after removing component-specific tax-exempt deductions): $W_i = (Y_i/Y)*W$.

Secondly, R is the parameter of the iteration in going from net to gross, as described below. Its role is even more important in the presence of missing data where modelling has to be used in conjunction with imputation as discussed in the next section.

Table 5. Basic relationship between net and gross amounts



Because of its more direct relationship with all other quantities, we take H$_i$ as the 'base' form in discussion of the model.

The modelling task

Table 5 summarises the essential modelling task. Input data may be provided for each income component (i) in any of the forms shown in the table (plus possibly a zero value), and the required model output is the quantity in both net and gross forms:

$$[0, N_i, (G_i, XT_i, XS_i, XTS_i, H_i)] \quad \Rightarrow \quad \left[0, \quad N_i.\text{and}.G_i\right]$$

Generally for a particular component, the relationship of gross taxable income H$_i$ with gross income G$_i$ or quantities like XST$_i$ after retention at source tends to be relatively simple, being dependent only on the income particular component (i) concerned, independently of other components and other persons in the same tax unit. In some cases, an iterative procedure may be involved. However, normally the iteration is very simple and converges quickly. There are no other parameters to be estimated; and the need for numerical iteration arises simply from the fact that the unknown quantity to be determined (H$_i$) appears in an implicit equation. The same applies for components which are taxed separately at a flat rate or a rate determined only by the level of income from that component, and of course also for tax exempt components. Hence the conversion of these quantities into H$_i$ can be done for the component concerned *independently of whether* *information on other components is missing and needs to be imputed.*[11] This case is shown in the first panel of Table 6.

Table 6. Calculation of H_i according to the form in which the component is specified.

Set H

given value P_i =	XS_i	$H_i = XS_i$	
	G_i	$H_i = G_i - \mathbf{S_i(G_i)}$	
	XT_i	$H_i = G_i - \mathbf{S_i(G_i)}$ where $G_i = XT_i + \mathbf{T_i(H_i)}$	Simple iteration, generally separately for each component
	XTS_i	$H_i = XTS_i + \mathbf{T_i(H_i)}$	

Set N

given value P_i =	N_i	$H_i = Y_i + \mathbf{D_i(H_i)}$ where $Y_i = [H_i - N_i + \mathbf{C_i(Y_i)}] / \mathbf{R}$	Double iteration (i) with assumed R, for each component in turn (ii) for determining R, common to all pooled components

By contrast, the relationship between H_i and N_i is more complex where income has to be pooled together over components and over persons in the tax units for the purpose of determining the amount of tax due. The modelling task can be accomplished *if there are no missing data for any of the components involved* (or any required imputations have already been performed). Going from known H_i to N_i is simpler since the relationships (the tax rules) are a function of the former. Going from given N_i to H_i required iterative solutions. The second panel of the table shows the relationship between H_i and the reported amount is 'final net' N_i. Going from N_i to H_i in fact involves a double iterative loop. The inner loop of iteration is applied with an assumed value of the parameter "tax rate" R defined above. Once this has been done for every income component in the group (including over all individuals in the same tax unit), an outer iterative loop obtains a convergent value of this parameter which is common to all those components.

Iterative procedure

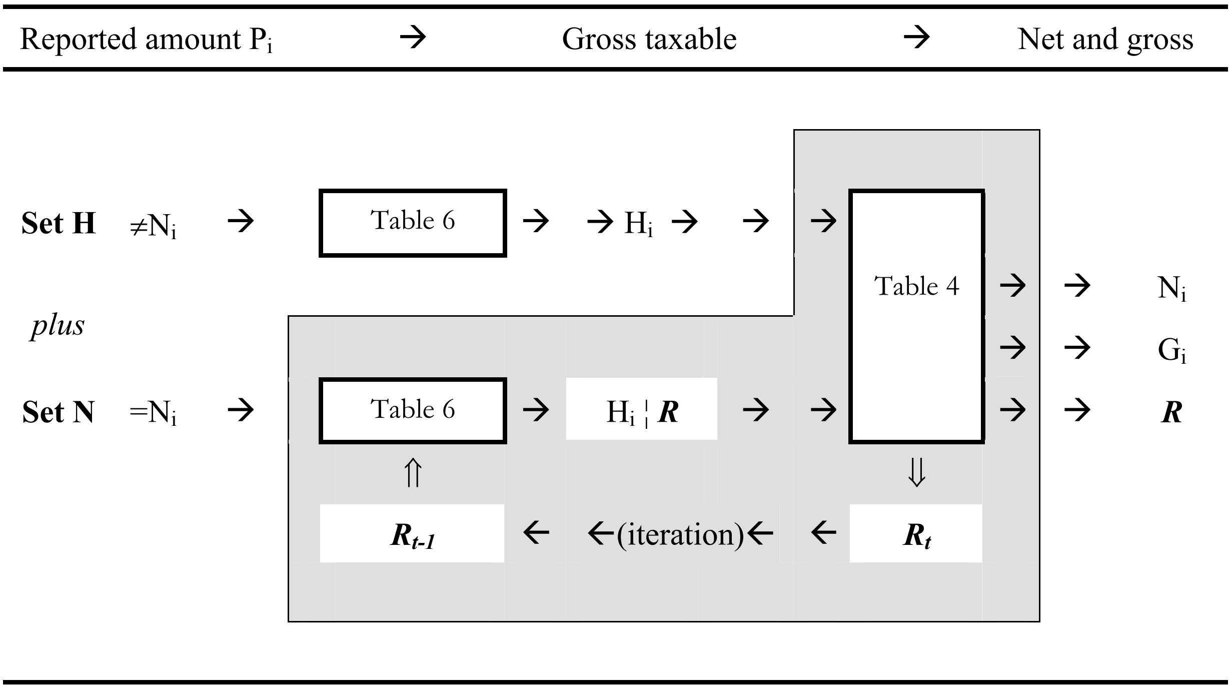
The N_i to H_i conversion process is therefore considerably more complex. Furthermore, this complexity is substantially increased in the presence of missing data, where the modelling and imputation procedures will have to be applied interactively, as discussed in the next section).

