Immervoll, Herwig; Lindström, Klas; Mustonen, Esko; Riihelä, Marja; Viitamäki, Heikki

Working Paper

Static data ageing techniques: Accounting for population changes in tax-benefit microsimulation models

EUROMOD Working Paper, No. EM7/05

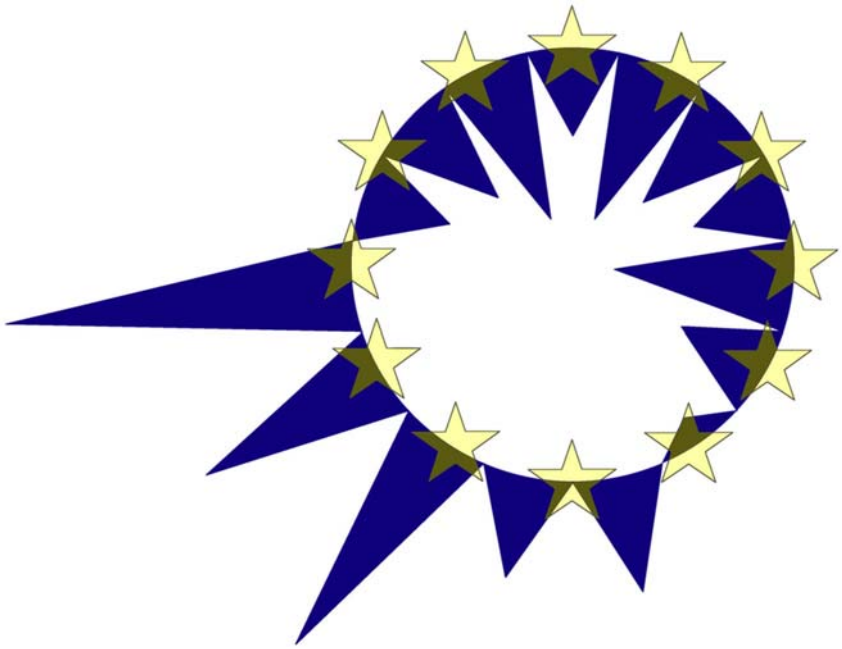
Provided in Cooperation with:
Institute for Social and Economic Research (ISER), University of Essex

Suggested Citation: Immervoll, Herwig; Lindström, Klas; Mustonen, Esko; Riihelä, Marja; Viitamäki, Heikki (2005) : Static data ageing techniques: Accounting for population changes in tax-benefit microsimulation models, EUROMOD Working Paper, No. EM7/05, University of Essex, Institute for Social and Economic Research (ISER), Colchester

This Version is available at:
https://hdl.handle.net/10419/68980

EUROMOD

WORKING PAPER SERIES



EUROMOD Working Paper No. EM7/05

STATIC DATA "AGEING" TECHNIQUES. ACCOUNTING FOR POPULATION CHANGES IN TAX-BENEFIT MICROSIMULATION MODELS.

Herwig Immervoll, Klas Lindström , Esko Mustonen, Marja Riihelä and Heikki Viitamäki

March 2005

Static data "ageing" techniques.
Accounting for population changes in tax-benefit microsimulation models.

Herwig Immervoll

Klas Lindström

Esko Mustonen

Marja Riihelä

Heikki Viitamäki[1]

Abstract

Tax-benefit microsimulation models are frequently used to analyse the distributional, budgetary and behavioural effects of social and fiscal policies in a period *t** using data from some previous period that have been adjusted ("aged") to approximate the population in period *t**. This paper considers which types of data adjustments are necessary and appropriate and discusses issues and limitations that affect the scope and interpretation of results based on aged data. It presents a simple conceptual framework for thinking about different types of data adjustments and illustrates the mechanics of ageing procedures in a case study using the EUROMOD tax-benefit model in conjunction with detailed Finnish household micro-data from two periods (1996 and 1998). The case study evaluates the performance of one particular ageing technique by comparing results from the 1998 dataset with those derived from aged 1996 data.

JEL Classification: C81; C88
Keywords: Data Ageing; Re-weighting; Microsimulation; Tax-Benefit Model

1. Introduction

Tax-benefit microsimulation models are frequently used to analyse the distributional, budgetary and behavioural effects of social and fiscal policies in a period *t** using micro-data from some another (usually earlier) period *t* that have been adjusted ("aged") to approximate the population in period *t**. There are several possible reasons why one may do this. The most common one is that data for period *t** are not available, for instance in the case where tax burdens and benefit entitlements are to be projected using some future policy scenario. In fact, assessing the effects of future policy reforms is the most common use of tax-benefit models. In such a situation, one has no choice but to use the most appropriate and up-to-date micro-data available. Given additional information about population differences between the data period *t* and the

[1] **Acknowledgements:** This paper was written as part of the MICRESA (Micro Analysis of the European Social Agenda) project, financed by the Improving Human Potential programme of the European Commission (SERD-2001-00099). We are grateful to Tim Callan and Joanna Gomulka for helpful comments and suggestions. The views expressed in this paper, as well as any errors, are the responsibilities of the authors. In particular, this applies to the interpretation of EUROMOD model results and any errors in its use. EUROMOD is continually being improved and updated and the results presented here represent work in progress. EUROMOD relies on micro-data from twelve different sources for fifteen countries. This paper only uses the Finnish dataset which is based on Income Distribution Survey made available by Statistics Finland. Statistics Finland does not bear any responsibility for the analysis or interpretation of the data reported here.

period of interest *t**, it is then necessary to devise a method for approximating the population of period *t** using the available mico-data and the "additional information" as a starting point.[2]

The purpose, scope and quality of such "ageing" techniques depend on a number of factors that are very specific to the research task at hand. As a result, it can be difficult to discuss the different methods in a systematic way and there have been few attempts to do so.[3] This is unfortunate since the usefulness of tools such as tax-benefit models depends crucially on the quality of the micro-data used as input. It is therefore essential that the characteristics of these data, including any adjustments, are well understood.

This paper discusses the feasibility and implications of techniques to approximate the population of period *t** using the period *t* data as a starting point. It considers which types of data adjustments are necessary and appropriate and discusses issues and limitations that affect the scope and interpretation of results based on aged data. We focus specifically on "static" ageing techniques which, for the purpose of this paper, are defined as methods attempting to align the available micro-data with other known information (such as changes in population aggregates, age distributions or unemployment rates), without modelling the *processes* that drive these changes (e.g., migration, fertility, or economic downturn). In particular, we focus on a technique known as "re-weighting" (altering the "weights" of different observations in the data) and attempt to position this approach relative to other data adjustment methods that can be substitutes or complements to re-weighting.

The paper is structured as follows. Section 2 discusses how differences between populations in periods *t* and *t** can be accounted for in theory. Drawing on this discussion, Section 3 then explores how the quality of a particular "ageing" technique might be evaluated in practice and presents a case study using Finnish household data in conjunction with the EUROMOD tax-benefit model. We run EUROMOD on actual period *t** data and compare the results with a scenario where data have been artificially aged using period *t* data as a starting point. Results are presented and discussed in Section 4. A final section concludes.

2. What happens between *t* and *t** and how can differences be accounted for in theory?

Tax-benefit models need to produce a "good" approximation of taxes and benefits payable in a particular period. It is of course possible that, for the purpose of achieving such a "good" approximation, data from period *t* are already sufficiently close to any "real" period *t** data. But it is useful to ask whether we can improve on this to any meaningful extent by further enhancing the degree to which data available for modelling describe the target population in period *t**.

In thinking about the correlation between period *t* data and period *t** populations, it is useful to separate the factors which determine how representative sample S_t taken in period *t* will be of population P_{t*} in period *t**. One can, for this purpose, look separately at

1. the degree to which S_t is representative of P_t; and
2. the processes causing P_t to differ from P_{t*}.

One intuitive interpretation of the observations in the sample is that they each "represent" a certain number of population members in the sense that the variable values recorded for a given observation approximate the

[2] In certain cases, no particular adjustments of period *t* data are needed. Users of tax-benefit models can be interested in policy changes between period *t* and t^* in relation to a population "frozen" at a particular point in time *t*. The aim of such an evaluation is generally to measure the "pure" mechanical policy effects independently of changes in the underlying population (which may occur independently or as consequence of policy measures).

[3] Creedy (2004) provides a recent discussion of certain aspects of data "ageing". The approach is largely descriptive focussing on one particular instance of data ageing rather than an evaluation of this technique or a more general discussion of conceptual issues.

characteristics of a certain fraction of the population. We will denote the group in the population "represented" by observation i in sample S_t as $G_{t,i}$. The changes in the population (point 2 above) can then be broken down further into

2a. Processes altering the average value of a variable in group $G_{t,i}$. Given the conceptualisation of each observation i in the sample as representing $G_{t,i}$ this translates into changing the value of the relevant variable of observation i in sample S_t.

2b. Processes causing the composition of $G_{t,i}$ to change. That is, some population members who have "fitted into" group $G_{t,i}$ may, due to changes in the population structure, be better represented by another group in population P_{t^*} (or may no longer be part of the population in time period t^* at all). Similarly, group $G_{t^*,i}$ may encompass population members which were not part of $G_{t,i}$.

Each of these factors will be discussed in turn.