Table 7 demonstrates the common structure of the iterative procedure. As noted at the bottom of the table, the income components may be divided into two sets, say 'N' and 'H', depending on whether the amount reported is 'final net' (N_i), or is in some other form (G_i, XS_i, XT_i, XTS_i, H_i) more directly convertible to the 'gross taxable' form H_i.

The procedure may be applied as follows. The required H_i quantities for set H are computed (only once), and form an input into the iterative cycle for parameter R required for set N. The parameter is best estimated by using information on all income components from both the sets.

[11] Sometimes, dependence of the relationship on other sources of income may also be involved, but mostly these simply in the form of upper limits which may apply to certain quantities pooled over more than one component.

Table 7
Common structure of the iterative model



Set of variables N: set of income components which are subject to income tax (irrespective of whether the component is also subject to social insurance deductions), and for which the 'final net' amount (P_i=N_i) has been specified in the data collected.

Set of variables H: all other income component (not subject to tax, or for which the data has been collected in a form P_i other than the 'final net' amount).

5 Imputation and Modelling in Conjunction

The issue

In order to appreciate the interaction between imputation and modelling systems, consider matrix Y, the set of imputation/modelling variables for a set of units[12]. We denote by y_i a particular variable in the set, or where necessary, by y_{ji} that variable for unit j in the data set. In any cell (j.i) of this matrix, the value of the variable may appear in any of the following forms:

$$[0, N_i, (G_i, XT_i, XS_i, XTS_i, H_i), X]_j \quad \Rightarrow \quad [0, N_i, H_i, X]_j$$

i.e., the value is either zero (no income from the source concerned), missing (X), specified as the net amount (N_i), or in one of the various possible forms defined in Table 6, which can be converted to the form H_i even in the presence of missing values on other variables for the unit.

Net-Gross modelling applies along individual rows of the matrix Y, to one unit at a time. However, in order to 'model' i.e. construct the full information on gross and net amounts for each component, it is necessary that there be no missing values (X) in any cell of the row for the unit (that is, as noted, conversion between N_i and H_i generally requires that

[12] In practice, imputation and modelling procedures will generally involve different sets of variables. However, as noted in Section 2, for simplicity and to explain the interaction between imputation and modelling more clearly, it is assumed throughout that the two systems involve an identical set of variables.

there are no such missing values). Hence modelling requires prior imputation of the missing values.

Though imputation for missing values can be carried out along columns of the matrix, i.e. variable by variable, it invokes the whole matrix in order to take into account the correlation between variables and between units. It is necessary that for each variable, the available information for all units is in the same form (always net, always gross etc.). The form may vary from one variable to another, but should be uniform for all units within each variable. Where this is not the case in the data as collected, imputation requires prior modelling to meet this requirement.

Hence the combined imputation and modelling system has to be interactive and iterative. The whole process is described step by step below. **Step (0)** provides the starting point by using the micro-simulation model to rationalise the input information, and provides some initial information on the basis of 'complete information cases', i.e. units with no missing data. **Step (1)** concerns the conversion of the collected data for each variable into a uniform form for all responding units. It is only on this basis that imputations for missing values for the variable can be performed. On the basis of these initial imputations for the current substantive variables (income components), imputations for lagged versions of these variables and for regressor variables are performed in **Step (2)**. Our primary concern is with *current substantive variables*. Hence the results for the regressor and lagged versions from Step (2) can be considered as 'final', without the need for further iterative refinement. **Step (3)** produces refined imputations for the substantive variables, using the regressors and lagged variables imputed earlier. The procedure is applied iteratively.

Step (0). Initial data conversion and modelling

<u>0.1</u> Variables (i) are ordered according increasing proportion of missing values among applicable cases.

<u>0.2</u> Units in Y are divided into two subsets:

A: complete information set (units with information available on all variables)

B: set with missing values on one or more variables

<u>0.3</u> The starting point of this process is provided by the "complete information" set of units (set A), for which there are no missing values on any variable and hence modelling can be carried out without involving imputation, and the data for set A reduced to the form on the right, giving amount both in gross taxable (H_i) and in net (N_i) forms for each income component (i), and also the unit's "tax rate" R defined earlier:

$$\left[0, N_i, (G_i, XT_i, XS_i, XTS_i, H_i)\right]_{j\in A} \quad \Rightarrow \quad \left[0, H_i.\text{and}.N_i.\text{and}.R\right]_{j\in A}.$$

<u>0.4</u> Using the model conversion routine, any of the following data forms can all be converted to the "gross taxable income" form H_i for a unit in set B

$$(G_i, XT_i, XS_i, XTS_i, H_i) \Rightarrow H_i,$$

so that the available information for units in set B is reduced to the form

$$\left[0, N_i, H_i, X\right]_{j\in B}.$$

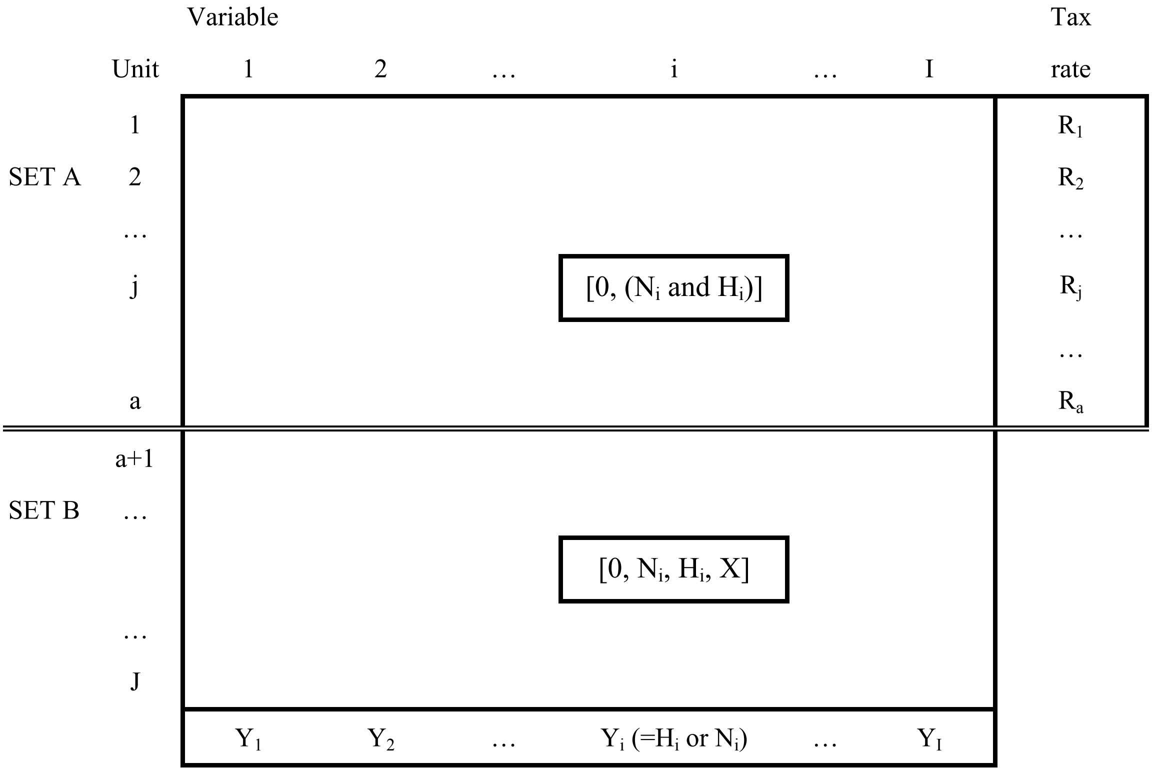
Step (1). Conversion to uniform form

1.1 For each variable (column), the predominating reporting form (Y_i) is determined. This is done on the basis of whether H_i or N_i is the more common form of reporting among units in **set B**:

$$\textbf{if count}(H_i)_{j\in B} > \textbf{count}(N_i)_{j\in B} \ \textbf{then}\ Y_i \equiv H_i \ \textbf{else}\ Y_i \equiv N_i$$

The resulting data form is shown in Table 8.