2.1 Post-stratification: making S_t statistically representative of P_t

While post-stratification is not an approach to "age" data it is worth considering in some detail here due to instructive parallels with the re-weighting approach to be discussed later on. Initially, a degree of representativity of S_t is achieved by introducing weights. These simply depend on the selection probabilities of each observation in the sample (and, thus, the sample design). Further modifications to these weights are then often made by the original data providers in order to exploit any knowledge about differential non-response.

Starting from the resulting "original weights", researchers may be able to exploit other available information to improve the representativity of the sample. If one knows the true number of population members with certain characteristics (the post-strata) then the original weights can be forced to correspond to these control totals. This is known as post-stratification due to some parallels to stratification, where the sample is designed in such a way as to guarantee a certain number of draws from certain population groups (the strata).[4] If control totals are available for a sufficient number of population sub-groups then it may be possible to also align certain distributional characteristics of sample and target population (such as the number of people in a certain sub-group belonging to each earnings decile).

For post-strata whose weighted totals (using the unadjusted "original" weights) fall short of (exceed) the control totals, weights are adjusted upwards (downwards) while the weights of other post-strata are adjusted in the opposite direction in order to keep weighted sample *totals* unaffected. For the upward adjustments, this amounts to the assumption that those population members in the relevant post-stratum who have not been included in the sample (which is the reason for the mismatch between sample total and control total) are reasonably well represented by population members in the post-stratum who do show up in the sample. Clearly, this assumption may be wrong. However the critical point is that, to the extent that the variables of interest to the data user (V) are correlated with variables defining the post-strata (X), it can be an *improvement* vis-à-vis not making any adjustments. This is because the assumption implicit in *not* taking into account available true control totals is that the non-observed members of the post-stratum are well-represented by the sample at large. The averages of variables correlated with X will be closer to their true period t^* value if weights are adjusted in accordance with the control totals. However, depending on the precise re-weighting approach used, the distributions of these variables may become distorted (and may be

[4] Similar to stratification, this process can increase the precision by which statistics of given variable can be computed (the gain in precision for a given variable x will, again similar to stratification, depend on how homogeneous the values of x are within the strata as compared to the inter-strata variation).

further from the "true" distribution than if unadjusted weights were used). In addition, re-weighting can cause distortions (of averages *and* distributions) of variable that are not correlated with X or are correlated with more than one variable X simultaneously.

It is useful to note that a mismatch of control totals and post-strata sample totals can be due to random sampling error; sampling bias or measurement error (in either the survey data, the control totals, or both). The importance of random sampling error will depend on the cell size of the selected post-stratum and can be quantified by computing the standard error of the estimated sample total. Sampling bias, on the other hand, will occur if the sample design is inadequate for the purpose of drawing a representative sample of the population of interest (e.g., if the sampling frame is different from population P or if differential non-response is not fully accounted for). Re-weighting will force the number of observations in each post-strata to match the control-totals. Note that if the reasons for the original mismatch is measurement error, then the alteration of changing weights is an attempt to correct one error by introducing another one. If, for instance, the true wage of observation i in the sample was 100 while the reported wage is 90 and we know from the true control totals that no population member does in fact have a wage of 90 then re-weighting will force the weight of i towards zero (and thereby also surrender information on any other characteristics of observation i which may be measured without error). Clearly, the appropriate adjustment in this case would be to the reported wage level rather than the weight attached to the observation.

2.2 Adjusting variable values of individual observations ("uprating")

If the average value $\bar{v}$ of a variable observed in a given group $G_{t,i}$ changes between t and t^* then for observation i to still be "representative" of $G_{t^*,i}$ in period t^*, this change will need to be reflected in the variable value recorded for that observation. For monetary variables, this can be achieved by "uprating" (ie, inflating or deflating) each value by an appropriate index describing how the value of the variable, averaged across the population group represented by i, has behaved between t and t^*. In doing so, it is important to separate changes in the average value of the variable averaged across members of $G_{t,i}$ from changes in the number of population members with certain variable values. To illustrate, let the wage and region for observation i be recorded as 100 and 1 in S_t and let observation i be the only observation drawn from region 1 who has non-zero wages. If the average wage level in region 1 rises to 110 and, at the same time, employment rates decrease by 10% then we would want the wage of observation i to be uprated by the change in average wage (+10%) rather than the change in total wages earned in region 1 (0%). The fact that employment rates in region 1 have decreased is one of the processes defined above as type 2b and should be dealt with separately.[5]

Of course, indices capturing the change in variable values separately for each group $G_{t,i}$ (and thus each observation i) will often not be available. One will, for instance, usually see more than one observation with non-zero wages in a given region and if there is only one index of average earnings available for the region as a whole then the same index will need to be applied to all wage earners of that region in the sample. In other words, we cannot hope to perfectly replicate the distribution of all relevant variable changes occurring between t and t^*.

2.3 Adjusting the relative sizes of sub-populations ("re-weighting")

Just as post-stratification aligns the sum of weights of a given population sub-group with external control totals, re-weighting can be used to align weighted frequencies of subgroups in sample S_t with external control totals relating to time period t^*. While the process of uprating discussed in the previous section aims at

[5] See Section 2.3 below. The weight of i could, for instance, be decreased by 10% while increasing the weights of all other working-age observations from region 1.

correcting the information of observations in sample S_t so that they are still approximately representative of equivalent population members in P_{t*}, re-weighting sample S_t can be used to correct for the difference in probabilities of drawing an observation i (which is already part of S_t) if another sample were to be taken of population P_{t*}. When moving from period t to t*, it is possible and, indeed, likely that both the probability of drawing observation i and the average values of variables in the group "represented" by this observation will have changed. To exploit all available information, it will generally be desirable to use both uprating and re-weighting.[6]

Clearly, there are many ways in which a sample could be weighted in order to match a given set of control totals. How then should the weights be re-computed? Since no exact solution exists to the re-weighting problem and since the original weights provided in the dataset prior to any re-weighting contain a great deal of information about the population, a natural approach is trying to achieve the control totals by changing the existing weights as little as possible. There are different ways of measuring the distance of a new weight to the original one. This is formalised as a distance function which the re-weighting algorithm then aims to minimise.[7]

Similar to uprating, one would ideally want separate pieces of reference information for each observation i. This would permit resetting the size of each group $G_{t,i}$ to a value appropriate for time period t* and thus allow for differential changes of group sizes across all group (the problem of finding these new weights separately for each i could be thought equivalent to resetting the weights of earlier observations in a later wave of panel data in a situation of zero attrition). Again, the information available on the true size of population sub-groups is likely to be much more limited allowing only a relatively small number of population sub-groups (each encompassing a number of groups $G_{t,i}$) to be re-sized according to the external control total.

3. Evaluating the performance of static ageing techniques

A schematic illustration of the ageing procedure is provided in Figure 1. The sample S_t is composed of j observations characterised by a set of variables V and X. Each sample observation represents a population sub-group with the weight of the observation corresponding to the number n_j of population members p represented by that observation. Different observations can be combined into i groups G with Variables X (e.g., region or age-group) used for grouping. Variables V are those relevant for tax-benefit calculations (e.g., incomes). Note that V and X may partly overlap (for instance, the number of children in a family may be used to define strata but will also have an impact on tax liability and benefit entitlements of this family).

The ageing process is supposed to reflect changes in the population between periods t (shown on the left) and t^* (on the right) so that the same sample S_t can, after certain adjustments, be used as an approximation of the population in period t*. As explained above, we can distinguish two types of changes between t and t^*. Changes in the group G averages of variables V (vectors $\overline{v}_t ; \overline{v}_{t^*}$) and changes in the number of population members (p in period t and q in period t*) that each group G represents.

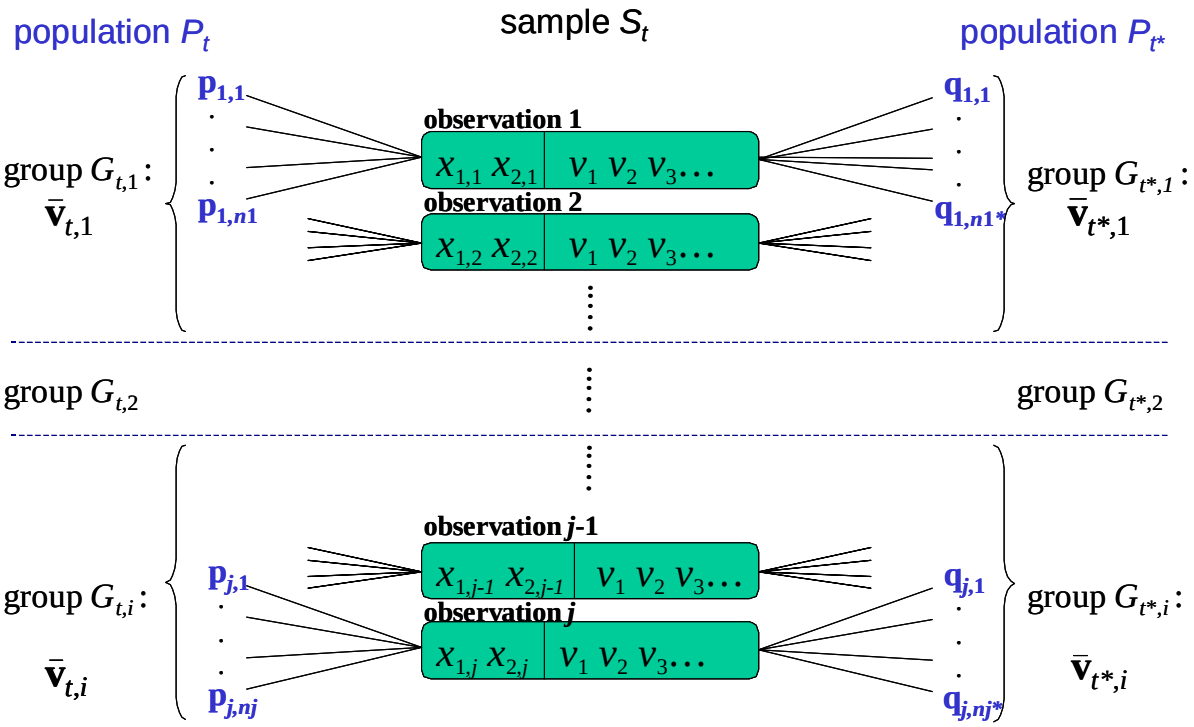
A straightforward way of evaluating the performance of static ageing techniques in the context of tax-benefit modelling is to use existing data sets from period t and t* in combination with a tax-benefit model containing policy rules for period t*. The simulation results of two scenarios, the results based on "aged" data versus results based on "real" period t* data, could then be compared. A natural approach to assess the quality of the aged dataset would then examine the types of output normally reported by tax-benefit models. This includes both aggregate and distributional statistics of tax burdens, benefit entitlements and household incomes.