Table 8. Data form after Step (1.1)



predominating reporting form (determined by set B only)

1.2 As a starting point, we may assign *the average* of R_j values for units in set A (say, R_0) to every unit in set B. This permits the conversion of all reported values (N_i or H_i) for each variable to its "predominating form" Y_i defined above using the modelling relationships.

For set A the information in any cell is already available in the "predominating form" from Step(0)[13]:

Set A → [0, Y_i].

For set B the resulting form is

Set B → [0, Y_i, $(Y_i|R_0)$, X].

Here, form "Y_i" indicates the original information was already in the "predominating form" Y_i (i.e., already specified as N_i for a variable with Y_i=N_i, and as H_i a variable with Y_i=H_i), so that no transformation using R is required. Form "$(Y_i|R_0)$" indicates that the original information was specified in the form different from the "predominating form" Y_i

[13] While the predominant form is determined from set B only, data are converted into this form for both sets A and B.

for the variable, so that transformation to Y_i conditional on the assumed R value was required. The resulting data form is shown in Table 9.

Table 9. Data form after Step (1.2)

		Variable						Tax
	Unit	1	2	…	i	…	I	rate
	1							R_1
SET A	2							R_2
	…							…
	j				$[0, Y_i]$			R_j
								…
	a							R_a
	a+1							R_0
SET B	…							R_0
					$[0, Y_i, (Y_i \mid R_0), X]$			R_0
	…							R_0
	J							R_0
		Y_1	Y_2	…	Y_i (=H_i or N_i)	…	Y_I	

predominating reporting form (determined by set B only)

1.3 Now consider the parallel set of lagged variables $Y^{(t-1)}$. The reporting forms can be simplified as in (1) above. Then rows and columns of this matrix are arranged identically to their arrangement in table Y (steps 2 and 3 above), units and variables in the same order as Y. Also, the R_j values and the *form* Y_i from matrix Y are imposed on $Y^{(t-1)}$. (The actual values are of course different; it is the choice between N_i and H_i forms which is impose from t on t-1.) These assigned parameters are then used to transform existing $Y^{(t-1)}$ cell values into the same form as that for Y in (1.2) above. The resulting form is:

$$\text{cell values of } Y^{(t-1)} \rightarrow \left[0, Y_i^{(t-1)}, \left(Y_i^{(t-1)} \mid R_j^{(t)}\right), X\right] .$$

Here (t) indicates that the R values are taken (copied) from the current (t) data set, as is the predominating reporting form Y_i. It has been considered preferable to borrow the arrangement, parameters and data forms from the current set Y, since it is not generally appropriate to apply the tax model for the current year to past data.

Step (2). Imputation of regressor and lagged variables

2.1 The resulting information is now in a form such that the imputation procedure of Section 3 above can be applied exactly as described to produce a preliminary complete set of all current (Y), lagged ($Y^{(t-1)}$) and other auxiliary variables (U). As noted there, the procedure is to consider the union $Z \equiv \left(Y, U, Y^{(t-1)}\right)$ as a single set, order variables in it according the proportion of values missing, and perform the "triangular" form of imputation (as described in Table 2).

2.2 The above is still based on rather crude estimates of R-values for the less than complete information subset B of Y (see Table 8). With missing values removed at (1) above, we apply the micro-simulation model to produce improved estimates of R_j for all units in Y (all rows of Table 9).

2.3 These improved values of R_j are imposed on corresponding rows of $Y^{(t-1)}$ table of lagged values.

2.4 All values imputed at (2.1) above are rejected, from all sets Y, $Y^{(t-1)}$ and U.

2.5 The remaining original values in both Y and $Y^{(t-1)}$ are transformed to the predominating form Y_i using the R_j values as defined in (2.2) and (2.3) above.

2.6 The imputation procedure at (2.1) is now repeated on the resulting data set. These are taken to be the final results for the lagged and regressor variable sets.

2.7 For the main data set Y, the results from (2.6) are used to re-estimate R_j values, but the imputed values of y_i themselves are rejected. The remaining original values are transformed to the predominating form Y_i using these improved R_j values.

Step (3). Imputation of target variables Y

The data form is now the same as that in Step (1) for set Y. The only difference is that, where necessary, all of the available values have been converted to the predominating form Y_i for variable i conditional on current values of parameter R_j specific to each case j, the cell values being:

cell values of Y → $\lfloor 0, Y_i, (Y_i \mid R_j), X \rfloor$.