---

[6] In that order if the variable to be uprated is also used for defining sub-groups $G_{t,i}$ and $G_{t*,i}$.

[7] Applying a different distance function can produce different weights but most common distance functions tend to produce comparable results (see Deville and Särndal, 1992).

Figure 1. Data ageing: Relationship between a sample drawn in period *t* and a population in period *t**.



3.1. A case study: Re-weighting Finnish household data using a "minimum distance" algorithm

In this section we present results from a re-weighting exercise using Finnish household data relating to one period in order to approximate later waves of the same dataset. The tool for re-weighting we use for this purpose is the *Clan97* software.[8] The sample design is important because *Clan97* is designed to work in conjunction with a limited set of designs such as pps-sampling, various type of network sampling and two-phase sampling schemes for stratification. *Clan97* is able to align sample weights with given control totals employing a method which changes weights as little as possible to achieve a set of target control totals. In this particular instance, we use a linear distance function in order to minimise the discrepancy between new and original weights.

While appealing in theory, the method of re-weighting can be problematic in practice. The minimum distance criterion ensures that weights are changes as little as possible in order to achieve a given set of control totals but it does not limit the alteration of weights in any absolute sense. In particular, achieving matching control totals for the chosen set of variables X can potentially distort the population total of variables V (and hence the aggregate statistics of tax and benefit amounts that depend on V). In addition re-scaling the weights to produce a given set of frequency totals can significantly alter the distributional information contained in the data (and hence the distribution of taxes, benefits and incomes calculated by the tax-benefit model). Distributional distortions can affect

- marginal distributions of the variables X one is controlling for;
- joint distributions between different variables X;

[8] Clan97 is a program designed to compute point- and standard error estimates in sample survey (Anderson and Nordberg, 1998). Clan97 is written in the SAS language and has been developed by Statistics Sweden.

- joint distributions between different variables V; and
- joint distributions between variables X and variables V.

3.1.1. Data sources

We aim to illustrate these effects by adjusting the sample weights contained in the 1996 wave of the Finnish Income Distribution Survey (IDS) so to align relevant totals with control totals derived from the 1998 wave of the same dataset. The re-weighted 1996 dataset is then compared with the "real" 1998 data.

The IDS is based on two-phase stratified sampling. The two-phase sampling can be applied when auxiliary information from administrative registers can be used to target sub-groups that are of particular interest. A rotating panel design is used so that each household is in the sample for two consecutive years. The sampling frame is the Finnish Central Population Register, so it includes the entire population. People living in institutions are, however, excluded. The first master sample is taken from the Central Population Register. Dwelling units are then constructed by adding all other people living in the same address with sampling person. The master sample is merged with the tax register and socio-economic class is derived for each individual. The final sample is then drawn using stratification according to socio-economic class of reference person (person with highest taxable income). Inclusion probabilities depend on the number of persons older than 15 years in household. Sampling rates differ between strata with groups such as entrepreneurs and high-income wage-earner households being assigned higher inclusion probabilities. The final sample contains around 10,000 households. Household weights are based on the inverse probability of inclusion in the sample. A non-response correction is applied using sample information on response rates in each stratum. Finally, the weights are calibrated (post-stratified) to match estimates of population structure and income totals.

3.1.2. Control totals

Information on the control totals of variables V is derived directly from the 1998 data. Using the same data source for the control totals and for the data to be re-weighted minimises the influence of any measurement issues that one may otherwise encounter and thus avoids clouding our view of the re-weighting mechanics we are interested in. It is nevertheless important to verify that the definitions of control variables are consistent in the two periods and we have made any relevant adjustments as required.

A number of technical problems are less likely if the number of control variables is kept low. The spread of weights is likely to increase with the number of variables that are controlled for. In extreme cases, the re-weighting algorithm may produce negative weights which is, of course, undesirable. We chose four types of control variables: gender, age, region and the number of recipients of different market and benefit incomes. While the re-weighting algorithm is used to adjust household weights, all our control totals are based on numbers of individual rather than households. That is, the re-weighting algorithm is solving for the "minimum distance" set of household weights that produces a given number of individual-based control totals. Since we are aiming to produce a good approximation of the 1998 population for the purpose of tax-benefit modelling, we have chosen categories that make a difference to tax burdens and benefit entitlements. For instance, in deciding on the age groups, we have chosen age cut-offs that are used in the legal entitlement rules governing the various policy instruments in 1998 (such as child benefits or old-age pension, indicated in brackets in the list below). These considerations have resulted in the following set of control frequencies (all control totals are shown in Table 1):

Eight age groups by gender:

- Younger than 3 years (age cut-off for child home care allowance)
- 3-6 years (day care payments)

- 7-16 years (child benefits)
- 17-27 years (student grants, student deduction, students housing benefit)
- 28-42 years
- 43-57 years
- 58-64 years (early retirement)
- Older than 64 years (old age pension)

By taking the number of persons in each age/gender group from the 1998 data, we also control for differences in population size between 1996 and 1998.

Four different types of employment status (identified by the receipt of income from the relevant source and with one particular individual counted in more than one category if she receives more than one type of relevant income at the same time).

- wage-earners
- unemployed (without separating to different benefit groups)
- farmers
- self-employed

Twelve regions (Finnish provinces):

- Uusimaa (south): category 1.
- Turku and Pori (south west): category 2.
- Häme (south central): category 4.
- Kymi (south east): category 5.
- Mikkeli (north central): category 6.
- North Karelian (north east): category 7.
- Kuopio (north east): category 8.
- Central Finland (central): category 9.
- Vaasa (north): category 10.
- Oulu (north): category 11.
- Lapland (north): category 12.
- Åland Islands: category 13.

4. Results

We used the calibrated 1996 weights, after adjustment for non-response, as the starting point for the re-weighting exercise.[9] By definition, re-weighting the 1996 data to match 1998 control totals causes all frequencies that are being controlled for to be perfectly aligned with the relevant target numbers from the 1998 dataset. Before turning to the effect of re-weighting on simulated taxes and benefits in a microsimulation model, it is, as a first step, interesting to see to what extent population characteristics that

[9] Results from a different scenario using the non-calibrated survey-design weights are available on request.

that are not, or only partly, controlled for are affected. We first examine a number of characteristics in isolation and then examine how the re-weighting process alters the joint distribution of relevant variables.

4.1. Results 1: Frequency totals for characteristics that were not controlled for

Tables 2a to 2e compare the size of relevant population groups between the 1996, 1998 and re-weighted 1996 datasets. The following characteristics were considered: a) family type; b) recipients of different income sources; c) employment status; and d) numbers of wage earners, farmers, self-employed and unemployed persons by region. Column 6 shows the relative change in the number of observations between the 1996 and 1998 datasets while column 7 shows the ratio between the number of observations in the re-weighted 1996 data and those in the actual 1998 dataset. The re-weighting process leads to a better correspondence with the 1998 data if the ratio in column 7 is closer to 100 than the ratio in column 6. The ideal outcome would be to have a perfect match with values 100 in column 7.

As one way of judging the distance from the 1998 target values, we have computed confidence intervals based on re-weighted 1996 data (column 4). Our hypothesis H_0 is that re-weighted 1996 results in column 4 are equal to values of 1998 data. Bold values are shown where we cannot reject H_0, i.e., where the re-weighting procedure performs well in the sense that it produces frequencies that are not significantly different from the 1998 value. It is important to note, however, that the 1998 value is derived from an independent sample and will be subject to statistical uncertainty of a similar magnitude.[10]

Table 2a compares the number of households corresponding to different family typologies. This is of interest since all control totals used for re-weighting relate to individual rather than household characteristics. The degree to which the shape of the 1998 population in terms of family typologies can be approximated will therefore depend on the correlation of family status with the individual-based variables used for the control totals. Since the total number of households is correlated with the total number of individuals (which is being controlled for), we see, compared to the unadjusted 1996 data, an improved match with the 1998 figures. However, the overall number of household derived from the re-weighted 1996 data is still significantly (in a statistical sense) different from the 1998 value. Looking at the different family types, we find that three groups (single parents and couples with and without children) are, again in a statistical sense, significantly different from the 1998 value. Yet, for all but one group, re-weighting either results in an improved match with the 1998 values (couples with and without children) or causes only very small changes relative to the unadjusted case (single parents, singles). The very heterogeneous group of “other” family types exhibits a slightly increased discrepancy after re-weighting which is, however, not statistically significant.