The first ("triangular") imputation

3.1 The first ("triangular") cycle of imputation is performed as in Table 2, without distinguishing between the two forms of known values Y_i, giving cell values in the form

cell values of Y after imputation → $\lfloor 0, Y_i, (Y_i \mid R_j), Y_i' \rfloor$.

where prime (') indicates imputed values. As explained earlier, these imputations are performed on variables in order of increasing proportion of missing values, and using imputed values for only variable previously imputed in this cycle, apart from the full U and $Y^{(t-1)}$ sets, of course.

3.2 The resulting complete set is used, with procedures of Section 4, to obtain improved values of R_j for each unit.

3.3 These R_j values are used to re-estimate the conditional values $(Y_i \mid R_j)$.

Imputation cycle for the full set

3.4 Next is performed imputation cycle for the full set, this involving a number of iterations as explained in Section 3. As explained, these imputations are performed on variables in order of increasing proportion of missing values, but using imputed values for all other variable previously imputed in any cycle, in addition of course to the full U and $Y^{(t-1)}$ sets.

3.5 The resulting complete set is used with the modelling procedure to obtain improved values of R_j for each unit.

3.6 These R_j values are used to re-estimate the conditional values $(Y_i \mid R_j)$.

3.7 The sequence is repeated a number of times.

6 Some empirical illustrations

It was noted in Section 3 that a fundamental requirement is that the imputation and modelling routines be interwoven and applied variable-by-variable in combination. We have done this by developing appropriate SAS routines to apply the imputation procedure (such as using the IVE-Ware program, used for earlier version of ECHP manual) *variable-by-variable*, rather than in one go to a whole set of variables in an automatic manner. Once the application of IVE-Ware has been broken down to the level of individual variables, it can also be used – as an important bye-product - to enhance the ECHP imputation procedures in a number of directions.

One of these is the possibility of varying the donor population from one variable to another as appropriate. Consider a population in which those with known information on an income component consist of two parts: (1) cases with known non-zero values, i.e. receiving that particular component of income; and (2) zeros, i.e. not receiving that income. Others (3) are known to receive income from that source, but the value is unknown and has to be imputed. The population in these categories differs from one variable (component) to another.

For two different such variables, the normal imputation procedure (dealing with the two variables in the same application of the IVE-Ware for instance) can of course take the permissible imputation range for each variable to be non-zero, since that fact is known to apply. However, to determine the underlying structural relationship between the imputed variable and correlates, the normal imputation procedure has to use the same population base. This in practice comes to using the whole population (1)+(2), that is, including the zeros in the donor population in each case. However, it is more appropriate that zeros among the donors are excluded from the regression since the recipient population (3) in the imputation consists only of non-zeros.

Table 10 illustrates the importance of this consideration. In this example, the information on the presence or otherwise, and the amount if applicable, for a component of net monthly employment income was available for the vast majority (about 4,500) of an ECHP sample, and was known to be non-zero but had to be imputed for a small number (30 cases). Including zeros among the donor population resulted in much less plausible set of imputed values, than that by excluding them as is *a priori* more appropriate. This is particularly striking concerning the distribution of imputed values according to quintiles of the distribution of given values. By including zeros among the donor population, 18 (60%) of the imputed cases lie at the extremities (with in 1%) of the permitted range. The distribution of imputed values by excluding zeros seems much more reasonable. The point is reinforced by observing the distribution of individual imputed values.

Table 10. Illustration of the effect of excluding non-relevant cases from the donor population

Log(income)							
	Non imputed	With zeros	Without zeros		Quintile	With zeros	Without zeros
Mean	11.84	11.20	11.82		1%	11	0
StDev	0.56	0.82	1.94		5%	14	1
Min	8.70	10.30	8.71		95%	9	4
Max	15.07	14.36	14.27		99%	7	3
N	4672	30	30				
Median	11.85						

Individual imputed values:

Obs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
With zeros	8.71	8.78	8.90	9.09	9.16	9.19	9.21	9.21	9.35	9.76	9.84	10.25	10.26	10.51	10.94
Without zeros	11.46	11.65	11.69	11.96	11.14	10.92	11.16	11.50	11.65	11.85	11.08	13.34	12.14	12.07	10.90
Obs	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
With zeros	11.10	11.31	11.49	11.58	11.67	12.47	12.77	13.02	13.14	13.89	13.89	14.03	14.12	14.24	14.27
Without zeros	11.16	12.71	11.65	10.30	12.17	14.36	11.56	11.69	11.54	12.38	12.48	11.71	11.61	13.32	11.48

Correlation between the two sets of imputed values: 0.2981

The same applies when supplementary information is available on the range within which a value must be imputed. IVE-Ware permits the specification of such ranges, a feature which has been used in ECHP application for instance. However, the problem of excluding zero donor values still remains. As an illustration, Table 11 demonstrates similar effect for another component of (log of) monthly employment income, where (only) 7 cases had to be imputed, each within a specified range. Including zeros among the donors (though it is known that the imputed cases are all non-zeros on this variable) gives results highly concentrated near the lower end of the permitted range in most cases. This is shown by the last two columns, where the imputed values are presented in terms of their location within the range, with the range rescaled as 0-1.