Table 2b shows frequencies for recipients of different income types – a variable that was partly controlled for in the re-weighting process. Unsurprisingly, we see near-perfect matches for categories used as calibration targets (earned income, unemployment benefits). Results for a number of other income sources are also quite close. The number of child benefit recipients has only changed slightly between 1996 and 1998. But even for bigger changes, one would expect the re-weighting algorithm to perform well as it controls for the number of people in the relevant age-groups.

Capital income is one income component that has not been controlled for. The large change in capital gains between 1996 and 1998 is mainly a result of the stock market boom during the late 1990s where the number of people holding stocks increased markedly. This type of change is difficult to control for in the calibration process since (1) the group of income receivers can be expected to be very heterogeneous; and (2) the number of observations is small. For incomes from capital as a whole, the change due to re-weighting is

[10] While it is possible to compute confidence intervals for frequency differences that take into account that the two samples are independent, we did not consider this crucial for the purpose of this case study.

small. Capital gains, which saw particularly large changes between 1996 and 1998, remain very far from the 1998 value.

An interesting case is the number of persons receiving unemployment benefits which has dropped markedly between 1996 and 1998. The total has been controlled for, leading to a close correspondence between the 1998 figures and those obtained from the re-weighted 1996 dataset (as well as a zero-sized confidence interval). The two components (minimum basic benefit and earnings-related part) are also closer to the 1998 value than before re-weighting. Significant discrepancies remain however. Formally, the two discrepancies must be pointing in different directions given that the size of these two groups determines the total number of unemployment benefit recipients.[11] In terms of economic processes, the compositional changes in the unemployed group are determined by diverging experiences in the labour market which accompanied the overall drop in unemployment rates. Despite the economic boom in the late 1990s unemployment rates were still high in 1998 with those receiving only basic unemployment benefit less likely to move off the benefit. This compositional change cannot be captured by a re-weighting scheme that forces the *total* number of recipients to match a particular target because the process does not distinguish between different types of unemployment benefit recipients. A priori, each recipient therefore has a similar probability of having her statistical weight reduced, resulting in an insufficient drop for recipients of earnings-related benefits and an overshoot for the number of individuals receiving basic benefits.

The number of people receiving social assistance drops between 1996 and 1998. Since wage earners (whose number increases) are less likely to receive these benefits than recipients of unemployment benefits (whose number decreases) and since both these variables are used as control totals, the re-weighting process is able to capture this decrease even if housing and social assistance benefits are not directly controlled for. This is an illustration of the point made in Section 2.1 that re-weighting will improve the match for variables that are *not* being controlled for if they are strongly correlated with the control variables. However, while one would suspect similar mechanisms in the case of housing benefits, we see that re-weighting actually worsens the match with 1998 values. This can be explained by a change in policy rules resulting in broadened entitlement to this benefit vis-à-vis 1996: While reduced unemployment would have resulted in lower numbers of housing benefit recipients (which is borne out by the frequencies obtained after re-weighting) the policy change more than compensated for this effect, resulting in an overall increase in the number of beneficiaries.

In Table 2c we consider individuals' main socio-economic status during the year. While some categories are closely related to the income variables we controlled for (unemployment benefits and income from employment, self-employment or farming activities), the correspondence is not exact since the *main* status in a given year as defined in the data depends on the level of these incomes. As in Table 2b, we find that re-weighting improves the match with 1998 values or leaves the frequencies roughly unchanged. An interesting case is that of pensioners. While our controlling for age should result in quite a good match with the 1998 value, we find that re-weighting actually produces a change in the wrong direction. While the adjusted 1996 weights indicate an increase in the number of pensioners, the 1998 wave of the IDS shows that the number has dropped. The reason for this is that in 1996 more than 200000 persons received an exceptional and small one-off lump sum pension payment that was no longer available in 1998.

4.2. Results 2: Frequency distributions

The totals we used as target values in the re-weighting process generally do not control for joint distributions (except in the case of gender/age group). So even if the total frequencies for particular values of two variables x_1 and x_2 (e.g. employment status and region) match 1998 target frequencies, their *joint* distribution

[11] A number of unemployed persons receive both types of benefits during a given year which is why the number of total recipients is less than the sum of the two components.

cannot generally be expected to show close matches with their 1998 counterparts. For any combination of x_1 and x_2, the adjustment of weights will take into account the 1996-1998 differences in the total number of observations with the relevant x_1 and x_2 values (for instance, the total number of unemployed and the total number of people living in region 1 may have increased). But in addition to changes in these marginal totals, the cell sizes of particular combinations may have changed independently (e.g. the number of unemployed in region 1 may have decreased) and these "interaction terms" x_1 X x_2 are not captured by controlling for each of the x_1 and x_2 frequencies separately. Table 2d shows the re-weighting results for different combinations of region x income receipt.[12] We mostly find that the match between re-weighted 1996 data and 1998 data is substantially improved indicating that the "interaction terms" for the two chosen variables are less important than the 1996-1998 change in the marginal totals. The worst match is found for the Helsinki region. This is a sub-region of a region used as a control-total (Uusimaa in the south, see Table 1) whereas all other regions in Table 2d encompass regions that are controlled for and whose totals therefore match the reference value. Even for Helsinki, however, we find statistically insignificant discrepancies between the frequencies derived from the re-weighted 1996 data and the respective 1998 values.

A more complete picture of how re-weighting alters the size of population groups with different combinations of characteristics is presented in Table 3, which shows cross-tabulations of certain control variables X, and Table 4 which tabulates control variables X against a number of other variables V that were not controlled for. In both tables, frequencies in each cell are compared between the 1998, the 1996 and the re-weighted 1996 datasets. The value in each cell shows to what extent re-weighting results in a better match with the 1998 reference values. The reduction in the distance to the 1998 value is expressed as percent of the 1998 frequency with a positive value indicating that re-weighting produces an improved correspondence and vice versa (empty cells are shown where there are fewer than 15 observations). For example, a value of +10% (-5%) would be shown for a cell where the 1998, 1996 and re-weighted 1996 frequencies are, respectively, 100, 85 and 95 (100, 105, 110). Un-shaded cells indicate that the match differs by less than 1% of the 1998 value.

In Table 3, the values along the diagonal for age and income source are positive by definition since the re-weighting process aligns the frequencies for these control variables. For instance, we see from that the number of persons receiving unemployment benefits (farming income) changes by 10% (13%) between 1996 and 1998 and that the re-weighting algorithm ensures that the adjusted 1996 weights reflect this change. For cells defined by a combination of different variables, there is no guarantee that the adjusted weights would result in an improvement. As discussed in the unemployment/region example above, specific cells of the joint distribution of two variables can be subject to changes that are not well captured by adjusting for the marginal totals.

While the re-weighting process only controls for aggregate frequencies in each of the shown categories, it also improves the correspondence with 1998 values for a large number of sub-groups that are not explicitly controlled for (lightly-shaded cells). Increasing discrepancies are, however, observed for a considerable number of cells (darker colour). A number of general patterns a worth noting. First, the extent to which frequencies for sub-group of a controlled-for category (e.g. unemployment benefit recipients) change between 1996 and 1998 will, on average, mirror the magnitude of the change for the category as a whole. Groups with larger overall movements between 1996 and 1998 (e.g. number of farmers, number of unemployed) therefore tend to be subject to larger changes vis-à-vis the unadjusted 1996 weights than groups whose frequency totals change less markedly (number of wage earners, number of self-employed). Whether these changes result in an improving or worsening correspondence with the target value depends on the

[12] While the income types correspond exactly to the income variables used for re-weighting, the region categories are more aggregated than those used as control variables

extent to which the changes in the size of a particular sub-group mirror the changes in the aggregates used as control totals. Where trends for a particular sub-group run counter the movement for the group as a whole, re-weighting will exacerbate group-size discrepancies. An example for such a diverging trend is the number of older men receiving unemployment benefits: while overall unemployment rates decreased between 1996 and 1998, data from labour force surveys show that the drop was substantially larger for older employees. Aligning the 1996 weights with the 1998 totals therefore fails to improve the match for the unemployed in the 58-64 age group and, instead, distorts the group size by a further 6% of the 1998 value. In addition to these "substantive" reasons for deteriorating cell-size matches, large distortions also result from insufficient numbers of observations (and the resulting large standard errors). For instance, the large negative values in cells [farm income; unemployment benefit] or [wage earner; age 65+] are caused by the small number of observations combining the respective characteristics rather than by any specific shortcoming of the calibration.

Each of the columns in Table 3 represents a category that is calibrated to match the 1998 frequency for the population as a whole. While all cells in a given column therefore show an improvement on average, this is not the case for Table 4, which shows results for sub-categories that are defined in terms of categories not controlled for. Compared to Table 3, we see a larger number of cells where re-weighting makes adjustments "in the wrong direction" (dark-shaded cells).