Table 11. Imputation within ranges: the effect of excluding zeros from the donor population

	Permitted range for imputation		imputed values		location within range	
Obs	low	high	Including zeros	Excluding zeros	Including zeros	Excluding zeros
1	8.85	11.29	8.96	9.85	0.045	0.408
2	9.96	12.99	11.77	12.81	0.596	0.938
3	10.49	13.08	10.69	11.20	0.077	0.274
4	8.01	12.06	8.02	11.99	0.003	0.985
5	8.85	12.48	8.87	11.59	0.005	0.755
6	10.49	14.25	11.32	12.59	0.220	0.559
7	11.58	12.73	11.81	12.45	0.199	0.757

The accompanying graph illustrates this clearly (X-axis is the location within the range, and Y is the imputed value in terms of log-income).

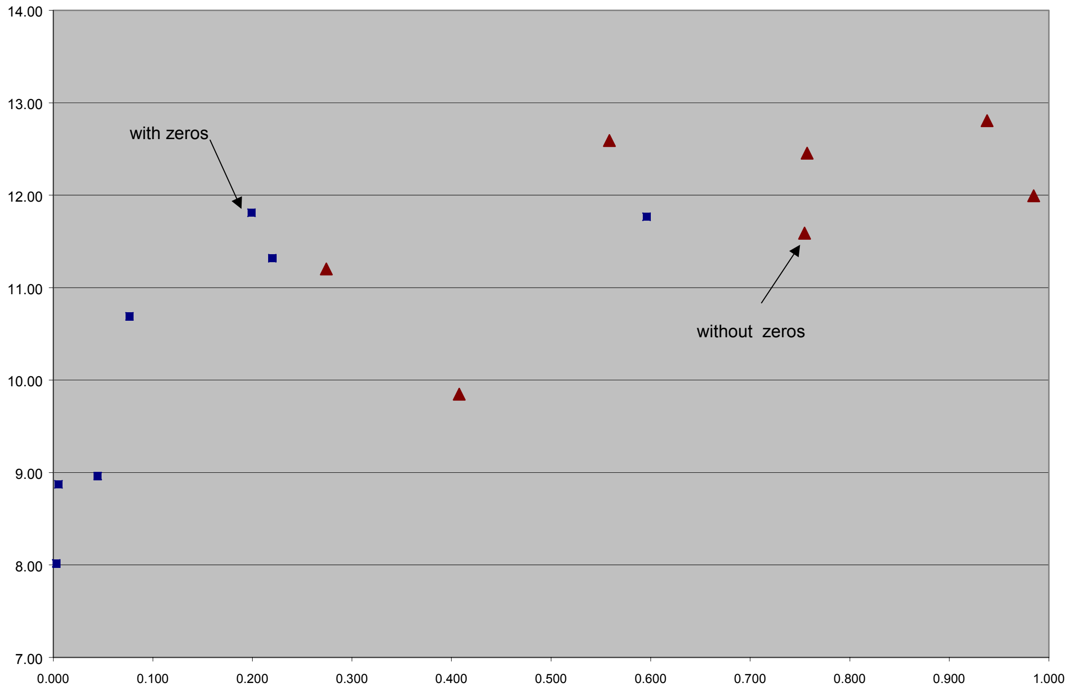


Table 12 shows the results for an imputation with a very large proportion of missing cases (the data relate to property income). Information was available for only 309 cases. Only 23 of those were in range '1', which was known to be the range for a vast majority (98.5%) of the 7,613 missing cases. An imputation based on such a weak base can hardly be expected to be very reliable. Nevertheless, the results seems to be much more consistent with the known cases when zero values among the donors are excluded.

Table 12. The effect in the case of a large number of imputations

	Observed		Imputed	
	all	in range '1'	with zeros	without zeros
Mean	10.0	4.4	1.1	5.9
Std Dev	2.5	1.6	0.9	1.1
Minimum	1.4	1.4	0.7	0.7
Maximum	16.7	6.6	9.2	16.7
Number	309	23	7,613	7,613

References

Eurostat (1996): The European Community Household Panel (ECHP): Survey methodology and implementation. Volume 1, Luxembourg: Office for Official Publication of the European Commission.

Eurostat (2001): DOC.PAN 164/2001-12, Imputation of income in ECHP.

Eurostat (2002): Commission Regulation concerning EU-Silc as regards the list of primary target variables, SILC107.