4.3. Results 3: Effects of re-weighting on typical tax-benefit model output

A significant deficiency of the evaluation offered by Tables 2, 3 and 4 is that they consider the *absolute* match for each of the sub-groups separately. Often, one is interested in the distribution itself and, thus, the *relative* difference between the frequencies. Even if all sub-groups of a particular category show a "better" absolute correspondence with the reference values, re-weighting may still result in undesirable distortions of the relative size of the sub-groups. In addition, the distribution within each of the sub-groups will also be altered when calibrating against changing frequency totals. Since incomes are correlated with a large number of characteristics, these potential distortions are especially relevant when the re-weighted data are used for analysing income distributions. Distributional analyses are the main use of tax-benefit models and it is therefore necessary to examine the effects of re-weighting on measures typically produced by these models.

Before using monetary variables of the 1996 dataset as an input into tax-benefit calculations for 1998, it is necessary to adjust the amounts recorded in the 1996 data. Using information from the 1996 and 1998 datasets, one could do this separately for each monetary variables v and for each of the i groups G_i in Figure 1. When "ageing" a dataset in practice, however, it is unlikely that detailed micro-data for the target year are available. The set of available uprating factors will generally much more limited and may often be restricted to the average change per income recipient. Rather than constructing detailed uprating factors from the 1998 micro-data, we have therefore used factors that would be available in practice. This includes separate indices for most income variables, including income from employment, self-employment, investments and social transfers.[13] These are, however, generally not differentiated by population sub-group so that all recipients are assumed to experience an equal increase or decrease.

A number of these indicators are shown in Table 5 for the 1996 and 1998 datasets as well as the re-weighted 1996 data. Monetary variables in both 1996 datasets were adjusted to 1998 levels using the same set of uprating factors. All income-related indicators relate to current cash incomes.[14] The numbers are based on

[13] Details are provided in the EUROMOD country report for Finland available on http://www.econ.cam.ac.uk/dae/mu/emod.htm.

[14] Monthly cash market incomes minus direct taxes plus cash social transfers. Household incomes are adjusted for differences in household size and composition using the 'modified OECD' equivalence scale (with weights 1 for the

household incomes as calculated by the EUROMOD tax-benefits model, that is, they are based on simulated tax and benefit amounts.[15] There has been a significant rise in inequality between 1996 and 1998 with the Gini coefficient (the standard error is 0.003), quantile ratios and poverty headcounts all pointing in the same direction. Moreover, the sub-group Ginis indicate that income disparities have increased in for all seven family types. Unsurprisingly, given the lack of distributional adjustments, neither the adjustment of the 1996 weights nor the uniform uprating of monetary variables is able to capture this increase reliably. While certain characteristics used in constructing the control totals, such as the number of unemployment benefit recipients, are clearly associated with household incomes, the changing sizes of these groups are only one of the influences on inequality. Adjusting for group sizes cannot capture changes in within-group inequality (such as those caused by the increasing importance of capital incomes). For headcount-based measures as well as income medians and means, the re-weighting procedure is doing better: very low incomes (particularly those below the 50% cut-off) appear to be reasonably correlated with the characteristics used in the re-weighting process.

In addition to the distribution of household incomes, tax-benefit models are commonly used to assess the distributive properties of different tax and benefit instruments. Results for the four main types of instrument simulated by EUROMOD are shown in Table 6. For income taxes, the pattern is quite similar to that of household incomes as a whole. While the re-weighting procedure produces substantially improved matches with 1998 aggregates and averages, it is ineffective at reproducing 1996-98 changes in the distribution. The concentration index, which measures how unequal an instrument is distributed across the household income spectrum, shows that 1998 income taxes (panel B) are clearly more progressive than in 1996 (panel A). This rise in progressivity is, however, not reflected by the numbers based on re-weighted 1996 data (panel C). For employees' social insurance contributions we see a better correspondence of the distributional statistics in panels B and C. One explanation can be seen in the fact that contributions are not influenced by capital incomes which is known to have been a particularly volatile income component in the 2-year period considered here.

Universal benefits, such as the main family transfers, are only mildly inequality reducing (indicated by the negative but small concentration index numbers). They do not depend on income and their distributional properties are therefore not affected by changing disparities of market incomes. Instead, these benefits are mainly determined by the number of children in different age groups and the sizes of these age-groups are controlled for as part of the re-weighting process. The distribution of benefit payments by income groups changes very little between 1996 and 1998 and the re-weighting method is performing well in this environment leading to close matches between the numbers in panels B and C. For means-tested benefits, however, the picture is entirely different. Re-weighting does not perform well at all. In fact both aggregate measures and distributional indicators are further from the 1998 values after re-weighting than before. Means-tested benefits are highly targeted and depend, simultaneously, on a large number of individual and family characteristics including income type, income levels, employment status, age, and family composition. Calibrating the data against a limited set of aggregate control totals is therefore unlikely to capture the complex processes that changes in benefit receipt over time.

Summary and Conclusions

This paper has considered methods for "ageing" micro-data used as input for tax-benefit microsimulation models. It has presented a conceptual framework for thinking about different types of data adjustments and

---

first adult, 0.5 for further adults and 0.3 for children aged under 14). For the purpose of constructing the income distribution measures, each individual is counted with her equivalised household income.

[15] A description of the EUROMOD model is available in Immervoll *et al.* (1999), Sutherland (2000) and on the Internet at http://www.econ.cam.ac.uk/dae/mu/emod.htm.

illustrated the empirical effects of data ageing in a case study involving Finnish household micro-data from two periods (1996 and 1998). The case-study has evaluated the performance of one particular ageing technique by comparing results from the 1998 dataset with those derived from the aged 1996 data. This has been done by determining how well certain measures computed from the reference 1998 dataset can be approximated using an older dataset that has been aligned with characteristics of the 1998 population. When interpreting the results, one should bear in mind that any distortions of particular population characteristics have to be weighed against improvements along the dimensions that the re-weighting process is controlling for.

A number of general observations are worth emphasising.

- There is no "one-size-fits-all" ageing technique. When aligning existing data to information from a different period, on needs to have a clear idea about the types of changes one would like to capture. For instance, controlling for changes in aggregate group sizes cannot generally be expected to improve the match for distributional patterns. Ageing techniques also should not be applied mechanically over different time-periods since structural changes in the population or the tax-benefit system will to a large extent determine whether a given set of alignments is appropriate or not.
- Adding information about the target population by altering the statistical weights in a dataset comes at the cost of potentially distorting information that the original weights represent. More specifically, "improvements" to marginal distributions usually distort joint distributions. While "minimum-distance" techniques maintain as much information as possible, the likelihood of such distortions grows with the number of dimensions used for re-weighting as well as the magnitude of the change along each individual dimension. If the size or number of relevant changes becomes very sizable then forcing the data to correspond to the observed values in the target period can compromise the representativity along relevant dimensions. In such a situation, ageing-techniques do not provide a reliable approximation of the population of interest (clearly, large changes will also render the "unadjusted" data non-representative of the target population).
- When using the aged dataset as an input for tax-benefit microsimulation models, it is essential that the choice of reference values used in the calibration process is informed by detailed knowledge about the mechanics of tax-benefit rules. The precision of simulated tax-benefit amounts will, for instance, rest on a good representation of those age-groups, family circumstances and employment situations that play a crucial role in determining tax burdens and benefit entitlements.
- More generally, inter-dependencies between categories used for defining control totals need to be understood. For instance, there is no need to control for the number of recipients of a particular benefit if this benefit is universally paid to all individuals in a certain age-group and the size of the age-group has already been used as a calibration target.
- When implementing ageing techniques in practice, a large number of data and definitional issues need to be addressed. Obviously, the external reference values used as calibration controls need to be conceptually similar to the variables recorded in the micro-data. On a more technical level, one needs to distinguish between changes concerning group *sizes* and changes in the *characteristics* of a group. Different calibration methods (adjusting the statistical weights versus adjusting recorded variable values) will be appropriate in each case.

References

Anderson C and L Nordberg, 1998, *A SAS-program for the computation of point- and standard errors in sample surveys*, Statistics Sweden.

Creedy J, 2004, “Survey reweighting for tax microsimulation modelling”, in: J A Bishop and Y Amiel, eds., *Studies on Economic Well-Being: Essays in the Honor of John P. Formby* (Research on Economic Inequality, Vol. 12), 229-49.

Deville J C and C E Särndal, 1992, "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, 87 (418), pp. 376-82.

Immervoll, H., O’Donoghue, C. and Sutherland, H. 1999. “An Introduction to EUROMOD”, EUROMOD Working Paper EM0/99, www.econ.cam.ac.uk/dae/mu/publications/emwp0.pdf

Sutherland H., 2000, “EUROMOD”, in Gupta A. and V. Kapur (eds), *Microsimulation in Government Policy and Forecasting*, Elsevier, 575-580.