Eurostat (2002): Commission Regulation fieldwork aspects and imputation procedures, EU-Silc n.93/02

Raghunathan, Solenberger, and Van Hoewyk (2002): IVE*ware*: Imputation and Variance Estimation Software. User guide, *Survey Methodology Program. Survey Research Center, Institute for Social Research, University of Michigan*.

Verma, V., Betti, G., Ballini, F., Natilli, M., and Galgani, S. (2003): Personal income in the gross and net forms: Applications of the *Siena Micro-Simulation Model (SM2)*. Paper presented to *Statistiche per l'analisi economica*, Campobasso, 2-3 ottobre 2003.

Working Papers già pubblicati

1. E. Battistin, A. Gavosto e E. Rettore, *Why do subsidized firms survive longer? An evaluation of a program promoting youth entrepreneurship in Italy*, Agosto 1998.
2. N. Rosati, E. Rettore e G. Masarotto, *A lower bound on asymptotic variance of repeated cross-sections estimators in fixed-effects models*, Agosto 1998.
3. U. Trivellato, *Il monitoraggio della povertà e della sua dinamica: questioni di misura e evidenze empiriche*, Settembre 1998.
4. F. Bassi*, Un modello per la stima di flussi nel mercato del lavoro affetti da errori di classificazione in rilevazioni retrospettive*, Ottobre 1998.
5. Ginzburg, M. Scaltriti, G. Solinas e R. Zoboli, *Un nuovo autunno caldo nel Mezzogiorno? Note in margine al dibattito sui differenziali salariali territoriali*, Ottobre 1998.
6. M. Forni e S. Paba, *Industrial districts, social environment and local growth. Evidence from Italy*, Novembre 1998.
7. B. Contini*, Wage structures in Europe and in the USA: are they rigid, are they flexible?*, Gennaio 1999.
8. B. Contini, L. Pacelli e C. Villosio, *Short employment spell in Italy, Germany and Great Britain: testing the "Port-of-entry" hypothesis*, Gennaio 1999
9. B. Contini, M. Filippi, L. Pacelli e C. Villosio*, Working careers of skilled vs. unskilled workers*, Gennaio 1999
10. F. Bassi, M. Gambuzza e M. Rasera, *Il sistema informatizzato NETLABOR. Caratteristiche di una nuova fonte sul mercato del lavoro*, Maggio 1999.
11. M. Lalla e F. Pattarin, *Alcuni modelli per l'analisi delle durate complete e incomplete della disoccupazione: il caso Emilia Romagna*, Maggio 1999.
12. A. Paggiaro*, Un modello di mistura per l'analisi della disoccupazione di lunga durata*, Maggio 1999.
13. T. Di Fonzo e P. Gennari*, Le serie storiche delle forze di lavoro per il periodo 1984.1-92.3: prospettive e problemi di ricostruzione*, Giugno 1999.
14. S. Campostrini, A. Giraldo, N. Parise e U. Trivellato*, La misura della partecipazione al lavoro in Italia: presupposti e problemi metodologici di un approccio "time use"*, Ottobre 1999.
15. A. Paggiaro e N. Torelli, *Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro*, Ottobre 1999.
16. A. D'Agostino, G. Ghellini e L. Neri*, A Multiple Imputation Method for School to Work Panel Data*, Ottobre 1999.
17. G. Betti, B. Cheli e A. Lemmi, *Occupazione e condizioni di vita su uno pseudo panel italiano: primi risultati, avanzamenti e proposte metodologiche*, Ottobre 1999.
18. B. Anastasia, M. Gambuzza e M. Rasera*, La durata dei rapporti di lavoro: evidenze da alcuni mercati locali del lavoro veneti*, Marzo 2000.
19. F. Bassi, M. Gambuzza e M. Rasera*, Struttura e qualità delle informazioni del sistema NETLABOR. Una verifica sui dati delle Scica delle province di Belluno e Treviso*, Marzo 2000.
20. N. Rosati, *Permanent and Temporary Inequality in Italy in the 1980s and 1990s*, Marzo 2000.
21. G. Betti, B. Cheli e A. Lemmi, *Analisi delle dinamiche di povertà e disoccupazione su uno pseudo panel italiano*, Marzo 2000.
22. A. D'Agostino, G. Ghellini e L. Neri, *Modelli statistici per l'analisi dei comportamenti di transizione scuola lavoro*, Marzo 2000.