Table 1. Control totals from original 1996 and 1998 datasets

	Share of population	Number of individuals	
	1998	1996	1998
Male			
age<3	0.04	99889	90051
2<age<7	0.05	124650	132868
6<age<17	0.13	336272	331058
16<age<28	0.15	348771	361930
27<age<43	0.22	570521	543055
42<age<58	0.23	556161	573235
57<age<65	0.07	159486	166755
age>64	0.11	271324	281248
Female			
age<3	0.03	89576	83979
2<age<7	0.05	134324	126332
6<age<17	0.12	311570	320171
16<age<28	0.13	341822	333644
27<age<43	0.21	550153	550371
42<age<58	0.21	547235	550143
57<age<65	0.07	176727	190611
age>64	0.17	444885	450682
Region			
1. Province of Uusimaa	0.27	1338702	1359839
2. Province of Turku and Pori	0.13	688191	682544
4. Province of Häme	0.14	723038	734380
5. Province of Kymi	0.06	329796	321435
6. Province of Mikkeli	0.04	210984	213883
7. Province of North Karelian	0.03	170255	162657
8. Province of Kuopio	0.05	239856	245711
9. Province of Central Finland	0.05	257388	253071
10. Province of Vaasa	0.09	437280	444143
11. Province of Oulu	0.09	450531	448992
12. Province of Lapland	0.04	192789	198949
13. Province of Åland	0	24553	20535
Employment status			
number of wage-earners		2503885	2584987
number of unemployed		802041	731366
number of self-employed		242782	244871
number of farmers		160343	142381

Source: Income Distribution Surveys 1996 and 1998, Statistics Finland

Table 2a Household type, number of households

(1)	(2)	(3)		(4)	(5)	(6)	(7)
	Original 1996 data	Re-weighted 1996 data		95% confidence interval	Original 1998 data	Col 2/ 5	Col 3/ 5
1. Single parents with children aged 17-	93 992	93 778	±	10 949	108 915	86	86
2. Couples without children	580 998	594 818	±	20 104	621 217	94	96
3. Singles	874 740	875 842	±	**36 280**	902 357	97	97
4. Couples with children aged 17-	535 797	534 146	±	12 818	515 701	104	104
5. Single parents children 18+	77 022	75 773	±	10 880	56 847	135	133
6. Couples children 18+	101 865	104 572	±	**8 785**	98 653	103	106
7. Unknown/Other	45 587	44 208	±	**8 267**	51 310	89	86
Total	2310 000	2323 137	±	10 949	2 355 000	98	99

Table 2b Recipients of different income items, number of individuals.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Original 1996 data	Re-weighted 1996 data	95% confidence interval	Original 1998 data	Col 2/5	Col 3/5
Earned income	2 762 068	2 826 539	**13151**	2 830 827	98	100
Capital income	1 575 017	1 601 036	27344	1 675 133	94	96
- Interest and dividend	927 603	946 488	31189	1 099 355	84	86
- Capital gains	146 895	149 611	**13959**	219 697	67	68
Unemployment benefit	802 041	731 366	**0**	731 366	109	100
-Earnings related unemployment benefit	464320	427924	**15686**	413 258	112	104
- Basic unemployment benefit	444 700	400 994	**15295**	410 282	108	98
Sickness allowance	122 181	120 745	**11660**	132 065	93	91
Maternity payment	155 056	143 315	9125	128 033	121	112
Child benefit	647 552	644 739	**12130**	638 989	101	101
Housing benefit	523 127	507 915	27981	553 526	95	92
Social assistance	319 829	298 555	**20197**	297 073	108	100
National pension	1 038 012	1 059 740	17127	987 940	105	107
Earnings-related pension	1 084 178	1 112 132	**21074**	1 102 605	98	101

Table 2c Socio-economic status, number of individual

(1)	(2)	(3)		(4)	(5)	(6)	(7)
	Original 1996 data	Re-weighted 1996 data		95% confidence interval	Original 1998 data	Col 2/5	Col 3/5
1. Farmers	108369	98339	±	**9191**	92093	118	107
2. Self-employed	213139	216418	±	**13501**	215309	99	101
3. Employee	1660459	1715449	±	24866	1765389	94	97
4. Pensioner	1093201	1118324	±	16428	1085117	101	103
5. Unemployed	414653	371652	±	**15469**	359044	115	104
6. Student	382177	395256	±	**18053**	397515	96	99
7. Inactive	29094	29229	±	5451	23093	126	126
8. Sick or Disabled	50414	46052	±	**5727**	47266	107	97
9. Other	663422	662183	±	**10954**	668080	99	99
0. Pre-school	448439	433230	±	**0**	433233	104	100
Total	5063367	5086131	±		5086139	100	100

Table 2d Number of wage earners, self-employed, farmers and unemployed by region

Number of wage earners by region

(1)	(2)	(3)		(4)	(5)	(6)	(7)
	Original 1996 data	Re-weighted 1996 data		95% confidence interval	Original 1998 data	Col 2/5	Col 3/5
Helsinki Region	512 119	533 703	±	**23 969**	550 518	93	97
Southern Finland	1 104 217	1 132 370	±	**26 338**	1 126 893	98	100
Middle Finland	586 917	606 611	±	**18 059**	601 827	98	101
Northern Finland	300 632	312 302	±	**13 310**	305 748	98	102
Total	2 503 885	2 584 986	±	**23 969**	2 584 987	97	100

Number of self-employed by region

(1)	(2)	(3)		(4)	(5)	(6)	(7)
	Original 1996 data	Re-weighted 1996 data		95% confidence interval	Original 1998 data	Col 2/5	Col 3/5
Helsinki Region	29 672	30 606	±	**5 702**	27 879	106	110
Southern Finland	102 153	102 586	±	**7 807**	106 987	95	96
Middle Finland	78 395	78 147	±	**7 136**	76 794	102	102
Northern Finland	32 562	33 532	±	**5 512**	33 212	98	101
Total	242 782	244 871	±	**5 702**	244 872	99	100

Number of farmers by region

(1)	(2)	(3)		(4)	(5)	(6)	(7)
	Original 1996 data	Re-weighted 1996 data		95% confidence interval	Original 1998 data	Col 2/5	Col 3/5
Helsinki Region	1 130	1 096	±	**990**	1 560	72	70
Southern Finland	70 823	61 971	±	**6 619**	61 915	114	100
Middle Finland	68 586	61 388	±	**6 290**	63 510	108	97
Northern Finland	19 803	17 926	±	**4 006**	15 396	129	116
Total	160 343	142 381	±	**990**	142 382	113	100

Number of unemployed by region

(1)	(2)	(3)		(4)	(5)	(6)	(7)
	Original 1996 data	Re-weighted 1996 data		95% confidence interval	Original 1998 data	Col 2/5	Col 3/5
Helsinki Region	113 282	103 908	±	**12 121**	95 675	118	109
Southern Finland	342 903	310 993	±	**15 057**	311 116	110	100
Middle Finland	222 519	202 740	±	**12 816**	208 410	107	97
Northern Finland	123 337	113 725	±	**10 061**	116 166	106	98
Total	802 041	731 366	±	**12 121**	731 366	110	100

Table 3. Joint distribution of control variables (crosstable X x X).

Match with frequencies from 1998 data: improvement from re-weighting versus unadjusted 1996 data