23. A. Paggiaro e U. Trivellato*, Assessing the effects of the "Mobility List" programme in an Italian region: do (slightly) better data and more flexible models matter?,* Marzo 2000.
24. F. Bassi, M. Gambuzza, M. Rasera e E. Rettore, *L'ingresso dei giovani nel mercato del lavoro: prime esplorazioni dall'archivio Netlabor*, Giugno 2000.
25. A. D'Agostino, G. Ghellini e L. Neri, *Percorsi di ingresso dei giovani nel mercato del lavoro*, Giugno 2000.
26. E. Battistin, E. Rettore e U. Trivellato, *Measuring participation at work in the presence of fallible indicators of labour force state,* Giugno 2000.
27. E. Battistin e E. Rettore, *Testing for the presence of a programme effect in a regression discontinuity design with non compliance*, Novembre 2000.
28. A. Ichino, M. Polo e E. Rettore, *Are judges biased by labor market conditions?*, Novembre 2000.
29. N. Rosati, *Further results on inequality in Italy in the 1980s and the 1990s*, Aprile 2001.
30. F. Bassi, M. Gambuzza e M. Rasera, *Imprese e contratti di assunzione: prime analisi da Netlabor*, Novembre 2001.
31. F. Bassi e U. Trivellato, *Gross flows from the French labour force survey: a reanalysis*, Novembre 2001.
32. A. Borgarello e F. Devicienti, *Trend nella distribuzione dei salari italiani 1985-1996*, Novembre 2001.
33. B. Contini, *Earnings mobility and labor market segmentation in Europe and USA: preliminary explorations*, Novembre 2001.
34. B. Contini e C. Villosio, *Job changes and wage dynamics*, Novembre 2001.
35. A. Borgarello, F. Devicienti e C. Villosio, *Mobilità retributiva in Italia 1985-1996*, Novembre 2001.
36. L. Pacelli, *Fixed term contracts, social security rebates and labour demand in Italy*, Novembre 2001.
37. B. Anastasia, M. Gambuzza e M. Rasera, *Le sorti dei flussi: dimensioni della domanda di lavoro, modalità di ingresso e rischio disoccupazione dei lavoratori extracomunitari in Veneto*, Novembre 2001.
38. N. Torelli e A. Paggiaro, *Estimating transition models with misclassification*, Novembre 2001.
39. G. Barbieri, P. Gennari e P. Sestito, *Do public employment services help people in finding a job? An evaluation of the italian case*, Novembre 2001.
40. A. Giraldo, E. Rettore e U. Trivellato, *The persistence of poverty: true state dependence or unobserved heterogeneity? Some evidence form the Italian survey on household income and wealth*, Novembre 2001.
41. A. Giraldo, E. Rettore e U. Trivellato, *Attrition bias in the bank of Italy's survey on household income and wealth*, Novembre 2001.
42. F. Devicienti, *Estimating poverty persistence in Britain*, Novembre 2001.
43. B. Contini, F: Cornaglia, C. Malpede, E. Rettore, *Measuring the impact of the Italian CFL programme on the job opportunities for the youths,* Novembre 2002.
44. E. Battistin, E. Rettore, *Another look at the regression discontinuity design,* Novembre 2002.
45. U. Trivellato, A. Giraldo, *Assessing the 'choosiness' of the job seekers. An exploratory approach and evidence for Italy,* Novembre 2002.
46. E. Rettore, U. Trivellato, A. Martini, *La valutazione delle politiche del lavoro in presenza di selezione: migliorare la teoria, i metodi o i dati?,* Novembre 2002.
47. B. Anastasia, D. Maurizio, *Misure dell'occupazione temporanea: consistenza, dinamica e caratteristiche di uno stock eterogeneo,* Novembre 2002.
48. S. Bragato, F. Occari, M. Valentini, *I problemi di contabilizzazione dei lavoratori extracomunitari. Una verifica nelle province di Treviso e Vicenza,* Novembre 2002.

49. A. Borgarello, F. Devicienti, *Trends in the Italian earnings distribution, 1985-1996,* Novembre 2002.
50. V. Verma, G. Betti, *Longitudinal measures of income poverty and life-style deprivation,* Novembre 2002.
51. F. Devicienti, *Downward nominal wage rigidity in Italy: evidence and consequences,* Novembre 2002.
52. D. Favaro, S. Magrini, *Gender wage differentials among young workers: methodological aspects and empirical results*, Settembre 2003.
53. R. Canu, G. Tattara, *Quando le farfalle mettono le ali. Osservazioni sull'ingresso delle donne nel lavoro dipendente,* Settembre 2003.
54. V. Verma, G. Betti, F. Ballini, M. Natilli, S. Galgani, *Personal income in the gross and net forms: applications of the Siena micro-simulation model (SM2),* Settembre 2003.
55. G. Betti, V. Verma, M. Natilli, F. Ballini, *Statistical imputation in conjunction with micro-simulation of income data,* Settembre 2003.

Le richieste di working papers vanno indirizzate alla Sig.ra Patrizia Piacentini, Dipartimento di Scienze Statistiche, via C. Battisti 241-243, 35121 Padova, oppure all'indirizzo e-mail patrizia.piacentini@unipd.it.