							Men								Women							
		all	*n unemployed (receiving benefits)*	*n wage earners*	*n farmers*	*n self-employed*	*0 to 2*	*3 to 6*	*7 to 16*	*17 to 27*	*28 to 42*	*43 to 57*	*58 to 64*	*65+*	*0 to 2*	*3 to 6*	*7 to 16*	*17 to 27*	*28 to 42*	*43 to 57*	*58 to 64*	*65+*
coREGION	*1*	**2%**	**10%**	**3%**	**-9%**	**-2%**	**-3%**	**6%**	0%	**4%**	**3%**	**-2%**	**5%**	**-5%**	**6%**	**3%**	**4%**	-1%	1%	**-3%**	**7%**	0%
	2	1%	**-3%**	**-2%**	**15%**	**1%**	**16%**	**-6%**	**-4%**	**3%**	**3%**	**-2%**	**-3%**	**-3%**	**8%**	**7%**	1%	**3%**	**-1%**	-1%	**5%**	-1%
	4	**2%**	**9%**	1%	**13%**	**2%**	**8%**	**5%**	0%	**5%**	**4%**	**-4%**	**-5%**	**-4%**	**4%**	**-5%**	**3%**	**-2%**	**1%**	0%	**-12%**	**2%**
	5	**3%**	**12%**	1%	**4%**	0%		-1%	**8%**	**-1%**	**9%**	0%	**-2%**	**2%**			0%	**7%**	**4%**	**-2%**	**5%**	1%
	6	**1%**	1%	**4%**	**15%**	**-2%**	**-8%**		**3%**	1%	**-4%**	**-3%**	**-6%**	**3%**			**4%**	**5%**	**3%**	0%	**-5%**	**-2%**
	7	**5%**	-1%	**2%**	**-15%**	1%			**10%**	**5%**	**-9%**	**2%**	1%	**2%**			**4%**	**13%**	**5%**	**6%**	**3%**	-1%
	8	**2%**	**-8%**	**-5%**	**15%**	1%			**-4%**	**-6%**	**6%**	**3%**	**4%**	**4%**		**2%**	**-7%**	**3%**	**-5%**	1%	**6%**	**-3%**
	9	**2%**	**-11%**	**2%**	**-1%**	1%		**-4%**	**1%**	**1%**	**8%**	**-1%**	**-4%**	**1%**		**-6%**	0%	**6%**	**2%**	-1%	**4%**	0%
	10	**2%**	**8%**	**4%**	**-7%**	**2%**	**-6%**	**-8%**	-1%	**-6%**	**-3%**	**3%**	**6%**	**-4%**	**-2%**	**-7%**	**2%**	0%	1%	**-1%**	**7%**	**-3%**
	11	0%	**10%**	**-3%**	**15%**	**3%**	**18%**	**5%**	**3%**	**3%**	**6%**	**2%**	**-6%**	**4%**	**-9%**	**8%**	**-2%**	**-2%**	**2%**	0%	**-11%**	**2%**
	12	**3%**	**-5%**	**5%**		**-6%**			**2%**	**-1%**	**2%**	**5%**	**5%**	**-2%**		0%	**-9%**	1%	**-3%**	**2%**	**9%**	**4%**
	13	**20%**		**21%**																		
	n unemployed (receiving UB)	**10%**	**10%**	**6%**	**-21%**	**-3%**				**7%**	**16%**	**5%**	**-6%**					**16%**	**-7%**	**-9%**	-1%	
	n wage earners	**3%**	**6%**	**3%**	**9%**	**5%**			**6%**	**5%**	**1%**	**3%**	**10%**	**-14%**			**-8%**	0%	**2%**	**2%**	**12%**	**-10%**
	n farmers	**13%**	**-21%**	**9%**	**13%**						**7%**	**-8%**	**7%**	-1%					**18%**	**15%**	**8%**	**-8%**
	n self-employed	**2%**	**-3%**	**5%**		**2%**				**4%**	**5%**	**4%**	**6%**	**4%**					**-2%**	0%	**6%**	**3%**
Men	*0 to 2*	**11%**					**11%**															
	3 to 6	**6%**						**6%**														
	7 to 16	**2%**		**6%**					**2%**													
	17 to 27	**4%**	**7%**	**5%**		**4%**				**4%**												
	28 to 42	**5%**	**16%**	**1%**	**7%**	**5%**					**5%**											
	43 to 57	**3%**	**5%**	**3%**	**-8%**	**4%**						**3%**										
	58 to 64	**4%**	**-6%**	**10%**	**7%**	**6%**							**4%**									
	65+	**4%**		**-14%**	-1%	**4%**								**4%**								
Women	*0 to 2*	**7%**													**7%**							
	3 to 6	**6%**														**6%**						
	7 to 16	**3%**		**-8%**													**3%**					
	17 to 27	**2%**	**16%**	0%														**2%**				
	28 to 42	0%	**-7%**	**2%**	**18%**	**-2%**													0%			
	43 to 57	1%	**-9%**	**2%**	**15%**	0%														1%		
	58 to 64	**7%**	-1%	**12%**	**8%**	**6%**															**7%**	
	65+	**1%**		**-10%**	**-8%**	**3%**																**1%**

Table 4. Joint distribution of control variables and other variables (crosstable X x V)

Match with frequencies from 1998 data: improvement from re-weighting versus unadjusted 1996 data

			HH type							marital status				children in HH		
		all	other no children	other with children	singles female	singles male	single parents	couples no children	couples with children	**single**	**married**	**separated**	**widowed**	**1**	**2**	**3+**
coREGION	*1*	**2%**	**-3%**	**-7%**	0%	0%	0%	**3%**	0%	0%	**-3%**	1%	0%	**-2%**	**-1%**	**-2%**
	2	1%	1%	**-3%**	0%	0%	**-2%**	1%	**4%**	**-1%**	0%	1%	0%	1%	**5%**	**-3%**
	4	**2%**	**-1%**	**6%**	**1%**	-1%	-1%	**3%**	0%	-1%	**-3%**	1%	**2%**	**2%**	0%	1%
	5	**3%**	**3%**	-1%	0%	**2%**	**12%**	**1%**	**-4%**	**4%**	0%	**5%**	**-3%**	**3%**	**7%**	**14%**
	6	**1%**	**2%**	**-6%**	-1%	**-2%**	**-4%**	**3%**	**1%**	**-3%**	**-3%**	0%	**-1%**	**2%**	0%	**-5%**
	7	**5%**	**6%**	**11%**	**-2%**	**-1%**		**-1%**	**-7%**	**-3%**	**3%**	**9%**	**1%**	**8%**	**-4%**	**15%**
	8	**2%**	**5%**	**-8%**	**-3%**	0%	**-8%**	**3%**	**3%**	**-3%**	**4%**	1%	**-3%**	0%	**-6%**	**-5%**
	9	**2%**	**6%**	**-3%**	0%	0%	**-6%**	**1%**	**5%**	**-1%**	0%	**3%**	**-2%**	0%	**4%**	**-7%**
	10	**2%**	**-2%**	**3%**	0%	0%	1%	**4%**	0%	0%	**-3%**	1%	-1%	**-2%**	0%	-1%
	11	0%	**-1%**	**-7%**	**-2%**	**-1%**	**-4%**	**2%**	**6%**	-1%	**1%**	**2%**	**-2%**	**-1%**	**3%**	**4%**
	12	**3%**	**-6%**	**-4%**	**3%**	**-3%**	**-1%**	**5%**	**2%**	**1%**	**4%**	**3%**	**1%**	-1%	**2%**	**-7%**
	13	**20%**			**-3%**			**-8%**	**31%**		**9%**				**-17%**	
	n unemployed (receiving UB)	**10%**	**18%**	**6%**	**-6%**	**7%**	**9%**	**-5%**	**-11%**	**11%**	0%	**11%**	**-7%**	**9%**	**11%**	**9%**
	n wage earners	**3%**	**-6%**	**-8%**	**-2%**	**2%**	**2%**	**3%**	**1%**	**2%**	**-4%**	**2%**	**2%**	**-4%**	**1%**	**2%**
	n farmers	**13%**	**12%**	**-12%**		**9%**		**-9%**	**20%**	**8%**	**12%**	**11%**		**14%**	**22%**	**-1%**
	n self-employed	**2%**	-1%	**-6%**	0%	0%		**3%**	0%	0%	**-3%**	0%	1%	**-3%**	**3%**	**1%**
Men	*0 to 2*	**11%**					**17%**		**-1%**					**7%**	**12%**	**8%**
	3 to 6	**6%**		**3%**			**7%**		**6%**					**-8%**	**-6%**	**5%**
	7 to 16	**2%**	**16%**	**-4%**			**3%**		**-2%**					**-1%**	**2%**	**-4%**
	17 to 27	**4%**	**-7%**	**-1%**		**-3%**		0%	**-4%**	**-3%**	0%	**4%**		**-2%**	**1%**	**2%**
	28 to 42	**5%**	**17%**	**2%**		**-1%**	0%	**-1%**	**-4%**	**7%**	**4%**	**-5%**	**-5%**	**2%**	**4%**	**2%**
	43 to 57	**3%**	**-3%**	**-8%**		1%		**1%**	**5%**	1%	**-4%**	0%	1%	**-5%**	**6%**	**4%**
	58 to 64	**4%**	**-2%**			**-3%**		**6%**	**3%**	**-4%**	**-2%**	**1%**	**2%**	**4%**		
	65+	**4%**	**1%**			**-2%**		**5%**		**1%**	**2%**	**2%**	**2%**			
Women	*0 to 2*	**7%**							**6%**					**8%**	**6%**	**-5%**
	3 to 6	**6%**		**5%**			**-6%**		**6%**					**-3%**	**7%**	**7%**
	7 to 16	**3%**	**7%**	**-6%**			**2%**		**2%**					**-4%**	**3%**	1%
	17 to 27	**2%**	**3%**	**-3%**	**2%**		**18%**	0%	0%	**-1%**	**2%**	1%		**4%**	**9%**	**4%**
	28 to 42	0%	1%	**-5%**	**2%**		0%	0%	**-1%**	**2%**	0%	**-2%**	0%	0%	0%	**-2%**
	43 to 57	1%	-1%	**-5%**	**-3%**		-1%	0%	**1%**	**-2%**	**-2%**	**-3%**	**-2%**	**-3%**	1%	**-1%**
	58 to 64	**7%**	**-6%**		**5%**			**6%**		**5%**	**-5%**	**5%**	**5%**			
	65+	**1%**	**3%**		1%			**4%**		**1%**	**3%**	0%	1%			

Table 5. Weights and income distribution by family type

A. 1996 data

HH-type	Frequency	min weight	max weight	std. Dev weight	mean HH size	mean equivalised HDI	median equivalised HDI	quantile ratio 80/20	quantile ratio 90/10	Gini	poverty head count 50	poverty head count 60	poverty head count 70
all	2310000	10	2524	510	2.19	1163	1054	1.904	2.683	0.228	157153	462982	900477
other, no own children	433345	13	2060	474	2.58	1267	1185			0.215	22374	75285	150975
other, own children	108557	10	1171	299	4.44	1170	1091			0.196	15628	39523	68530
Singles, female	498615	31	2524	702	1.00	929	819			0.225	44427	126496	216686
Singles, male	376125	19	1433	641	1.00	1041	877			0.276	45117	93692	145838
Single Parents	71556	28	1103	532	2.53	1027	956			0.158	2415	6167	27589
Couples, no own children	421457	17	984	324	2.00	1376	1209			0.242	7510	29182	85705
couples, own children	400346	13	1032	263	3.91	1252	1175			0.207	19681	92638	205154

B. 1998 data

HH-type	Frequency	min weight	max weight	std. Dev weight	mean HH size	mean equivalised HDI	median equivalised HDI	quantile ratio 80/20	quantile ratio 90/10	Gini	poverty head count 50	poverty head count 60	poverty head count 70
all	2355000	10	1683	505	2.16	1212	1073	1.947	2.710	0.246	171110	465814	929680
other, no own children	236120	16	1529	386	2.87	1312	1196			0.234	17000	46279	99728
other, own children	93074	10	744	237	4.62	1185	1080			0.216	7160	33238	74729
Singles, female	521678	27	1683	698	1.00	947	829			0.234	55808	137651	223177
Singles, male	380679	22	1413	650	1.00	1072	931			0.262	41432	87206	140769
Single Parents	84913	26	1252	525	2.48	1019	959			0.172	7069	17195	46754
Couples, no own children	621217	23	1296	378	2.00	1434	1266			0.257	18860	57925	149845
couples, own children	417319	17	936	277	3.89	1330	1202			0.225	23780	86321	194677

C. 1996 data reweighted

HH-type	Frequency	min weight	max weight	std. Dev weight	mean HH size	mean equivalised HDI	median equivalised HDI	quantile ratio 80/20	quantile ratio 90/10	Gini	poverty head count 50	poverty head count 60	poverty head count 70
all	2323137	8	2576	504	2.19	1178	1071	1.895	2.663	0.228	170006	473699	907593
other, no own children	433765	11	1788	456	2.58	1286	1197			0.215	25282	73134	152273
other, own children	113139	8	950	278	4.44	1184	1103			0.194	14122	40875	68350
Singles, female	501273	30	2576	706	1.00	936	822			0.226	48850	133444	223888
Singles, male	374568	18	1281	631	1.00	1052	895			0.275	49591	95352	143470
Single Parents	70551	25	1141	501	2.52	1036	965			0.158	3084	6217	33199
Couples, no own children	436276	15	1088	335	2.00	1392	1221			0.243	6986	32880	90864
couples, own children	393564	10	1079	260	3.91	1272	1197			0.206	22091	91797	195550

Table 6. Simulated income component.

Simulated Taxes

	total cash amount	mean equivalised instrument	median equivalised instrument	Concentration index	mean quintile 1	mean quintile 2	mean quintile 3	mean quintile 4	mean quintile 5
A. 1996 data									
all	1341453740	366	284	0.415	75	191	312	467	894
other, no own children	305994358	395	325	0.396	76	171	283	428	842
other, own children	106569682	391	329	0.196	125	238	332	494	885
Singles, female	120324555	241	141	0.520	68	198	347	495	915
Singles, male	121143982	322	202	0.524	75	201	333	516	921
Single Parents	26732929	247	187	0.441	60	132	290	504	853
Couples, no own children	296583421	469	352	0.429	79	174	288	439	904
couples, own children	364104812	435	354	0.365	107	219	325	476	914
B. 1998 data									
all	1435303121	387	296	0.426	80	199	326	485	962
other, no own children	187709217	407	318	0.413	74	171	277	455	917
other, own children	100055117	415	330	0.216	131	249	348	523	1074
Singles, female	129673924	249	157	0.515	73	216	360	543	929
Singles, male	128487698	338	232	0.500	86	203	358	499	991
Single Parents	27729618	220	184	0.403	60	157	298	357	818
Couples, no own children	457740200	491	375	0.435	76	182	303	457	948
couples, own children	403907347	464	370	0.367	112	220	337	503	1008
C. 1996 data reweighted									
all	1386670608	375	294	0.412	77	199	324	479	913
other, no own children	316064628	407	333	0.393	77	180	292	440	862
other, own children	114498208	401	335	0.194	128	246	349	507	915
Singles, female	122968432	245	143	0.520	73	205	362	504	922
Singles, male	123129867	329	211	0.522	74	215	346	527	935
Single Parents	27272447	256	192	0.435	62	143	314	509	896
Couples, no own children	313857118	480	363	0.429	79	181	297	454	923
couples, own children	368879907	448	363	0.361	111	227	333	490	934

children: aged 16-

Simulated Own Social Insurance Contributions

	total cash amount	mean equivalised instrument	median equivalised instrument	Concentration index	mean quintile 1	mean quintile 2	mean quintile 3	mean quintile 4	mean quintile 5
A. 1996 data									
all	298678976	81	62	0.383	15	43	75	114	184
other, no own children	70768600	92	79	0.377	15	40	69	110	184
other, own children	25279654	93	85	0.294	31	60	86	121	186
Singles, female	26248268	53	29	0.533	13	41	83	121	184
Singles, male	26181275	70	37	0.523	13	44	78	130	188
Single Parents	6342797	59	48	0.432	9	34	79	121	171
Couples, no own children	57936771	92	64	0.424	13	32	55	90	179
couples, own children	85921611	103	92	0.302	30	57	86	122	187
B. 1998 data									
all	313572745	84	68	0.377	16	47	82	118	187
other, no own children	40249159	87	77	0.368	13	39	68	109	172
other, own children	21876108	92	85	0.281	33	64	93	121	175
Singles, female	28269050	54	30	0.523	13	48	92	127	179
Singles, male	28684635	75	49	0.495	16	48	94	129	198
Single Parents	7042070	56	57	0.397	10	45	84	92	161
Couples, no own children	95562311	103	82	0.390	17	39	67	108	189
couples, own children	91889411	106	98	0.301	24	58	90	127	191
C. 1996 data reweighted									
all	310886726	83	66	0.376	16	45	79	117	188
other, no own children	73793027	96	85	0.369	16	42	72	114	188
other, own children	27392531	96	89	0.285	32	63	91	126	190
Singles, female	26918946	54	29	0.531	14	44	87	121	187
Singles, male	26801465	72	39	0.516	14	47	83	132	190
Single Parents	6528000	61	54	0.419	10	38	84	121	176
Couples, no own children	61513978	94	67	0.422	13	34	58	93	184
couples, own children	87938779	107	98	0.292	34	60	90	125	192

children: aged 16-

Table 6. (continued).

Simulated Universal Benefits

	total cash amount	mean equivalised instrument	median equivalised instrument	Concentration index	mean quintile 1	mean quintile 2	mean quintile 3	mean quintile 4	mean quintile 5
A. 1996 data									
all	112169236	22	0	-0.047	14	27	30	24	18
other, no own children									
other, own children	16450481	56	39	-0.099	74	58	55	50	48
Singles, female									
Singles, male									
Single Parents	10754922	94	69	-0.009	88	98	94	93	86
Couples, no own children									
couples, own children	79327336	88	95	-0.069	97	99	92	82	76
B. 1998 data									
all	111156255	21	0	-0.046	15	26	28	25	17
other, no own children									
other, own children	15192191	58	39	-0.115	71	64	55	48	49
Singles, female									
Singles, male									
Single Parents	12348202	92	69	-0.011	88	97	85	101	87
Couples, no own children									
couples, own children	81531988	87	95	-0.063	91	102	92	80	76
C. 1996 data reweighted									
all	110909932	22	0	-0.049	14	26	30	24	18
other, no own children									
other, own children	17063943	56	39	-0.104	74	58	55	49	46
Singles, female									
Singles, male									
Single Parents	10591060	94	69	-0.012	89	94	97	88	87
Couples, no own children									
couples, own children	77938209	88	95	-0.073	99	99	92	81	76

Simulated Means-tested Benefits

	total cash amount	mean equivalised instrument	median equivalised instrument	Concentration index	mean quintile 1	mean quintile 2	mean quintile 3	mean quintile 4	mean quintile 5
A. 1996 data									
all	204411929	49	0	-0.198	71	60	46	32	24
other, no own children	27927468	35	0	-0.272	80	40	33	22	14
other, own children	23313108	83	67	-0.127	127	82	78	66	70
Singles, female	10123732	20	0	-0.270	26	25	8	2	8
Singles, male	15779630	42	0	-0.480	79	23	14	5	2
Single Parents	25771010	233	192	-0.173	297	284	167	127	117
Couples, no own children	2223750	4	0	-0.727	26	2	2	1	0
couples, own children	99273232	112	95	-0.166	183	135	104	87	81
B. 1998 data									
all	220858522	51	0	-0.228	83	58	46	34	23
other, no own children	14444324	31	0	-0.320	88	21	20	18	18
other, own children	22092147	88	69	-0.161	117	103	83	66	58
Singles, female	10649400	20	0	-0.442	35	7	8	5	5
Singles, male	15825693	42	0	-0.502	77	30	17	4	3
Single Parents	31381827	240	167	-0.172	325	271	156	159	119
Couples, no own children	6640600	7	0	-0.789	45	5	3	1	0
couples, own children	119824531	130	95	-0.196	234	173	120	94	86
C. 1995 data reweighted									
all	198128488	47	0	-0.196	68	58	44	32	23
other, no own children	26897824	34	0	-0.270	76	40	33	21	14
other, own children	23771456	80	63	-0.127	124	80	76	67	63
Singles, female	9582747	19	0	-0.267	23	25	7	2	8
Singles, male	14902494	40	0	-0.488	74	23	14	6	2
Single Parents	24681737	226	178	-0.175	298	264	157	121	118
Couples, no own children	2110045	3	0	-0.720	24	2	2	0	0
couples, own children	96182187	110	95	-0.166	184	131	103	86	80